JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 8 (2020), Issue 10 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Review

Clinical Decision Support Systems for Pressure Ulcer Management: Systematic Review

Sabrina Magalhaes Araujo[1], RN, MSc; Paulino Sousa[2,3], RN, MScN, PhD; Inês Dutra[4,5], PhD

[1]Medical Informatics, Faculty of Medicine and Faculty of Sciences, University of Porto, Porto, Portugal

[2]Nursing School of Porto, Porto, Portugal

[3]Health Information Systems & Electronic Health Records, Center for Health Technology and Services Research, University of Porto, Porto, Portugal

[4]Department of Computer Science, Faculty of Sciences, University of Porto, Porto, Portugal

[5]Artificial Intelligence for Health Care, Center for Health Technology and Services Research, University of Porto, Porto, Portugal

Corresponding Author:
Paulino Sousa, RN, MScN, PhD
Nursing School of Porto
Rua Dr. António Bernardino de Almeida
Porto, 4200-072
Portugal
Phone: 351 225073500
Email: paulino@esenf.pt

Abstract

Background: The clinical decision-making process in pressure ulcer management is complex, and its quality depends on both the nurse's experience and the availability of scientific knowledge. This process should follow evidence-based practices incorporating health information technologies to assist health care professionals, such as the use of clinical decision support systems. These systems, in addition to increasing the quality of care provided, can reduce errors and costs in health care. However, the widespread use of clinical decision support systems still has limited evidence, indicating the need to identify and evaluate its effects on nursing clinical practice.

Objective: The goal of the review was to identify the effects of nurses using clinical decision support systems on clinical decision making for pressure ulcer management.

Methods: The systematic review was conducted in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) recommendations. The search was conducted in April 2019 on 5 electronic databases: MEDLINE, SCOPUS, Web of Science, Cochrane, and CINAHL, without publication date or study design restrictions. Articles that addressed the use of computerized clinical decision support systems in pressure ulcer care applied in clinical practice were included. The reference lists of eligible articles were searched manually. The Mixed Methods Appraisal Tool was used to assess the methodological quality of the studies.

Results: The search strategy resulted in 998 articles, 16 of which were included. The year of publication ranged from 1995 to 2017, with 45% of studies conducted in the United States. Most addressed the use of clinical decision support systems by nurses in pressure ulcers prevention in inpatient units. All studies described knowledge-based systems that assessed the effects on clinical decision making, clinical effects secondary to clinical decision support system use, or factors that influenced the use or intention to use clinical decision support systems by health professionals and the success of their implementation in nursing practice.

Conclusions: The evidence in the available literature about the effects of clinical decision support systems (used by nurses) on decision making for pressure ulcer prevention and treatment is still insufficient. No significant effects were found on nurses' knowledge following the integration of clinical decision support systems into the workflow, with assessments made for a brief period of up to 6 months. Clinical effects, such as outcomes in the incidence and prevalence of pressure ulcers, remain limited in the studies, and most found clinically but nonstatistically significant results in decreasing pressure ulcers. It is necessary to carry out studies that prioritize better adoption and interaction of nurses with clinical decision support systems, as well as studies with a representative sample of health care professionals, randomized study designs, and application of assessment instruments appropriate to the professional and institutional profile. In addition, long-term follow-up is necessary to assess the effects of clinical decision support systems that can demonstrate a more real, measurable, and significant effect on clinical decision making.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD42019127663; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=127663

(JMIR Med Inform 2020;8(10):e21621) doi:10.2196/21621

KEYWORDS

pressure ulcer; decision support systems, clinical; systematic review

Introduction

Background

A pressure ulcer is an injury resulting from tissue compression and inadequate perfusion to the skin and underlying structures, usually over a bony prominence [1,2]. Pressure ulcer management performed by health care professionals involves phases of prevention, classification, diagnosis, and treatment. The implementation in clinical practice of appropriate strategies for pressure ulcer prevention is indispensable for improving the quality of nursing care.

The clinical decision-making process in pressure ulcer care phases is complex, and its quality depends on both the professional's experience and the availability of accurate knowledge [3]. Decision making should follow evidence-based practices, represented by the management of individualized care for each patient and integrating the use of the best evidence from scientific research [4,5]. The decisions made by nurses should be based on their clinical judgment, with consideration of recommendations in pressure ulcer management guidelines and a view to appropriate clinical practice [1].

Evidence-based guidelines for pressure ulcer prevention and treatment are widely available but are often overlooked or complex to implement in clinical practices. Schaarup et al [6] point out that many randomized controlled trials have concluded that health care professionals are often forced to rely only on their experiences when making wound care decisions because of the low evidence base in studies.

In order to guide professionals in decision making and following recommended guidelines, health information technology that has been incorporated into the clinical workflow, such as clinical decision support systems, may be used. These electronic systems are designed to generate patient-specific assessments or recommendations by comparing characteristics with a knowledge base to directly assist health care professionals in clinical decision making [7]. These systems can be classified into 2 types: (1) knowledge-based clinical decision support systems, expert systems based on inference mechanisms, and (2) nonknowledge-based clinical decision support systems, an inductive system with the application of artificial intelligence (machine learning), such as the use of artificial neural networks [8]. The main methodologies for clinical decision support systems are machine learning, knowledge representation, visualization techniques, and text mining [9].

Knowledge acquisition for these systems is related to the identification and assessment of the best available knowledge [3], making their effectiveness dependent on high-quality clinical research evidence that is up-to-date, easily accessible, and interpretable by computers [4]. The use of clinical decision support systems, in addition to assisting decision makers, can increase the quality of care provided [6,8,10] and reduce errors [8,10,11]. However, there is still limited evidence available on the widespread use of these systems [12], and the quality or relevance of research evidence may restrict their effectiveness [4].

Objective

The purpose of this systematic review was to identify the effects of nurses using clinical decision support systems on clinical decision making for pressure ulcer management. Evaluation of these effects can clarify whether the incorporation of these systems in the workflow improves clinical nursing practice and nurses' knowledge.

Methods

Protocol Registration

This systematic review was conducted in accordance with recommendations by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [13]. A protocol was developed to guide this review and was registered in the International Prospective Register of Systematic Reviews (PROSPERO CRD42019127663) [14].

Search Strategy

The literature search was conducted in April 2019 on 5 electronic databases: MEDLINE/PubMed, Scopus, Cochrane, Web of Science, and CINAHL. The search strategy is reported in detail in Multimedia Appendix 1. Search results were exported and managed in EndNote (Clarivate Analytics). Reference lists of eligible articles were also screened manually for additional studies.

Study Selection

In the first selection phase, studies were screened by assessing titles, abstracts, and keywords, after removing duplicates. The second phase of the full-text review was independently performed by 2 reviewers applying predefined inclusion and exclusion criteria. Eligibility criteria are presented in Textbox 1. The study design of the articles was not limited to high-quality randomized trials to increase the sample of clinical decision support systems publications on pressure ulcers. Qualitative, quantitative, and mixed method studies were included. There was no restriction on the year of publication.

Articles were reviewed by 2 nurses (SA, PS), and using the criteria, those evaluated as appropriate were included. Any disagreement between the reviewers was resolved by consensus or by a third author (ID) through discussion. Cohen κ statistic was calculated to quantify the agreement between reviewers.

Textbox 1. Eligibility criteria.

Inclusion criteria

- described a computer-based clinical decision support systems used by health care professionals for pressure ulcer management
- addressed a clinical decision support systems that generated patient-specific recommendations

Exclusion criteria

- studies that were not written in English
- systems developed to aid teaching only and not to clinical practice
- clinical decision support systems for use on skin lesions or wounds other than pressure ulcers
- clinical decision support systems for use on a smartphone or any other device than the computer
- clinical decision support systems that only generated evaluation results, without specific recommendations
- clinical decision support systems that have not been evaluated or implemented in a real clinical setting

Data Extraction

First author, journal of publication, year, country, study design, aim, pressure ulcer phase (prevention, classification, diagnosis, treatment) for the clinical decision support system application, health care setting involved, participants, type of clinical decision support system and guidelines used, main function of the clinical decision support system, identified evidence, and results of included studies were extracted by one reviewer and confirmed by another.

Clinical Decision Support Systems Classification

Two types were considered in the classification of the clinical decision support systems [8]: knowledge-based (deductive system based on inference engines, usually in the form of *if-then* rules) and non–knowledge based (inductive system with application of artificial intelligence). Another classification used in this review divided the clinical decision support systems into 5 groups, according to their methodologies: machine learning (artificial neural networks, logistic regression, support vector machines), knowledge representation (ontology-based systems, guideline-based, fuzzy logic), information visualization (visualization algorithms to encode abstract concepts and information), text mining (natural language processing and information retrieval), and multipurpose (various attributes and characteristics of existing domains, includes decision trees and Bayesian logic) [8,9].

Study Quality

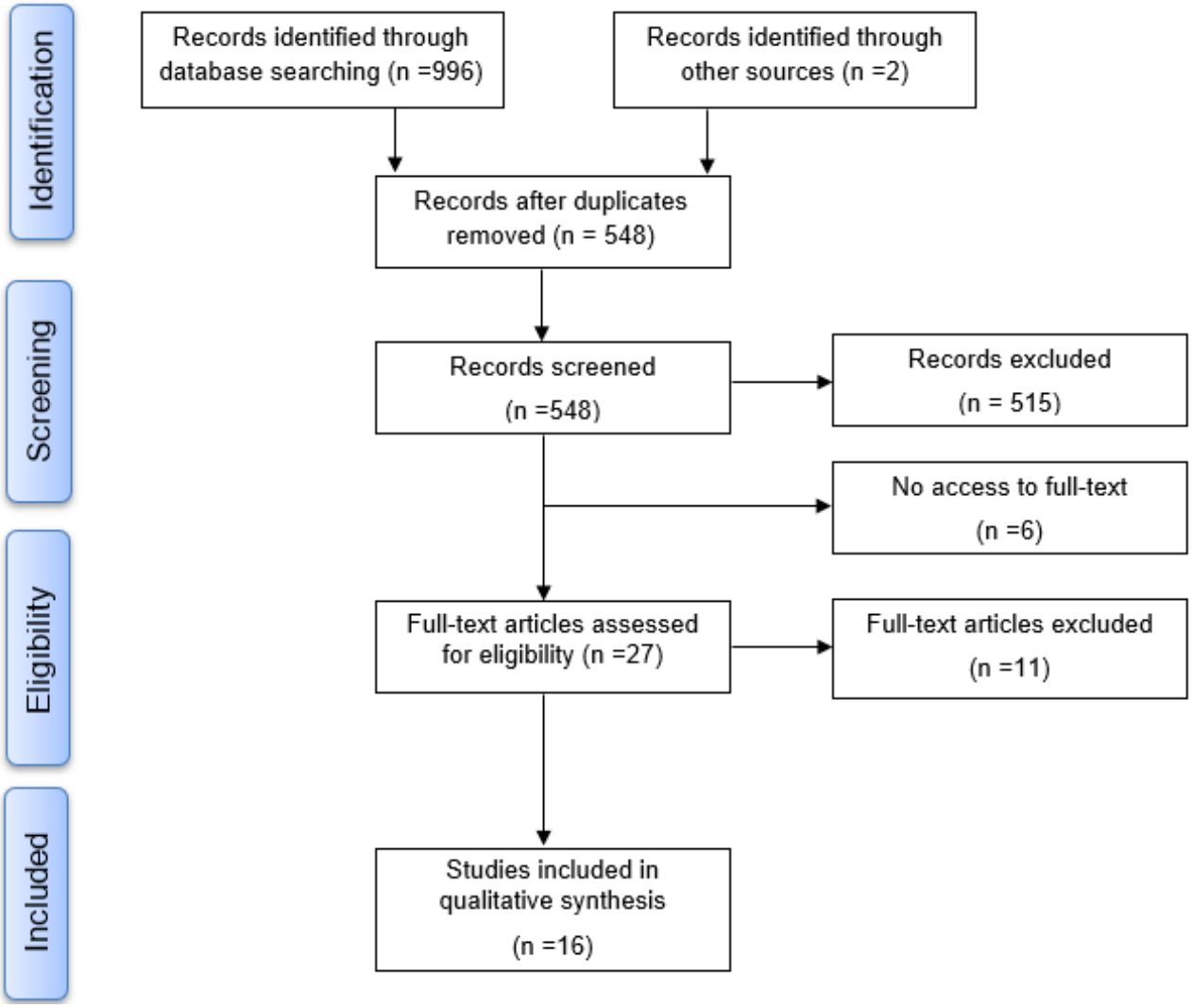
The methodological quality of included studies was assessed using the revised version of the Mixed Methods Appraisal Tool (MMAT) [15]. The MMAT contains a checklist with 5 questions to assess methodological quality for each study design category, defined by MMAT with a number from 1 to 5: (1, qualitative; 2, quantitative randomized controlled trials; 3, quantitative nonrandomized; 4, quantitative descriptive; 5, mixed methods). Each criterion must be answered as “yes,” “no,” or “can't tell.” The studies were analyzed separately and were considered to be of high quality when meeting 100% (5/5) of the criteria, considerable quality with 80% (4/5) of the criteria, moderate quality with 60% (3/5) of the criteria, low quality with 40% (2/5) of the criteria, and very low quality with 20% (1/5) of the criteria.

Results

Search Results

The search strategy yielded a total of 996 articles, and 2 additional articles were identified manually, resulting in 998 articles. After removing duplicates, 548 articles were analyzed in the first phase, in which, 515 articles were excluded; therefore, 33 articles were eligible for the second phase. Access to 6 articles was not possible, and 11 were excluded for different reasons (see Multimedia Appendix 2). Hence, 16 studies [16-31] met all the eligibility criteria and were included in this review. A flow diagram of the selection process is presented in Figure 1.

Figure 1. Flow diagram of the selection process.



Kappa Statistics

When analyzing the selection of the 16 studies included in the qualitative synthesis, the value obtained from the Cohen κ coefficient was 0.67. This value represents substantial strength of agreement between reviewers [32].

Study Quality

In assessing methodological quality using MMAT [15], included studies were classified according to category, and each group was analyzed separately for quality assessment. Methodological quality results are presented in Multimedia Appendix 3. Of the 16 studies, 3 used qualitative research, only 1 was a randomized controlled trial, 8 studies used a nonrandomized quantitative approach, and 2 studies used mixed methods. A total of 5 studies were rated as high-quality, 6 studies were rated as considerable quality, 2 studies were rated as moderate quality, and 1 study was rated as low quality. There were no studies rated as very low quality; however, 2 articles [16,18] did not receive a classification because all 5 criteria analyzed obtained a “can't tell” answer. These 2 studies did not meet any quality criteria in their study category. Both described clinical decision support systems for the care of pressure ulcers but did not describe methodology used for analysis and data collection, which made assessment with MMAT unfeasible. No study was excluded based on quality assessment.

Study Characteristics

General characteristics of included studies are shown in Multimedia Appendix 3. Included studies were conducted between 1995 to 2017 in the following countries: United States of America [16,17,20,21,25-27], Italy [18], Canada [19], Norway [22-24], South Korea [28], Belgium [29], and Singapore [30,31]. The studies were published in 9 different journals and in symposium proceedings, most of which related to health informatics (9/16, 60%), followed by nursing sciences (3/16, 20%). The clinical decision support systems were implemented to support nurses' clinical decisions in multiple clinical and health care settings such as nursing homes [22-24,26,27,29]; hospital inpatient units (medical-surgical) [16-18,20,21,30,31]; acute, home, and extended care [19]; intensive care [28]; and long-term care facilities [25]. The clinical decision support systems were used in pressure ulcer prevention [18,20,21,25-29], prevention and treatment [16,17,19], pressure ulcer prevention and evaluation of nutritional status [22-24], and treatment [30,31]. Interventions in the studies were based on the implementation of clinical decision support systems in clinical

practice with follow-up periods ranging from 1 month [18] to 12 months [19,27] or more [31].

All included studies describe knowledge-based systems—13 out of 16 systems were classified as knowledge representation, with methodologies such as decision rules (*if-then* model) [20,30,31], guideline modeling language (GLIF, Guideline Interchange Format) to validate the logic of enhanced decision rules [21], or clinical practice guidelines represented through the graphic editor GUIDE, written in Java [18]; and 3 out of 16 systems were classified as multipurpose, with 2 using decision trees [30,31] and 1 using a Bayesian network model [28].

In 7 out of 13 systems classified as knowledge representation, the clinical decision support systems were developed based on Agency for Healthcare Research and Quality guidelines for pressure ulcer prevention and treatment [16-19,25-27]. The Braden Scale [16,19-21] and the Risk Assessment Pressure Scale [22-24], both for pressure ulcer risk screening, also appear as evidence bases. The Pressure Sore Status Tool [19], an instrument for pressure ulcer evaluation; the American Medical Directors Association guidelines for pressure ulcer prevention [25,26]; and opinions of pressure ulcer experts on the decision-making rules of the clinical decision support systems [16,18-21,29] were other knowledge described in the articles. In addition, literature reviews to identify the best evidence for pressure ulcer care were also used to create the systems [16,19,23,28,29,31]. The classification, evidence base, and function of the clinical decision support systems are detailed in Multimedia Appendix 4.

Clinical Decision Support Systems in Analysis

Effects on Nurses' Clinical Decision Making

Few studies evaluated the effects on nurses' decision making. Nurses acknowledged advantages after a month of testing the implementation of a computerized guideline for pressure ulcer prevention in a general medicine ward; users reported that the daily prevention work-plans generated by the clinical decision support systems and the detailed storage of actions were useful in making decisions for planning patient discharge [18].

On the other hand, nurses at a public tertiary hospital in Singapore reported low credibility and confidence in the implemented clinical decision support systems [30]. This assessment, influenced by the workplace culture, had consequences for the adoption of the system and for nurses' decision making. Instead of what was recommended by the clinical decision support system, many nurses preferred to follow their past experiences or opinions of leaders and wound experts when determining the treatment modalities for the wound [30]. The same was observed in the study by Clarke et al [19] in which some nurses perceived the care plans generated by the clinical decision support systems as elementary, preferring to trust on their own assessment skills.

Regarding the knowledge acquired by professionals after the implementation of clinical decision support systems, which could have a positive effect on decision-making skills in the care of pressure ulcers, the results were paradoxical. Clarke et al [19] observed an increase in knowledge about pressure ulcers prevention, treatment strategies, resources required, and the importance of interdisciplinary teams in the daily planning of interventions. However, in the studies by Zielstorff et al [17] and Beeckman et al [29], the results showed no significant improvement in nurses' knowledge about pressure ulcer prevention and treatment, when comparing the knowledge assessment instrument results applied to health care professionals in the intervention and control groups, before and after the implementation of clinical decision support systems.

Factors That Influence the Use or Intention to Use and Successful Implementation in Clinical Practice

Nurses had favorable attitudes toward use when a clinical decision support system [28] was implemented in an intensive care unit using data from the electronic health record to predict hospital-acquired pressure ulcers. In nursing homes, some nursing personnel who were comfortable with computer technology evaluated the use of clinical decision support systems with positive feedback, while others expressed resistance to use [23]. In the studies, various reasons that influenced nurses' adoption of the systems to support clinical decision making in pressure ulcer care were observed. Professional, organizational, and software-design barriers affected the use of clinical decision support systems by nurses. The main advantages and difficulties of using the clinical decision support systems that were assessed by users are presented in Textbox 2.

Textbox 2. Advantages and difficulties assessed by users in using clinical decision support systems to care for pressure ulcers.

Advantages

- Easy to use [17,18,23,30]
- Detailed documentation [18,19,24,25]
- Improved planning [18,19,25]
- Workload assessment [18]
- Useful at the patient discharge [18]
- Education [18]
- Facilitates handing on duties to the next shift nurses [18]
- Implementing and following the protocols [16,29]
- Improved the recording of nursing assessments and comprehensiveness [24,25]

Difficulties

- Lack of flexibility [18,23]
- Lack of logical flow [23,30]
- Lack of time for data input [17,18,23]
- Lack of computer skills [19,23]
- Lack of training [19,23]
- Lack of computer infrastructure [19]
- Lack of information about the clinical decision support systems implementation [23]
- Resistance to use computers [23]
- Workplace culture [30]
- Lack of trust and credibility in clinical decision support systems [30]
- Frustration with clinical decision support systems use [19,30]

The factors associated with successful clinical decision support system implementation in clinical practice were involvement of the administrator or head of nursing in the process [25,26], emphasizing the importance of leadership that was actively engaged; the presence of an internal champion [26] as a key nurse [29], who can be a persuasive leader as the force for change; and participation of an interdisciplinary team, facilitators, and a quality improvement team [25,26,29] in the health care organization. In addition, consideration of clinical workflow [18,31], training and previous education activities for professionals on the use of clinical decision support systems [19,22-25,28,29] and the importance of preventing pressure ulcers [28,29] performed before implantation of the clinical decision support systems were also described in the articles as factors associated with success.

Clinical Effects on Pressure Ulcer Incidence and Prevalence

Preliminary results in one study [16], indicated a significant reduction, from 7% to 2%, in pressure ulcer incidence in the case units, 6 months postimplementation of a clinical decision support system for pressure ulcer prevention in an American hospital. In the study by Olsho et al [27], this clinical effect occurred in nursing homes that jointly implemented 4 components (nutrition, weight summary, priority, trigger summary), avoiding approximately 2.6 pressure ulcers per 100 patients per month (P=.035).

In 7 long-term institutions that implemented a clinical decision support system [25], there was a decrease in the percentage of high-risk residents with pressure ulcers from 13.0% (before implementation) to 8.7% (12 months after implementation), with a combined reduction of 33%. However, quality control decreased in 5 facilities and increased slightly in 2 facilities that did not implement all the system reports.

In the intervention group of an intensive care unit, adoption to the clinical decision support systems [28] for pressure ulcer prevention allowed a 21% to 4% reduction in the prevalence of hospital-acquired pressure ulcer and decreased the length of stay by approximately one-third (7.6 to 5.2 days). Beeckman et al [29] also observed a decrease in the prevalence of pressure ulcers after using a clinical decision support systems in the experimental group. The result was clinically meaningful but nonstatistically significant. Therefore, no overall significant effect was found on pressure ulcer prevalence [29].

Discussion

Principal Results

As for the impact on nurses' knowledge with the use of clinical decision support systems, only 3 included studies evaluated this

effect and obtained paradoxical results. There was no description of the time of data collection to assess knowledge, nor of the type of assessment used, in the study [19] that identified an increase in nurses' knowledge after the intervention. In studies in which this effect was not identified, few nurses participated in the posttest [17], and there were limitations in the knowledge questionnaire applied before and after the clinical decision support systems implementation [29]. The assessment instrument for nurses was used with health care professionals who had no nursing education background and may have been too difficult, resulting in low scores on the instrument [29].

Evidence of the effect of clinical decision support systems on clinical knowledge is still insufficient, with evaluations carried out after short periods of system implementation that may not demonstrate measurable effects [17] as well as with small sizes in the assessed sample.

As for the factors that influenced the use or intention to use clinical decision support systems and the success of implementations in included articles, the professionals played important roles in the process. Several professional and organizational barriers were identified in the adoption of the clinical decision support systems, as well as in nurses' relationships with the use of the systems. Relying on their own assessments, instead of the recommendations generated by the clinical decision support systems, was an observation found only in studies that analyzed the use of systems in pressure ulcer treatment.

Gerrish et al [33] reported that nurses rely heavily on communication with colleagues and their personal experience rather than formal sources of knowledge. Dowding et al [34], also described that nurses report relying on their experience when dealing with tasks in which decisions seemed more familiar and using the clinical decision support for situations with which they had little experience.

The interaction between the nurse and the technology must be considered by involving end users during all stages of the implementation and in evaluations of the system [34,35]. The user's computer knowledge and training on the clinical decision support systems also directly affected the adoption of the systems. Ammenwerth et al [36] identified that a professional's computer knowledge and previous acceptance of the nursing process were 2 factors that were significant predictors of user acceptance of computerized nursing systems. The other factors observed were the fit between the nursing workflow and the functionality of the system [36].

An important basis for clinical decision support system design is an understanding of the clinical care process and local workflow. Decision support can be provided continuously throughout the care process, at the most effective level of nursing care (from the user's initial assessment to the outcome evaluation) [37]. The use of clinical decision support systems allowed increased compliance with pressure ulcer prevention protocols, improving professional attitudes, in addition to encouraging more complete documentation and more comprehensive nursing assessments [24,25]. The other benefits included consistency in the quality of nursing care and greater access to information on best practices [38].

Clinical decision support system implementation must be based on models of technology adoption, evidence-based practices, and conceptual models in nursing practice. The success of clinical decision support system implementation will clearly depend on the analysis of critical success factors, and modeling efforts should allow for the broadest and most effective use of the systems [39]. Only 4 studies [19,23,28,29] addressed the use of some model or conceptual framework as a guide, organizing implementation strategies and elucidating the variables found.

Clarke et al [19] used 5 phases of the adoption of innovation [40] and 5 factors influencing the rate of adoption of innovations [41] models; Fossum et al [23] applied the Task Technology Fit model [42]; to measure the user's attitude toward the system, Cho et al [28] used the United Theory of Acceptance and Use of Technology [43] model questionnaire; and Beeckman et al [29] used a model for effective implementation [44].

To trigger improvement in nursing practice, it is important that clinical decision support systems have following characteristics: automatic provision of decision support, facilitating clinical practice and decreasing the professional's effort; provision of recommendations, rather than just evaluations; and provision of decision support at the time and location of clinical decision making [45,46]. According to Kawamoto et al [45], nursing practices improved significantly in 94% of the analyzed trials when all these characteristics were present in the clinical decision support system.

Automatic prompting in clinical decision support systems can improve integration into the workflow and provide the opportunity to correct inadvertent deficiencies in care [47]. The decision support system [16] that used an alert logic had a positive impact in reminding nurses about the completion of each patient's processes. Only 6 out of 50 admissions were completed on the system without prompting alerts. The availability of this tool in clinical decision support systems affects the performance of professionals [10,47]. However, these reminders should be relevant to the patient's profile so that the user does not reject them [10]; interfaces with many alerts can generate frustration when using the clinical decision support systems, decreasing workflow, quality, efficiency, and safety in providing patient care [10].

As for the clinical effects from using a clinical decision support systems, the reduction in the pressure ulcer incidence was considered to be of low evidence. One of the studies [16] with this finding did not meet any MMAT quality criteria in its study category. In the other [27], the analysis was subject to several important limitations, and there was an imprecision associated with the estimate when the 95% confidence interval was applied.

In reducing pressure ulcer prevalence, there was a possible bias in the study by Cho et al [28] from the long time elapsed between the intervention and the observation, which may have positively influenced the results of both the reduction of pressure ulcers and the length of intensive care unit stay [48]. The study by Fossum et al [22] showed no effect on patient outcomes in relation to pressure ulcer risk and prevalence. However, all the groups that were evaluated had smaller samples than those recommended by power analysis calculations. The positive

clinical effects shown in the included studies were mostly clinically significant but without statistical significance.

Assessing and interpreting the clinical effects generated by the clinical decision support system intervention, as well as obtaining results with strong evidence in clinical practice, can be a difficult task. This can happen because clinical decision support systems are knowledge-based, using, for example, expert opinions and prevention scales when creating the algorithms. There is still no strong evidence that the risk of developing pressure ulcer decreases with the use of pressure ulcer risk assessment instruments (such as the Braden scale) when compared to less standardized risk assessment based on nurses' clinical judgment [49].

Thus, if the evidence from the system's knowledge base has scientific limitations, the clinical effects generated by clinical decision support system may also be limited. There is also a difficulty in identifying, in the widely available literature, the best knowledge to be used to create this type of system [3]. In this way, clinical decision support systems will only be able to facilitate the implementation of evidence-based care when the systems can follow the literature in identifying high-quality studies and incorporate the best evidence to generate more appropriate recommendations [4].

Limitations

This systematic review was limited by the eligibility of heterogeneous studies, publication bias, location bias, and nonconducted meta-analysis. There was a plurality of methodological approaches, not limited to randomized controlled trials. However, this is often a necessary approach to expand the understanding of clinical acceptance influenced by clinical decision support system development and deployment [50].

In addition, most of the studies evaluated were not randomized, with an inherent risk of bias. However, the quasi-experimental design is often used in many medical informatics articles to evaluate the benefits of specific interventions when it is not logistically feasible or ethical to conduct a randomized controlled trial [51]. Finally, the analysis of the results was limited, with some included studies that published only preliminary results [16-19].

Directions for Future Studies

Effects of clinical decision support systems used by nurses in the management of pressure ulcers lack results of strong evidence in the literature. It is necessary to carry out studies that prioritize better adoption and interaction of nurses with these systems by making this the focus during the development of clinical decision support systems and in planning implementation strategies, as well as having studies with representative samples of health care professionals, randomized designs, and the application of assessment instruments appropriate to the professional profile and consistent with the health care organization. Longer periods should be used for the evaluation of the effects of the clinical decision support systems, which may have a more real, measurable, and significant effect on clinical decision making. In addition, these studies should be accompanied by the creation and implementation of systems based on recommendations and successful models, for better adoption by nurses to clinical decision support systems in the pressure ulcers treatment.

Conclusions

Evidence in the available literature is still insufficient regarding the effects of nurses who use clinical decision support systems on clinical decision making for pressure ulcer prevention or treatment. No significant effects were found on nurses' knowledge following the integration of clinical decision support systems into workflows, with assessments made for a brief period of up to 6 months of implementation. Clinical effects, such as outcomes in the incidence and prevalence of pressure ulcers, remain limited, and most were clinically significant but nonstatistically significant.

Acknowledgments

This article was supported by National Funds through *Fundação para a Ciência e a Tecnologia*, within the Center for Health Technology and Services Research (CINTESIS), Research and Development Unit (reference UIDB/4255/2020).

Authors' Contributions

SA completed the title and abstract search, interpretation of results, and writing of the manuscript. PS supervised the project. Both SA and PS completed the review of full-text papers and data extraction. ID contributed to the analysis of inclusion of articles when there was no consensus. All authors contributed to the final version of the manuscript.

Conflicts of Interest

PS currently serves as a JMIR Medical Education reviewer.

Multimedia Appendix 1
The search strategy.
[PDF File (Adobe PDF File), 31 KB - medinform_v8i10e21621_app1.pdf]

Multimedia Appendix 2
Reasons for records exclusion in the screening and eligibility phase.

[PDF File (Adobe PDF File), 40 KB - medinform_v8i10e21621_app2.pdf]

Multimedia Appendix 3
General characteristics of the included studies.
[PDF File (Adobe PDF File), 134 KB - medinform_v8i10e21621_app3.pdf]

Multimedia Appendix 4
Characteristics of the clinical decision support systems described in the included studies.
[PDF File (Adobe PDF File), 81 KB - medinform_v8i10e21621_app4.pdf]

References

1. National Pressure Ulcer Advisory Panel, European Pressure Ulcer Advisory Panel, Pan Pacific Pressure Injury Alliance. In: Haesler E, editor. Prevention and Treatment of Pressure Ulcers: Quick Reference Guide. Osborne Park, Australia: Cambridge Media; 2014.
2. Garcia T. Classificação internacional para prática da enfermagem (CIPE): versão. Porto Alegre: Artmed; 2018.
3. Zolhavarieh S, Parry D, Bai Q. Issues associated with the use of semantic web technology in knowledge acquisition for clinical decision support systems: systematic review of the literature. JMIR Med Inform 2017 Jul 05;5(3):e18 [FREE Full text] [doi: 10.2196/medinform.6169] [Medline: 28679487]
4. Sim I, Gorman P, Greenes RA, Haynes RB, Kaplan B, Lehmann H, et al. Clinical decision support systems for the practice of evidence-based medicine. J Am Med Inform Assoc 2001;8(6):527-534 [FREE Full text] [doi: 10.1136/jamia.2001.0080527] [Medline: 11687560]
5. DiCenso A, Cullum N, Ciliska D. Implementing evidence-based nursing: some misconceptions. Evidence-Based Nursing 1998 Apr 01;1(2):38-39. [doi: 10.1136/ebn.1.2.38]
6. Schaarup C, Pape-Haugaard LB, Hejlesen OK. Models used in clinical decision support systems supporting health care professionals treating chronic wounds: systematic literature review. JMIR Diabetes 2018 Jun 21;3(2):e11 [FREE Full text] [doi: 10.2196/diabetes.8316] [Medline: 30291078]
7. Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA 1998 Oct 21;280(15):1339-1346. [doi: 10.1001/jama.280.15.1339] [Medline: 9794315]
8. Shahsavarani A, Abadi E. Kalkhoran MH Clinical decision support systems (CDSSs): state of the art review of literature. Int J Med Rev 2015;2(4):299-308 [FREE Full text]
9. Fraccaro P, O Sullivan D, Plastiras P, O Sullivan H, Dentone C, Di Biagio A, et al. Behind the screens: Clinical decision support methodologies – a review. Health Policy and Technology 2015 Mar;4(1):29-38. [doi: 10.1016/j.hlpt.2014.10.001]
10. Bates DW, Cohen M, Leape LL, Overhage JM, Shabot MM, Sheridan T. Reducing the frequency of errors in medicine using information technology. J Am Med Inform Assoc 2001;8(4):299-308 [FREE Full text] [doi: 10.1136/jamia.2001.0080299] [Medline: 11418536]
11. Castillo RS, Kelemen A. Considerations for a successful clinical decision support system. Comput Inform Nurs 2013 Jul;31(7):319-326. [doi: 10.1097/NXN.0b013e3182997a9c] [Medline: 23774450]
12. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med 2012 Jul 03;157(1):29-43. [doi: 10.7326/0003-4819-157-1-201207030-00450] [Medline: 22751758]
13. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009 Jul 21;6(7):e1000100 [FREE Full text] [doi: 10.1371/journal.pmed.1000100] [Medline: 19621070]
14. Araujo S, Sousa P, Dutra I. Clinical decision support systems for pressure ulcer management: a systematic review (protocol). PROSPERO CRD4127663. 2019. URL: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42019127663 [accessed 2019-04-10]
15. Hong Q, Pluye P. Fàbregues S Mixed methods appraisal tool (MMAT), version 2018. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada. URL: http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/127916259/MMAT_2018_criteria-manual_2018-08-01_ENG.pdf [accessed 2019-05-15]
16. Willson D, Ashton C, Wingate N, Goff C, Horn S, Davies M, et al. Computerized support of pressure ulcer prevention and treatment protocols. Proc Annu Symp Comput Appl Med Care 1995:646-650 [FREE Full text] [Medline: 8563366]
17. Zielstorff RD, Estey G, Vickery A, Hamilton G, Fitzmaurice JB, Barnett GO. Evaluation of a decision support system for pressure ulcer prevention and management: preliminary findings. Proc AMIA Annu Fall Symp 1997:248-252 [FREE Full text] [Medline: 9357626]
18. Quaglini S, Grandi M, Baiardi P, Mazzoleni MC, Fassino C, Franchi G, et al. A computerized guideline for pressure ulcer prevention. Int J Med Inform 2000 Sep;58-59:207-217. [doi: 10.1016/s1386-5056(00)00088-5] [Medline: 10978922]

19. Clarke HF, Bradley C, Whytock S, Handfield S, van der Wal R, Gundry S. Pressure ulcers: implementation of evidence-based nursing practice. J Adv Nurs 2005 Mar;49(6):578-590. [doi: 10.1111/j.1365-2648.2004.03333.x] [Medline: 15737218]
20. Kim H, Choi J, Thompson S, Meeker L, Dykes P, Goldsmith D, et al. Automating pressure ulcer risk assessment using documented patient data. Int J Med Inform 2010 Dec;79(12):840-848. [doi: 10.1016/j.ijmedinf.2010.08.005] [Medline: 20869303]
21. Choi J, Kim H. Enhancement of decision rules to increase generalizability and performance of the rule-based system assessing risk for pressure ulcer. Appl Clin Inform 2013;4(2):251-266 [FREE Full text] [doi: 10.4338/ACI-2012-12-RA-0056] [Medline: 23874362]
22. Fossum M, Alexander GL, Ehnfors M, Ehrenberg A. Effects of a computerized decision support system on pressure ulcers and malnutrition in nursing homes for the elderly. Int J Med Inform 2011 Sep;80(9):607-617. [doi: 10.1016/j.ijmedinf.2011.06.009] [Medline: 21783409]
23. Fossum M, Ehnfors M, Fruhling A, Ehrenberg A. An evaluation of the usability of a computerized decision support system for nursing homes. Appl Clin Inform 2011;2(4):420-436 [FREE Full text] [doi: 10.4338/ACI-2011-07-RA-0043] [Medline: 23616886]
24. Fossum M, Ehnfors M, Svensson E, Hansen LM, Ehrenberg A. Effects of a computerized decision support system on care planning for pressure ulcers and malnutrition in nursing homes: an intervention study. Int J Med Inform 2013 Oct;82(10):911-921. [doi: 10.1016/j.ijmedinf.2013.05.009] [Medline: 23827767]
25. Horn SD, Sharkey SS, Hudak S, Gassaway J, James R, Spector W. Pressure ulcer prevention in long-term-care facilities: a pilot study implementing standardized nurse aide documentation and feedback reports. Adv Skin Wound Care 2010 Mar;23(3):120-131. [doi: 10.1097/01.ASW.0000363516.47512.67] [Medline: 20177165]
26. Sharkey S, Hudak S, Horn SD, Barrett R, Spector W, Limcangco R. Exploratory study of nursing home factors associated with successful implementation of clinical decision support tools for pressure ulcer prevention. Adv Skin Wound Care 2013 Feb;26(2):83-92. [doi: 10.1097/01.ASW.0000426718.59326.bb] [Medline: 23337649]
27. Olsho LEW, Spector WD, Williams CS, Rhodes W, Fink RV, Limcangco R, et al. Evaluation of AHRQ's on-time pressure ulcer prevention program: a facilitator-assisted clinical decision support intervention for nursing homes. Med Care 2014 Mar;52(3):258-266. [doi: 10.1097/MLR.0000000000000080] [Medline: 24374408]
28. Cho I, Park I, Kim E, Lee E, Bates DW. Using EHR data to predict hospital-acquired pressure ulcers: a prospective study of a Bayesian Network model. Int J Med Inform 2013 Nov;82(11):1059-1067. [doi: 10.1016/j.ijmedinf.2013.06.012] [Medline: 23891086]
29. Beeckman D, Clays E, Van Hecke A, Vanderwee K, Schoonhoven L, Verhaeghe S. A multi-faceted tailored strategy to implement an electronic clinical decision support system for pressure ulcer prevention in nursing homes: a two-armed randomized controlled trial. Int J Nurs Stud 2013 Apr;50(4):475-486. [doi: 10.1016/j.ijnurstu.2012.09.007] [Medline: 23036149]
30. Khong PCB, Hoi SY, Holroyd E, Wang W. Nurses' clinical decision making on adopting a wound clinical decision support system. Comput Inform Nurs 2015 Jul;33(7):295-305. [doi: 10.1097/CIN.0000000000000164] [Medline: 26066306]
31. Khong P, Lee L, Dawang A. Modeling the construct of an expert evidence-adaptive knowledge base for a pressure injury clinical decision support system. Informatics 2017 Jul 12;4(3):20. [doi: 10.3390/informatics4030020]
32. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
33. Gerrish K, Ashworth P, Lacey A, Bailey J. Developing evidence-based practice: experiences of senior and junior clinical nurses. J Adv Nurs 2008 Apr;62(1):62-73. [doi: 10.1111/j.1365-2648.2007.04579.x] [Medline: 18352965]
34. Dowding D, Mitchell N, Randell R, Foster R, Lattimer V, Thompson C. Nurses' use of computerised clinical decision support systems: a case site analysis. J Clin Nurs 2009 Apr;18(8):1159-1167. [doi: 10.1111/j.1365-2702.2008.02607.x] [Medline: 19320785]
35. Randell R, Dowding D. Organisational influences on nurses' use of clinical decision support systems. Int J Med Inform 2010 Jun;79(6):412-421. [doi: 10.1016/j.ijmedinf.2010.02.003] [Medline: 20233670]
36. Ammenwerth E, Mansmann U, Iller C, Eichstädter R. Factors affecting and affected by user acceptance of computer-based nursing documentation: results of a two-year study. J Am Med Inform Assoc 2003;10(1):69-84 [FREE Full text] [doi: 10.1197/jamia.m1118] [Medline: 12509358]
37. Lee S. Features of computerized clinical decision support systems supportive of nursing practice: a literature review. Comput Inform Nurs 2013 Oct;31(10):477-495. [doi: 10.1097/01.NCN.0000432127.99644.25] [Medline: 23958964]
38. Anderson JA, Willson P. Clinical decision support systems in nursing: synthesis of the science for evidence-based practice. Comput Inform Nurs 2008;26(3):151-158. [doi: 10.1097/01.NCN.0000304783.72811.8e] [Medline: 18438151]
39. Greenes RA, Bates DW, Kawamoto K, Middleton B, Osheroff J, Shahar Y. Clinical decision support models and frameworks: Seeking to address research issues underlying implementation successes and failures. J Biomed Inform 2018 Dec;78:134-143. [doi: 10.1016/j.jbi.2017.12.005] [Medline: 29246790]
40. Rogers E. Diffusion of Innovations 4th ed. New York: The Free Press; 1995.
41. Romano CA. Diffusion of technology innovation. ANS Adv Nurs Sci 1990 Dec;13(2):11-21. [doi: 10.1097/00012272-199012000-00003] [Medline: 2124787]

42. Goodhue DL, Thompson RL. Task-technology fit and individual performance. MIS Quarterly 1995 Jun;19(2):213. [doi: 10.2307/249689]
43. Venkatesh, Morris, Davis, Davis. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
44. Grol R, Wensing M. Effective implementation: a model. In: Improving Patient Care: The Implementation of Change in Clinical Practice. London: Elsevier; 2005.
45. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005 Apr 02;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
46. Randell R, Mitchell N, Dowding D, Cullum N, Thompson C. Effects of computerized decision support systems on nursing performance and patient outcomes: a systematic review. J Health Serv Res Policy 2007 Oct;12(4):242-249. [doi: 10.1258/135581907782101543] [Medline: 17925077]
47. 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]
48. Jeffery AD. Methodological Challenges in Examining the Impact of Healthcare Predictive Analytics on Nursing-Sensitive Patient Outcomes. Comput Inform Nurs 2015 Jun;33(6):258-264. [doi: 10.1097/CIN.0000000000000154] [Medline: 25899442]
49. Chou R, Dana T, Bougatsos C, Blazina I, Starmer AJ, Reitel K, et al. Pressure ulcer risk assessment and prevention: a systematic comparative effectiveness review. Ann Intern Med 2013 Jul 02;159(1):28-38. [doi: 10.7326/0003-4819-159-1-201307020-00006] [Medline: 23817702]
50. Kaplan B. Evaluating informatics applications--some alternative approaches: theory, social interactionism, and call for methodological pluralism. Int J Med Inform 2001 Nov;64(1):39-56. [Medline: 11673101]
51. Harris AD, McGregor JC, Perencevich EN, Furuno JP, Zhu J, Peterson DE, et al. The use and interpretation of quasi-experimental studies in medical informatics. J Am Med Inform Assoc 2006;13(1):16-23. [doi: 10.1197/jamia.M1749] [Medline: 16221933]

Abbreviations

MMAT: Mixed Methods Appraisal Tool

Edited by C Lovis; submitted 19.06.20; peer-reviewed by M Padilha, D Parry; comments to author 19.08.20; revised version received 27.08.20; accepted 06.09.20; published 16.10.20.

Please cite as:
Araujo SM, Sousa P, Dutra I
Clinical Decision Support Systems for Pressure Ulcer Management: Systematic Review
JMIR Med Inform 2020;8(10):e21621
URL: http://medinform.jmir.org/2020/10/e21621/
doi:10.2196/21621
PMID:33064099

Viewpoint

Clinical Decision Support May Link Multiple Domains to Improve Patient Care: Viewpoint

David Kao[1], MD; Cynthia Larson[2], BSN; Dana Fletcher[3], MA; Kris Stegner[4], PhD, MBA

[1]Department of Cardiology, University of Colorado School of Medicine, Aurora, CO, United States

[2]UCHealth, Aurora, CO, United States

[3]Evida Clinical Consulting, Inc, Golden, CO, United States

[4]G(x)P Advisors, Inc, Thornton, CO, United States

Corresponding Author:
David Kao, MD
Department of Cardiology
University of Colorado School of Medicine
12700 East 19th Avenue
Aurora, CO, 80045
United States
Phone: 1 720 848 5300
Email: david.kao@cuanschutz.edu

Abstract

Integrating clinical decision support (CDS) across the continuum of population-, encounter-, and precision-level care domains may improve hospital and clinic workflow efficiency. Due to the diversity and volume of electronic health record data, complexity of medical and operational knowledge, and specifics of target user workflows, the development and implementation of comprehensive CDS is challenging. Additionally, many providers have an incomplete understanding of the full capabilities of current CDS to potentially improve the quality and efficiency of care delivery. These varied requirements necessitate a multidisciplinary team approach to CDS development for successful integration. Here, we present a practical overview of current and evolving applications of CDS approaches in a large academic setting and discuss the successes and challenges. We demonstrate that implementing CDS tools in the context of linked population-, encounter-, and precision-level care provides an opportunity to integrate complex algorithms at each level into a unified mechanism to improve patient management.

(JMIR Med Inform 2020;8(10):e20265) doi:10.2196/20265

KEYWORDS

clinical decision support; population medicine; evidence-based medicine; precision medicine; care management; electronic health records

Introduction

The US health care sector has marched steadily toward adoption and standardization of health information technology systems over the past decade with much anticipation of their potential [1]. Clinicians face the increasing challenge of incorporating new guidelines into clinical practice as the volume and complexity of electronic health data continue to grow [2-7]. However, evidence of improvements in the quality, safety, and efficiency of patient care stemming from electronic medical records (EMRs) is mixed [3,5,8].

Clinical decision support (CDS) systems are broadly defined as information and tools used in patient care, such as reminders, alerts, and guidelines [9]. Automated CDS is developed to assist clinician decision-making and improve patient care by leveraging the breadth of electronic health data combined with up-to-date practice recommendations in the context of local workflow requirements. Ideally, CDS applications will seamlessly translate enhanced decision-making into action to optimize health care delivery [10]. For maximal impact, information and processes are delivered to providers on an ongoing basis; ideally, they are fully integrated with institutional workflows [11,12].

The increasing investment in the combination of traditional evidence-based and precision medicine also requires innovation in CDS approaches [13]. Incorporating ‘omic medicine and collectively characterizing and measuring molecular data from fields including molecular diagnostics, environmental exposures, and lifestyle behaviors will require considerable assistance from EMRs, CDS tools, and other expert systems given the scope

and complexity of the data involved. Given that contemporary EMRs are not equipped to capture and access 'omic "big data," one important function of CDS will be to access and use data from multiple disparate sources to generate recommendations ranging from discrete decisions such as choosing a medication to long-term chronic disease prevision. Additionally, future CDS applications should ideally operate using consistent, portable CDS knowledge bases to facilitate shared implementation and querying strategies between institutions that leverage data along the spectrum from highly individualized 'omics to population-level evidence-based medicine without requiring multiple, possibly inconsistent implementations at different institutions.

CDS is uniquely positioned to support population-, encounter-, and precision-level medicine as a continuum of care delivery through EMR and clinical informatics systems. Implementing CDS tools across the spectrum of these three clinical care domains may potentially improve efficiency and quality of care for patients. Medicare bundled payments and other pay-for-performance models incentivize efficient and consistent care transitions from the emergency department (ED) to inpatient settings to outpatient settings [14]. Therefore, new CDS solutions should reflect the reality of integrated care delivery. However, significant effort is required in planning and development to ensure that CDS applications align with end-user workflows to maximize efficiency and provider uptake. In addition, effective provider education and change management are critical for CDS implementation. Here, we summarize and provide examples of the implementation of CDS tools within each clinical care domain as well as across all three domains and the challenges that were encountered during this implementation at a large medical center.

Overview of the Three CDS Health Care Domains

Population-Level CDS

The goal of population health is to improve long-term outcomes of patient cohorts by means of preventive interventions, patient engagement, care coordination, and other activities outside clinical visits. Population-level CDS requires integration with very different workflows that are unique to a broad range of providers, such as care managers, social workers, or patient outreach staff, all independent of discrete, face-to-face encounters. Potential population health CDS applications may include identifying target patients (eg, patient registries) and monitoring of long-term treatment profiles (eg, guideline adherence dashboards), clinical outcomes (eg, urgent clinic visits or hospitalization), and care transitions (eg, post-hospitalization follow-up). By leveraging data as it enters the EMR irrespective of timing, CDS can support many providers in surveilling and delivering integrated patient- and disease-focused care to target patient populations. The range of population health CDS is currently limited in part by a lack of clear and consistent workflows. As these standards evolve, CDS will likely become critical for successful population health management given the challenges of the unpredictable timing of important clinical events, complexity of data, and novel interpretation methods that will be required.

Encounter-Level CDS

Encounter-level CDS is the most common and familiar type of CDS, and the most evidence has been obtained to date regarding its efficacy. Encounter-level applications generally provide important information, give reminders, or suggest a course of care during a discrete clinical interaction such as a clinic visit, a hospitalization, an elective procedure, or even a telephone call. This model allows providers to take immediate action that affects patient care, thereby providing the right information to the right person at the right time through the EMR, which is the right channel to enact a recommendation. Encounter-level CDS applications are often high-value use cases because practice guidelines, performance metrics, and safety measures often have important implications for patient outcomes, reimbursement, or public reporting of performance; also, direct suggestions within a clinical workflow (eg, when an order is placed) can change a provider's action in a timely manner. Examples of encounter-level CDS include alerts regarding drug-drug or drug-allergy interactions, risk-based vital sign monitoring recommendations, or reminders in computerized provider order entry systems to place orders for tests or medications [8,15,16]. In addition, encounter-level CDS systems are increasingly able to support complex, interactive applications to standardize care delivery for specific clinical scenarios in accordance with evidence-based recommendations.

Precision-Level CDS

Precision-level CDS integrates complex, voluminous, and disparate data regarding a patient's specific characteristics in multiple domains to provide clinical guidance [17]. Although the term "precision medicine" is often associated with genetic- or genomic-guided medical therapy, it applies to any individualized management strategy based on a patient's unique combination of traits, such as demographics, clinical and family history, physical traits (eg, weight, blood tests), activity, mental health, socioeconomic status, and environment, among many others [18]. Telemonitoring and remote care delivery education are also considered to be precision medicine tools in chronic disease management that are intended to optimize access to care and empower patients to manage their own health [19,20]. These tools facilitate communication between the patient and the provider about the patient's individualized risks and needs; they also formulate beneficial and achievable treatment plans and provide personalized education.

Applications of CDS Domains in Health Care

At our institution, we have used dozens of frameworks to implement hundreds of applications representing the three CDS domains outlined above. Herein, we present the successes and challenges we observed with the application of these frameworks.

Population-Level CDS

Our medical center is expanding its population health care management services to evolve with the growing focus of payers on management of high-risk patient groups and on performance metrics regarding integrated care delivery. This is being accomplished in part by the creation of EMR patient registries for chronic diseases such as diabetes, chronic obstructive pulmonary disease (COPD), and heart failure. Using regional health information exchanges, the statewide all-payer claims database, and our institutional clinical data warehouse, these registries integrate comprehensive data regarding clinical encounters, management changes, medication use, and outcomes both within and outside of our health care system for patients in each registry. We are developing CDS alerts based on these diverse data and on patient assessments and previous intervention outcomes to trigger tasks and education goals for the patient. For example, patients in a diabetes registry may have visit documentation and lab results stored in the EMR. If a patient's routine assessments include elevated hemoglobin A1C and poor familiarity with home monitoring equipment, the care manager receives a notification recommending goal-setting with the patient (eg, for home monitoring and stabilizing the patient's hemoglobin A1C) and adding tasks (eg, an order for a blood glucose monitor, future lab tests, or patient education). Although the role of CDS in the population health domain is growing, challenges remain due to the lack of established standards and provider workflows for care management, resulting in inconsistent delivery of notifications to the right people or at the right time and in the right format.

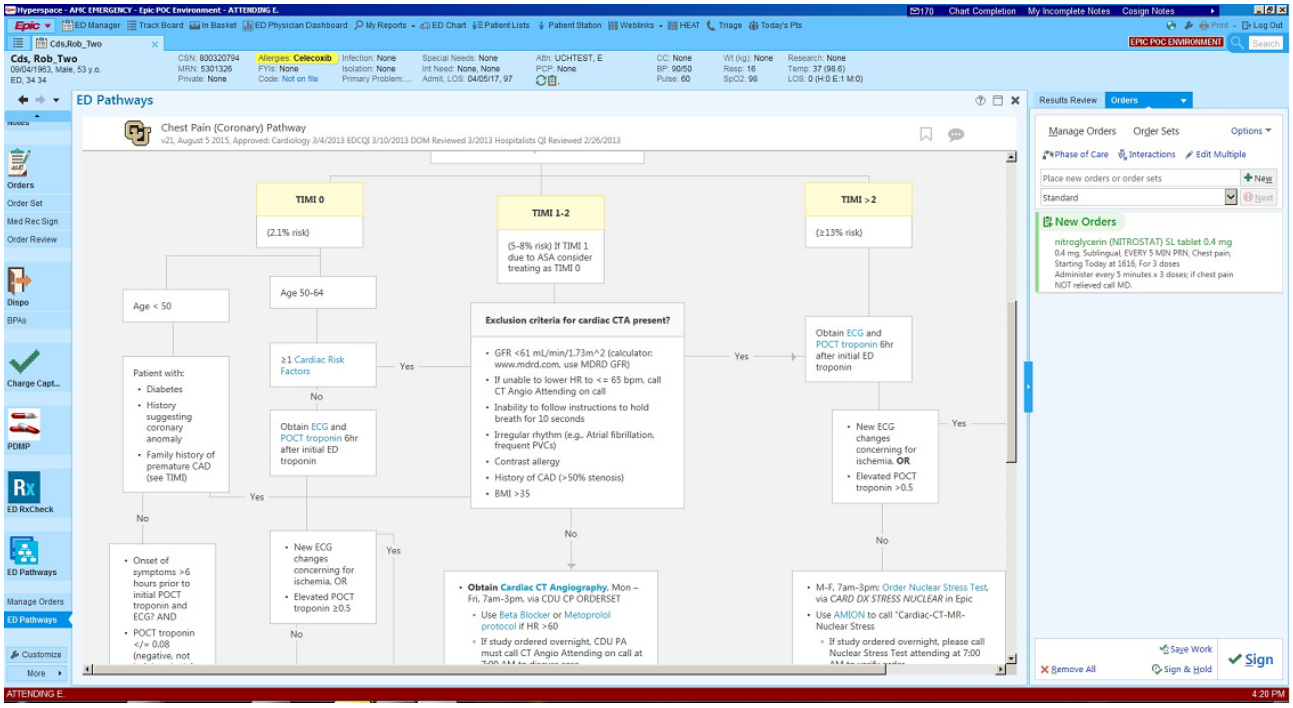
Encounter-Level CDS

Variability in management of common diagnoses such as pneumonia, heart failure, and COPD exacerbation may lead to poor outcomes. Conversely, improved standardization of care can improve outcomes [21,22]. Care pathways are one method of encouraging treatment standardization to improve patient outcomes while permitting flexibility within specific patient presentations [23]. They are often presented as workflows, prompting providers to adhere to recommended care practices. Historically, these pathways have been represented as static flowcharts that the provider can follow in a stepwise fashion, often as a printed diagram or as a digital picture. However, this method is inefficient and the flowcharts are often overlooked, as the decision aid is not made directly available in a provider's workflow. At our institution, we have embedded evidence-based care pathways within the EMR using web-based content in the clinical workflow for evaluation and treatment of patients presenting with common chief concerns such as chest pain or headache. The interface provides an interactive decision tree–like diagram similar to historic representations of care pathways to facilitate standardized patient management. However, the visual tool is implemented using a combination of native EMR functionality and third-party vendor technology, which allows the provider to place orders, perform calculations, and await results. CDS care pathways have been implemented most extensively in our ED. For example, the EMR will present a chest pain care pathway to the provider within the EMR record of a patient reporting a chief concern of chest pain (Figure 1). The provider navigates through the diagram workup steps and initial treatment (eg, aspirin, beta-blocker, nitroglycerin), and an interactive calculator is presented for the Thrombolysis In Myocardial Infarction score, which is a simple risk-stratification algorithm for non–ST-segment elevation myocardial infarction and unstable angina [24].

The care pathway links to an order placement queue for actions such as medications or additional tests, which is provided inline within the EMR. Importantly, orders can be placed directly from the flow diagram rather than requiring the provider to switch back and forth between applications or refer back to a printed diagram when taking a clinical action. Instead, the decision support is provided directly within the provider's workflow. Later in the encounter, the provider can return to the pathway at any given step (e.g., when new test results become available). By embedding order recommendations based on risk stratification within the care pathway, adherence to institutional care standards becomes the most efficient clinical workflow for the provider while preserving the ability to take alternative actions that may be warranted by the clinical scenario. Seamless integration of CDS technology within the EMR workflow provides a straightforward way to standardize care that promotes adherence to guideline-based therapy.

Important challenges in care pathway development include limitations of native EMR functionality in terms of dynamic data collection (eg, provider-entered data), visualization technology, and ability to translate clinical decisions into action. At our institution, implementation of these care pathways was made possible only by partnering with a third-party CDS vendor who developed the EMR interface as well as the data capture, analysis, and visualization technologies to achieve seamless clinical workflow integration. Based on our experiences with third-party technology for care pathways, we anticipate that the ability of a system to integrate with content and technology vendors will greatly expand options for innovative use of CDS.

Figure 1. Example of an encounter-level care pathway clinical decision support application using synthetic patient data.



Well-described predictive models can provide a specific assessment of a patient's risk using a real-time EMR. Implementing CDS based on risk scores can support evaluation and intervention using guideline-directed therapy. EMRs now include some risk models as part of their native functionality, such as the Length of stay, Acuity of admission, Comorbidities, Emergency department (LACE+) model or Early Warning Scores (EWS) [25-29]. Simple applications can alert case managers or nursing supervisors, respectively, when a patient's risk score exceeds a specific threshold, prompting further patient evaluation. For example, we use EWS on medical-surgical wards for early identification of patient deterioration. Score-driven CDS is calculated and presented to the patient's nurse manager, who generates a rapid response alert to evaluate the patient. Recent advances in EWS CDS systems have enabled real-time collection of patient vital signs and more diverse clinical data, more frequent calculation of EWS, and use of more complex models that predict clinical deterioration, allowing interventions to be initiated earlier to prevent or minimize adverse outcomes.

Important considerations include the timing, amount, and reliability of the data used for these scores. Data collection can be achieved directly from the EMR via third-party collection devices imported to the EMR or by manual entry, and significant challenges exist in extracting accurate data from the EMR or when providers are required to enter additional data to support complex predictive models [30]. New tools are in development to address these challenges and improve the quality of data used in these risk scores, coordinate various data sources, optimize entry and extraction of data with the EMR, and support implementation of larger libraries of complex models.

As with all CDS, a significant challenge in generating specific application requirements is optimization of CDS-workflow integration. Application requirements can be highly specific to disease processes, clinical venues, provider groups, or institutions, all of which must be considered. Furthermore, the same data (eg, risk assessment) may need to be delivered in a variety of ways depending on the care venue and providers. For the same CDS result, in some cases, an interruptive alert during a clinic visit may be preferred for physicians, a work list element may be preferred for care mangers, and an educational email may be preferred for a patient. The variable requirements of different CDS applications require a wide range of technical functionality. We addressed this by developing an array of CDS tools, including the EMR, third-party software, and custom applications built in-house in collaboration with a clinical champion.

Precision-Level CDS

Many existing CDS applications have been directed at providers based on coarse classifications (eg, presence of diabetes or hospitalization for acute myocardial infarction). With increasing focus on shared decision-making, emerging CDS tools are directed at providing individualized assessments and predictions to facilitate complex discussions between patients and providers. Dependence on many varied data inputs can discourage clinicians from using such models on a practical basis despite their superior predictive ability. Novel data streams such as genomic analyses and fitness trackers, or the variety of data requirements for highly accurate predictive algorithms, can eclipse human capacity to comprehensively process data. CDS analytics are therefore essential to synthesize the amount, breadth, and complexity of data necessary for precision medicine, such as in the cases of genomic medicine, data from wearable devices, and complex patient-reported outcome instruments.

As the least well-developed of the three domains, precision medicine CDS faces many challenges. There is relatively little knowledge or understanding of how to implement and use applications requiring myriad data points to generate highly

patient-specific management plans, and developing this knowledge is outside the capacity of most health care organizations. Precision medicine strategies may need to both educate and support providers in decision-making where the foundational knowledge to create the CDS is not widespread. Finally, optimal methods of patient engagement (eg, format and mode of delivery) using this information are not yet well understood.

At our institution, we have approached these complex challenges by developing tools such as an interactive CDS application in collaboration with members of the Surgical Outcomes and Applied Research Program within the Department of Surgery. This application is based on the published Surgical Risk Preoperative Assessment System (SURPAS), which is a set of risk predictive algorithms developed from the American College of Surgeons National Surgical Quality Improvement Program data to predict 11 adverse postoperative outcomes using 8 preoperatively available predictor variables [31].

The SURPAS intervention allows individualized inputs into the model that provide a precise and personalized risk assessment instead of a categorized level of risk. Using this method ensures precision medicine CDS in which the model will not make the same recommendations for different patients. Within the EHR, SURPAS automatically combines previously existing EHR data for each patient with provider-entered patient data to calculate the patient's procedure-specific likelihood of 11 different surgical complications. Then, the patient- and procedure-specific risk assessments are compared to national averages ranging from renal injury to 30-day mortality. During the preoperative office visit, patients are presented with their individualized risks of postoperative adverse outcomes as an infographic education tool, which streamlines the risk discussion and consent process, encourages patient engagement, and alerts providers to individual patient risk profiles that may guide preparations for postoperative care (Figure 2). Finally, risk estimates can be imported directly into clinic notes to document the basis for patient-centered care decisions.

Precision-level CDS can also provide recommendations in the context of both complex risk assessment and a patient's current management. For example, our institution leverages EMR data to generate alerts based on risk modeling to identify patients for initiation or modification of cholesterol management protocols. In 2013, the Adult Treatment Panel IV on cholesterol management updated practice guidelines for patients with prevention of atherosclerotic cardiovascular disease (ASCVD). The guideline now provides recommendations for determining the appropriate intensity of statin therapy based on four recommended risk groups, one of which is defined by a complex ASCVD risk model [32]. We have built a multimodal CDS application that uses extracted data, including the ASCVD risk score, to classify patients according to four specified risk groups to determine the recommended intensity of statin therapy, if any. The algorithm then evaluates the presence and intensity of ongoing statin therapy to generate a recommendation to the provider only if a change in statin therapy is indicated. The results and recommendations are visualized as alerts presented to the provider via the EMR during the encounter to facilitate guideline-based care (Figure 3).

Figure 2. Screenshot of the Surgical Risk Preoperative Assessment System, a personalized risk assessment clinical decision support application used to guide postoperative care, with synthetic patient data.

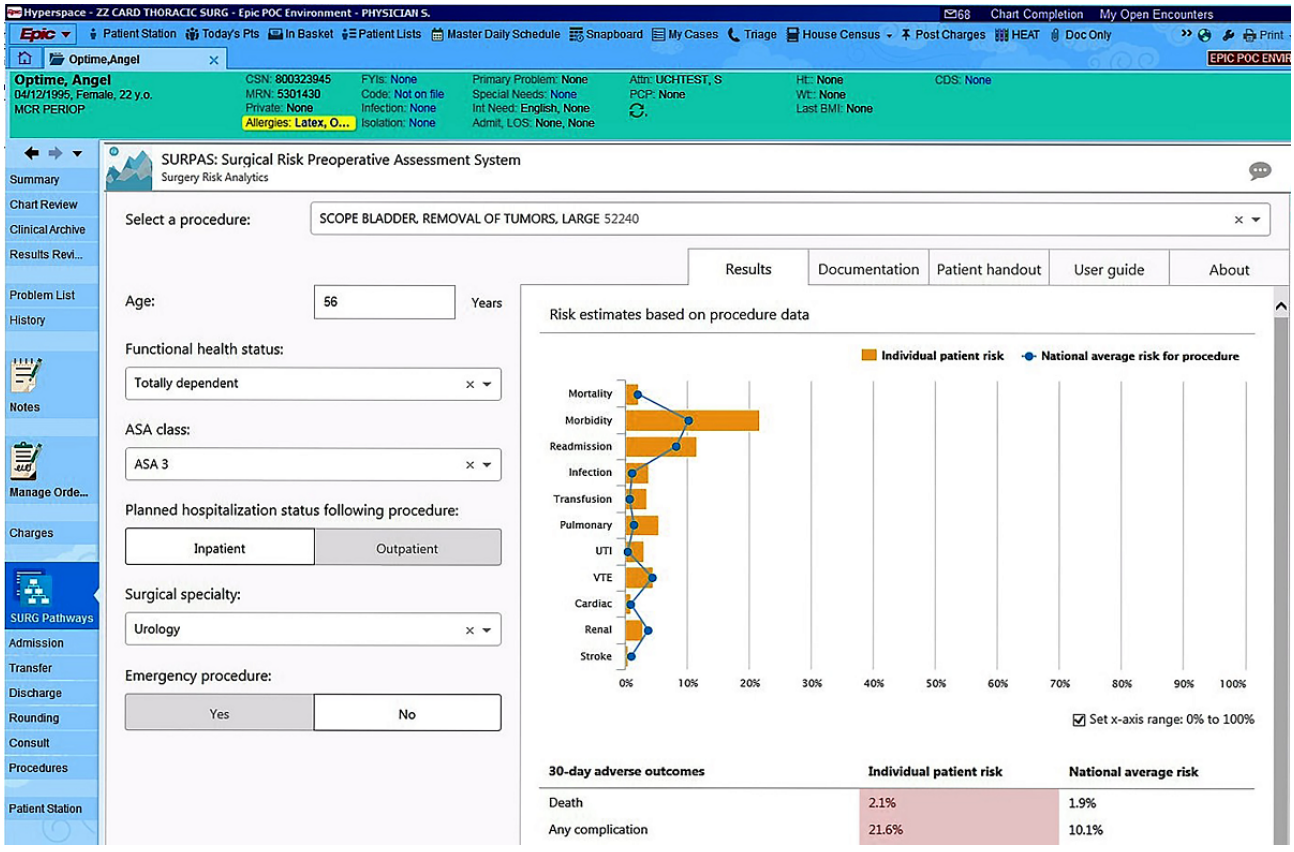
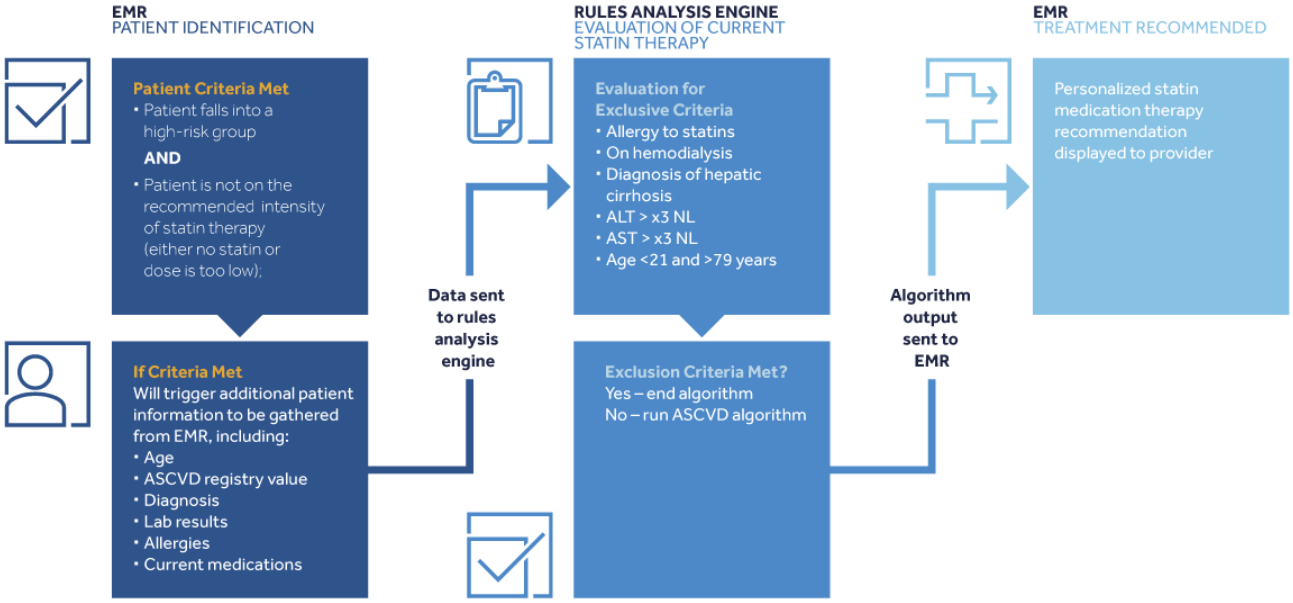


Figure 3. An atherosclerotic cardiovascular disease risk scoring algorithm to classify patients according to their risk group. ALT: alanine aminotransferase; ASCVD: atherosclerotic cardiovascular disease; EMR: electronic medical record; NL: normal.



Discussion

CDS Implementation Across All Three Domains

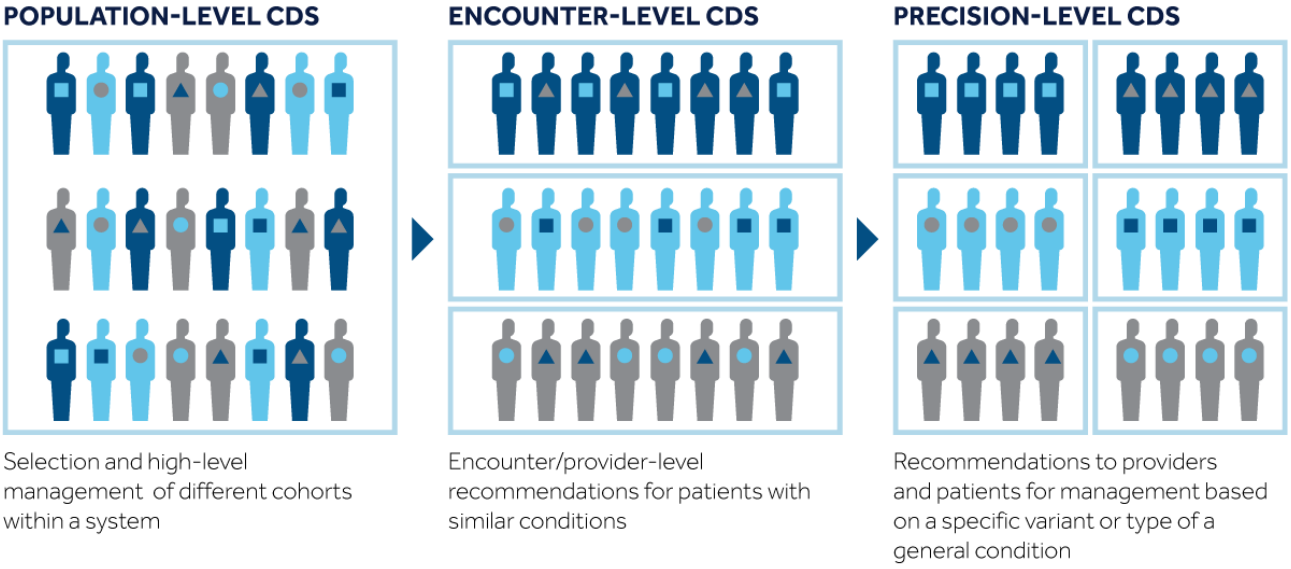
Care delivery across all three clinical care domains using integrated CDS applications has the potential to improve efficiency and quality for individual patients; however, significant planning and development effort is required to ensure that applications align with several types of end-user workflows. The three different levels of CDS have distinct, valuable roles with specific requirements and functionality (Table 1). All 3 levels should be used in concert when optimizing and coordinating a patient's care. Figure 4 presents a construct demonstrating the relationship of these care domains. Although the diagram flows from population to encounter to precision domains, in practice, workflows may practically or temporally move across the domains in any order, depending on the algorithm and the condition in question. The person requesting an application as well as the intended end users should ideally be engaged in the development process to optimize integration into the providers' workflow and to maximize provider uptake. Historically, the available technology has determined the types of CDS applications that can be requested; however, as the CDS capabilities of an institution grow, clinical use cases increasingly determine the selection of technical approaches from the organization's "toolbox."

Table 1. Salient features of each domain according to each level of CDS.

Feature	CDS[a] level		
	Population	Encounter	Precision
Alert timing	Asynchronous	Synchronous	Both
Data timing	Cumulative	At time of encounter	Preemptive
Basis	Evidence-based	Evidence-based	Individualized
Strategies	Dashboards, iterative risk scores, work lists	Interruptive alerts, risk scores, care pathways, passive alerts	Either population- or encounter-based
Clinical examples	Diabetic foot exam, HIV drug efficacy, annual cholesterol	Heart failure guidelines, QT-prolonging drugs, performance metrics, deterioration index	Telemonitoring and mobile health, *CYP2D6*/opiate metabolism, polygenic risk scores, BRCA cancer screening

[a]CDS: clinical decision support.

Figure 4. Overview of the integration and relationship of CDS across three domains of health care. CDS: clinical decision support.



As an example, our institution is developing an end-to-end, multi-tiered precision medicine framework that will use complex predictive models that integrate clinical data with a patient's genetic profile (ie, a single gene or a few genes) or genomic profile (ie, many or all genes) to assist in clinical decision-making. First, patients undergo genome-wide single nucleotide polymorphism analysis to produce clinically actionable genetic data, such as genes that increase disease risk or impact drug effectiveness. These data are then returned to the patient's individual record in the EMR, which may occur long before the genes become clinically relevant (eg, when a clinician attempts to prescribe a specific medication whose effectiveness could be influenced by the patient's genotype). CDS applications will use these data to support population care management, encounter-based tools, integrated third-party data collection methods, and genetic and genomic data, producing result displays that are customized separately for patients and providers. At a population level, high-risk patients will be identified with the help of select genetic data that are relevant for assessing the patients' risk of developing certain diseases. Care pathways relevant to these high-risk populations may recommend increased intensity of disease screening or review of possible preventive interventions. Assigned care managers will coordinate long-term care management for those patients; this may include appropriate proactive care plans such as lifestyle modifications, genetic counseling, family education, or frequent diagnostic screening tests. At the encounter level, CDS will alert providers when ordering a medication affected by a patient's genetics by providing real-time feedback to the clinician using evolving evidence regarding relevant pharmacogenetic markers. Patient education and engagement will be enhanced by combining genetic information with the results of risk modeling algorithms. This information will be shared with patients along with educational resources and genetic counseling as appropriate, facilitating collaborative decision-making between the patient and provider. This framework has required contributions from a wide array of experts, including basic genomic scientists, the pathology laboratory and clinical laboratory systems, clinical informaticians, pharmacists, EMR technical staff, patient representatives, hospital legal representation, and ethics committees. We have also greatly increased the technical capacity of our EMR to return the genetic data in a format that can be used by all three levels of CDS.

To illustrate how CDS can impact an individual patient through the three care domains, we present the trajectory of LK, a hypothetical patient with COPD, at different stages of care management in Table 2. Although this specific example is theoretical, we have developed similar applications at our institution at each level of care for different disease domains as described above. We use the example of LK to cohesively illustrate how all these applications could be integrated to support a comprehensive, holistic approach to the care of her chronic disease. Management of LK's COPD used population-level CDS via the registry (COPD), management protocols (annual spirometry and symptoms), and alerts regarding both planned (care provider) and unplanned (ED or inpatient) patient encounters to coordinate care (eg, home health evaluation). The encounter-based CDS helped providers choose and implement the correct guideline-based medications for different scenarios (corticosteroid for worsening COPD, cefepime when hospitalized for COPD, hydrocodone vs codeine based on CYP2D6 genotype, influenza vaccine on hospital discharge). Precision-level CDS helped refine management based on LK's specific characteristics including her genetics (selection of cefepime based on age and forced expiratory volume in first second of expiration [FEV1] selection of hydrocodone, and use of capnography monitoring based on CYP2D6). In this case, each CDS domain provided a level of integrated care coordination to manage LK's COPD and contributed significantly to LK's management and improved quality of care.

Table 2. Integration of CDS across the population, encounter, and precision care domains of LK, a hypothetical 68-year-old female patient with COPD.

Care management action	Associated CDS[a] level
LK is assigned a care management team (disease registry) that monitors her clinical status using annual office spirometry.	Population
After 3 years, longitudinal analytics alert LK's care managers that her spirometry is declining and her symptoms are increasing.	Population
Based on this trend, the team schedules an appointment with her health care provider. The provider considers starting a long-acting beta-agonist alone, but when he tries to order one, he is prompted to start an inhaled corticosteroid in accordance with present guidelines.	Encounter
After 6 months, LK has a severe COPD[b] exacerbation. She contacts her care team through an EMR[c], and they advise her to go to the emergency department.	Population
When LK is admitted to the hospital, the EMR recommends intravenous cefepime because she meets the criteria for complicated COPD based on her age of older than 65 years and a recent spirometry FEV_1[d] measurement of less than 50% predicted. During her hospitalization, LK develops a rib fracture from coughing and has severe pain. A genomic analysis performed two years earlier as part of the institution's precision medicine program determined that she had multiple copies of the *CYP2D6* gene, indicating an increased likelihood of excessive sedation from codeine-containing cough syrups due to rapid conversion into morphine.	Encounter and precision
The hospitalist is alerted to her pharmacogenetic status and prescribes hydrocodone instead of codeine for management of pain and cough, and capnography monitoring is used to monitor for respiratory depression or failure.	Encounter and precision
LK is ready for discharge after 5 days. Based on her known COPD and hospitalization, the EMR recommends an influenza vaccine prior to discharge.	Population and encounter
The discharging team arranges follow-up with LK's primary care provider. Her chronic care managers receive an alert that she is being discharged and contact her three days later. Through a video call, they learn that she is having trouble with daily activities due to deconditioning and the rib fracture. A home health evaluation is arranged, and physical therapy and home health nursing are prescribed. LK improves over the next 2 weeks and returns to her baseline surveillance schedule.	Population

[a]CDS: clinical decision support.

[b]COPD: chronic obstructive pulmonary disease.

[c]EMR: electronic medical record.

[d]FEV_1: forced expiratory volume in first second of expiration.

Conclusions

CDS is challenging to design and implement; however, significant progress has been made, with improvements in timely and workflow-specific management recommendations, EMRs, and resources created by third-party vendors. Conceptualizing CDS tools in the context of linked population-, encounter-, and precision-level health care affords an opportunity to integrate complex algorithms at each level into a unified mechanism for improving care across all levels of patient management.

Acknowledgments

Special thanks to Robert Meguid, MD (Surgical Outcomes and Applied Research Program, University of Colorado) for supplying information about the SURPAS program and to Steven W Lee, PhD (Medtronic, Inc) for providing medical writing and editorial support. This work was funded by Medtronic, Inc.

Authors' Contributions

All authors prepared, reviewed, and approved the final manuscript.

Conflicts of Interest

CL and DF are consultants of Medtronic, Inc. KS is a former employee of Medtronic, Inc. DK has no conflicts of interest to declare.

References

1. Shekelle PG, Morton SC, Keeler EB. Costs and benefits of health information technology. Evid Rep Technol Assess (Full Rep) 2006 Apr(132):1-71. [doi: 10.23970/ahrqepcerta132] [Medline: 17627328]
2. Report to Congressional Requesters: Patient safety: Hospitals face challenges implementing evidence-based practices. US Government Accountability Office. 2016 Feb. URL: https://patientsafetymovement.org/wp-content/uploads/2016/03/GAO_Report.pdf [accessed 2020-09-14]

3. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med 2012 Jul 03;157(1):29-43 [FREE Full text] [doi: 10.7326/0003-4819-157-1-201207030-00450] [Medline: 22751758]
4. Fillmore CL, Bray BE, Kawamoto K. Systematic review of clinical decision support interventions with potential for inpatient cost reduction. BMC Med Inform Decis Mak 2013 Dec 17;13:135 [FREE Full text] [doi: 10.1186/1472-6947-13-135] [Medline: 24344752]
5. Jones SS, Rudin RS, Perry T, Shekelle PG. Health information technology: an updated systematic review with a focus on meaningful use. Ann Intern Med 2014 Jan 07;160(1):48-54 [FREE Full text] [doi: 10.7326/M13-1531] [Medline: 24573664]
6. Roshanov PS, Misra S, Gerstein HC, Garg AX, Sebaldt RJ, Mackay JA, CCDSS Systematic Review Team. Computerized clinical decision support systems for chronic disease management: a decision-maker-researcher partnership systematic review. Implement Sci 2011 Aug 03;6:92 [FREE Full text] [doi: 10.1186/1748-5908-6-92] [Medline: 21824386]
7. Roshanov PS, You JJ, Dhaliwal J, Koff D, Mackay JA, Weise-Kelly L, CCDSS Systematic Review Team. Can computerized clinical decision support systems improve practitioners' diagnostic test ordering behavior? A decision-maker-researcher partnership systematic review. Implement Sci 2011 Aug 03;6:88 [FREE Full text] [doi: 10.1186/1748-5908-6-88] [Medline: 21824382]
8. Stabile M, Cooper L. Review article: the evolving role of information technology in perioperative patient safety. Can J Anaesth 2013 Feb;60(2):119-126. [doi: 10.1007/s12630-012-9851-0] [Medline: 23224715]
9. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc 2003;10(6):523-530 [FREE Full text] [doi: 10.1197/jamia.M1370] [Medline: 12925543]
10. Campbell RJ. The five rights of clinical decision support: CDS tools helpful for meeting meaningful use. Journal of AHIMA 2013;84(10):42-47.
11. Kindig D, Stoddart G. What is population health? Am J Public Health 2003 Mar;93(3):380-383. [doi: 10.2105/ajph.93.3.380] [Medline: 12604476]
12. Redekop WK, Mladsi D. The faces of personalized medicine: a framework for understanding its meaning and scope. Value Health 2013;16(6 Suppl):S4-S9 [FREE Full text] [doi: 10.1016/j.jval.2013.06.005] [Medline: 24034312]
13. Afzal M, Riazul Islam SM, Hussain M, Lee S. Precision medicine informatics: principles, prospects, and challenges. IEEE Access 2020;8:13593-13612. [doi: 10.1109/access.2020.2965955]
14. Snow V, Beck D, Budnitz T, Miller DC, Potter J, Wears RL, American College of Physicians, Society of General Internal Medicine, Society of Hospital Medicine, American Geriatrics Society, American College of Emergency Physicians, Society of Academic Emergency Medicine. Transitions of Care Consensus Policy Statement American College of Physicians-Society of General Internal Medicine-Society of Hospital Medicine-American Geriatrics Society-American College of Emergency Physicians-Society of Academic Emergency Medicine. J Gen Intern Med 2009 Aug 3;24(8):971-976 [FREE Full text] [doi: 10.1007/s11606-009-0969-x] [Medline: 19343456]
15. Cornu P, Steurbaut S, De Beukeleer M, Putman K, van de Velde R, Dupont AG. Physician's expectations regarding prescribing clinical decision support systems in a Belgian hospital. Acta Clin Belg 2014 Jun;69(3):157-164. [doi: 10.1179/2295333714Y.0000000015] [Medline: 24820921]
16. Cresswell KM, Bates DW, Williams R, Morrison Z, Slee A, Coleman J, et al. Evaluation of medium-term consequences of implementing commercial computerized physician order entry and clinical decision support prescribing systems in two 'early adopter' hospitals. J Am Med Inform Assoc 2014 Oct;21(e2):e194-e202 [FREE Full text] [doi: 10.1136/amiajnl-2013-002252] [Medline: 24431334]
17. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015 Feb 26;372(9):793-795 [FREE Full text] [doi: 10.1056/NEJMp1500523] [Medline: 25635347]
18. Nayak L, Ray I, De RK. Precision medicine with electronic medical records: from the patients and for the patients. Ann Transl Med 2016 Oct;4(Suppl 1):S61 [FREE Full text] [doi: 10.21037/atm.2016.10.40] [Medline: 27868029]
19. Guidi G, Pettenati M, Miniati R, Iadanza E. Heart failure analysis dashboard for patient's remote monitoring combining multiple artificial intelligence technologies. In: Conf Proc IEEE Eng Med Biol Soc. 2012 Presented at: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society; August 28-September 1, 2012; San Diego, CA p. 2210-2213. [doi: 10.1109/EMBC.2012.6346401]
20. Kao DP, Lindenfeld J, Macaulay D, Birnbaum HG, Jarvis JL, Desai US, et al. Impact of a telehealth and care management program on all-cause mortality and healthcare utilization in patients with heart failure. Telemed J E Health 2016 Jan;22(1):2-11 [FREE Full text] [doi: 10.1089/tmj.2015.0007] [Medline: 26218252]
21. Di Palo KE, Piña IL, Ventura HO. Improving provider adherence to guideline recommendations in heart failure. Curr Heart Fail Rep 2018 Dec;15(6):350-356. [doi: 10.1007/s11897-018-0411-y] [Medline: 30238398]
22. Hortmann M, Heppner H, Popp S, Lad T, Christ M. Reduction of mortality in community-acquired pneumonia after implementing standardized care bundles in the emergency department. Eur J Emerg Med 2014 Dec;21(6):429-435. [doi: 10.1097/MEJ.0000000000000106] [Medline: 24384619]
23. Chawla A, Westrich K, Matter S, Kaltenboeck A, Dubois R. Care pathways in US healthcare settings: current successes and limitations, and future challenges. Am J Manag Care 2016 Jan;22(1):53-62 [FREE Full text] [Medline: 26799125]

24. Antman EM, Cohen M, Bernink PJ, McCabe CH, Horacek T, Papuchis G, et al. The TIMI risk score for unstable angina/non-ST elevation MI: a method for prognostication and therapeutic decision making. JAMA 2000 Aug 16;284(7):835-842. [doi: 10.1001/jama.284.7.835] [Medline: 10938172]
25. Amarasingham R, Velasco F, Xie B, Clark C, Ma Y, Zhang S, et al. Electronic medical record-based multicondition models to predict the risk of 30 day readmission or death among adult medicine patients: validation and comparison to existing models. BMC Med Inform Decis Mak 2015 May 20;15:39 [FREE Full text] [doi: 10.1186/s12911-015-0162-6] [Medline: 25991003]
26. Escobar GJ, LaGuardia JC, Turk BJ, Ragins A, Kipnis P, Draper D. Early detection of impending physiologic deterioration among patients who are not in intensive care: development of predictive models using data from an automated electronic medical record. J Hosp Med 2012;7(5):388-395. [doi: 10.1002/jhm.1929] [Medline: 22447632]
27. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. QJM 2001 Oct;94(10):521-526. [doi: 10.1093/qjmed/94.10.521] [Medline: 11588210]
28. Wang H, Robinson RD, Johnson C, Zenarosa NR, Jayswal RD, Keithley J, et al. Using the LACE index to predict hospital readmissions in congestive heart failure patients. BMC Cardiovasc Disord 2014 Aug 07;14:97 [FREE Full text] [doi: 10.1186/1471-2261-14-97] [Medline: 25099997]
29. van Walraven C, Wong J, Forster A. LACE+ index: extension of a validated index to predict early death or urgent readmission after hospital discharge using administrative data. Open Med 2012;6(3):e80-e90 [FREE Full text] [Medline: 23696773]
30. Amiri M, Kelishadi R. Comparison of models for predicting outcomes in patients with coronary artery disease focusing on microsimulation. Int J Prev Med 2012 Aug;3(8):522-530 [FREE Full text] [Medline: 22973481]
31. Meguid RA, Bronsert MR, Juarez-Colunga E, Hammermeister KE, Henderson WG. Surgical Risk Preoperative Assessment System (SURPAS): I. Parsimonious, Clinically Meaningful Groups of Postoperative Complications by Factor Analysis. Ann Surg 2016 Jun;263(6):1042-1048. [doi: 10.1097/SLA.0000000000001669] [Medline: 26954897]
32. Stone NJ, Robinson JG, Lichtenstein AH, Goff DC, Lloyd-Jones DM, Smith SC, 2013 ACC/AHA Cholesterol Guideline Panel. Treatment of blood cholesterol to reduce atherosclerotic cardiovascular disease risk in adults: synopsis of the 2013 American College of Cardiology/American Heart Association cholesterol guideline. Ann Intern Med 2014 Mar 04;160(5):339-343 [FREE Full text] [doi: 10.7326/M14-0126] [Medline: 24474185]

Abbreviations

ASCVD: atherosclerotic cardiovascular disease
CDS: clinical decision support
COPD: chronic obstructive pulmonary disease
ED: emergency department
EMR: electronic medical record
EWS: Early Warning Scores
FEV1: forced expiratory volume in first second of expiration
LACE+: Length of stay, Acuity of admission, Comorbidities, Emergency department
SURPAS: Surgical Risk Preoperative Assessment System

Edited by M Focsa; submitted 14.05.20; peer-reviewed by M Afzal, A Gupta; comments to author 24.06.20; revised version received 31.07.20; accepted 01.08.20; published 16.10.20.

Please cite as:
Kao D, Larson C, Fletcher D, Stegner K
Clinical Decision Support May Link Multiple Domains to Improve Patient Care: Viewpoint
JMIR Med Inform 2020;8(10):e20265
URL: https://medinform.jmir.org/2020/10/e20265
doi:10.2196/20265
PMID:33064106

Original Paper

Implementation of a Cohort Retrieval System for Clinical Data Repositories Using the Observational Medical Outcomes Partnership Common Data Model: Proof-of-Concept System Validation

Sijia Liu[1], PhD; Yanshan Wang[1], PhD; Andrew Wen[1], MSc; Liwei Wang[1], MD, PhD; Na Hong[1], PhD; Feichen Shen[1], PhD; Steven Bedrick[2], PhD; William Hersh[3], MD; Hongfang Liu[1], PhD

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

[2]Department of Computer Science and Electrical Engineering, Oregon Health & Science University, Portland, OR, United States

[3]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, United States

Corresponding Author:
Hongfang Liu, PhD
Department of Health Sciences Research
Mayo Clinic
200 First Street SW
Rochester, MN, 55901
United States
Phone: 1 507 293 0057
Email: liu.hongfang@mayo.edu

Abstract

Background: Widespread adoption of electronic health records has enabled the secondary use of electronic health record data for clinical research and health care delivery. Natural language processing techniques have shown promise in their capability to extract the information embedded in unstructured clinical data, and information retrieval techniques provide flexible and scalable solutions that can augment natural language processing systems for retrieving and ranking relevant records.

Objective: In this paper, we present the implementation of a cohort retrieval system that can execute textual cohort selection queries on both structured data and unstructured text—Cohort Retrieval Enhanced by Analysis of Text from Electronic Health Records (CREATE).

Methods: CREATE is a proof-of-concept system that leverages a combination of structured queries and information retrieval techniques on natural language processing results to improve cohort retrieval performance using the Observational Medical Outcomes Partnership Common Data Model to enhance model portability. The natural language processing component was used to extract common data model concepts from textual queries. We designed a hierarchical index to support the common data model concept search utilizing information retrieval techniques and frameworks.

Results: Our case study on 5 cohort identification queries, evaluated using the precision at 5 information retrieval metric at both the patient-level and document-level, demonstrates that CREATE achieves a mean precision at 5 of 0.90, which outperforms systems using only structured data or only unstructured text with mean precision at 5 values of 0.54 and 0.74, respectively.

Conclusions: The implementation and evaluation of Mayo Clinic Biobank data demonstrated that CREATE outperforms cohort retrieval systems that only use one of either structured data or unstructured text in complex textual cohort queries.

(JMIR Med Inform 2020;8(10):e17376) doi:10.2196/17376

KEYWORDS

cohort retrieval; information retrieval; common data model; electronic health records; natural language processing

Introduction

The widespread adoption of electronic health records has enabled the use of clinical data for clinical research and health care delivery [1]. Many institutions have established clinical data repositories in conjunction with cohort retrieval tools (eg, Informatics for Integrating Biology & the Bedside) to support the use of clinical data for research including retrospective studies as well as feasibility assessment or patient recruitment for clinical trials. However, electronic health record–based clinical research has been hampered by poor research reproducibility caused by the heterogeneity and complexity of both health care institutions and electronic health record systems.

For structured electronic health record data, to ensure a standardized and logically unified representation of electronic health record data across multiple institutions (and across multiple sites), many large-scale clinical research networks such as Accrual to Clinical Trials [2], Electronic Medical Records and Genomics [3], and National Patient-Centered Clinical Research [4] have adopted common data models aimed at producing comparable and reproducible results with the same research methods [5,6]. Our prior investigation [7] demonstrated the generalizability of one of the common data models, the Observational Medical Outcomes Partnership (OMOP) Common Data Model, in achieving structural and semantic consistency of electronic health record data in multi-institutional research with the Observational Health Data Sciences and Informatics (OHDSI) program [7].

In electronic health records, a significant portion of relevant patient information is embedded as unstructured text, and natural language processing techniques such as information extraction are critical when using these data for clinical research [8-11]. Many clinical natural language processing systems have been developed to extract information from text for various downstream applications [12,13] but have challenges in performance and portability [14-17]. Information retrieval, a technique used in search engines for storing, retrieving, and ranking documents from a large collection of text documents based on users' queries, can provide an alternative approach to leverage clinical narratives for cohort retrieval as it is less semantic-dependent and can involve end users in the loop [18,19]. The combination of natural language processing and information retrieval is a promising solution for cohort retrieval from unstructured clinical text, and there are several review articles [5,20] about information retrieval or natural language processing techniques for case detection.

However, most of the current clinical data repository implementations do not support searches on both structured and unstructured text, seamlessly. An efficient and comprehensive patient-level search engine on both structured and unstructured data from electronic health record is, therefore, still highly demanded by health care practitioners and researchers. In this paper, we describe a proof-of-concept implementation of a cohort retrieval system—Cohort Retrieval Enhanced by Analysis of Text from Electronic Health Records (CREATE)—in which the same query to search both structured (electronic health record represented using the OMOP Common Data Model) and unstructured text (leveraging a concept extraction system) are used. Cohort retrieval in CREATE is conducted in 2 phases: the first phase filters patients using structured data, and the second phase retrieves and ranks results at either a document or a patient level. The functionality of the system was tested using a previously assembled query collection [21] on a corpus composed of the electronic health record data from the Mayo Clinic Biobank cohort [22].

There are generally 2 approaches to search unstructured text for purposes such as patient care, clinical research, and traceability of medical care [23]. The first approach is based on a text search. For example, the Electronic Medical Record Search Engine (EMERSE) from the University of Michigan [24] is a full-text search engine with the goal of facilitating the retrieval of information for clinicians, administrators, and clinical or translational researchers based on clinical narratives. However, EMERSE does not support queries using structured electronic health record data such as demographic information, lab tests, and medications. Dr. Warehouse, proposed by Garcelon et al [25], is a free-text search engine using Oracle Text to index its documents. The system is based on relational databases and relies on ranking after retrieval, which may limit its capability to deploy state-of-the-art information retrieval methods such as best match 25 or Markov random fields. The other approach to searching unstructured text is to extract concepts using natural language processing systems. For example, SemEHR [26] is a semantic search engine based on a Fast Healthcare Interoperability Resources [27] representation of clinical semantic concepts extracted from a clinical natural language processing system named Bio-YODIE. The system showed a high performance in retrieving patients given queries of single concepts, such as Hepatitis C and HIV, in local electronic health record and lab test results when evaluated with the MIMIC-III (Medical Information Mart for Intensive Care) data set [26].

National NLP Clinical Challenges 2018 (Shared-Task Track 1) [28] also contributed to standardized evaluations of cohort retrieval systems from electronic health records. The evaluation data set includes clinical narrative texts of 288 patients for concept extraction, temporal reasoning, and inferencing. The official evaluation indicated that the top systems used rule-based and hybrid systems for the problems and led the directions of future system development for similar tasks. The 2018 corpus consists of semistructured and narrative text. The structured data are provided via sections of semistructured text rather than in structured formats. Therefore, cohort retrieval systems for the 2018 corpus require additional components to handle the semistructured metadata, which may not be applicable to systems for real-world electronic health record data.

Several studies [29-31] have addressed the challenge of how to represent textual cohort criteria or queries via syntactic parsing or sequence labeling. The main focuses of proposed methods were to provide the functions of automatic parsing and modeling of textual queries of end-to-end retrieval systems. To further extend querying to support the customization of parsed results by end users, our cohort retrieval system has the following design principles: (1) the adoption of common data models to facilitate cohort retrieval using both structured and unstructured

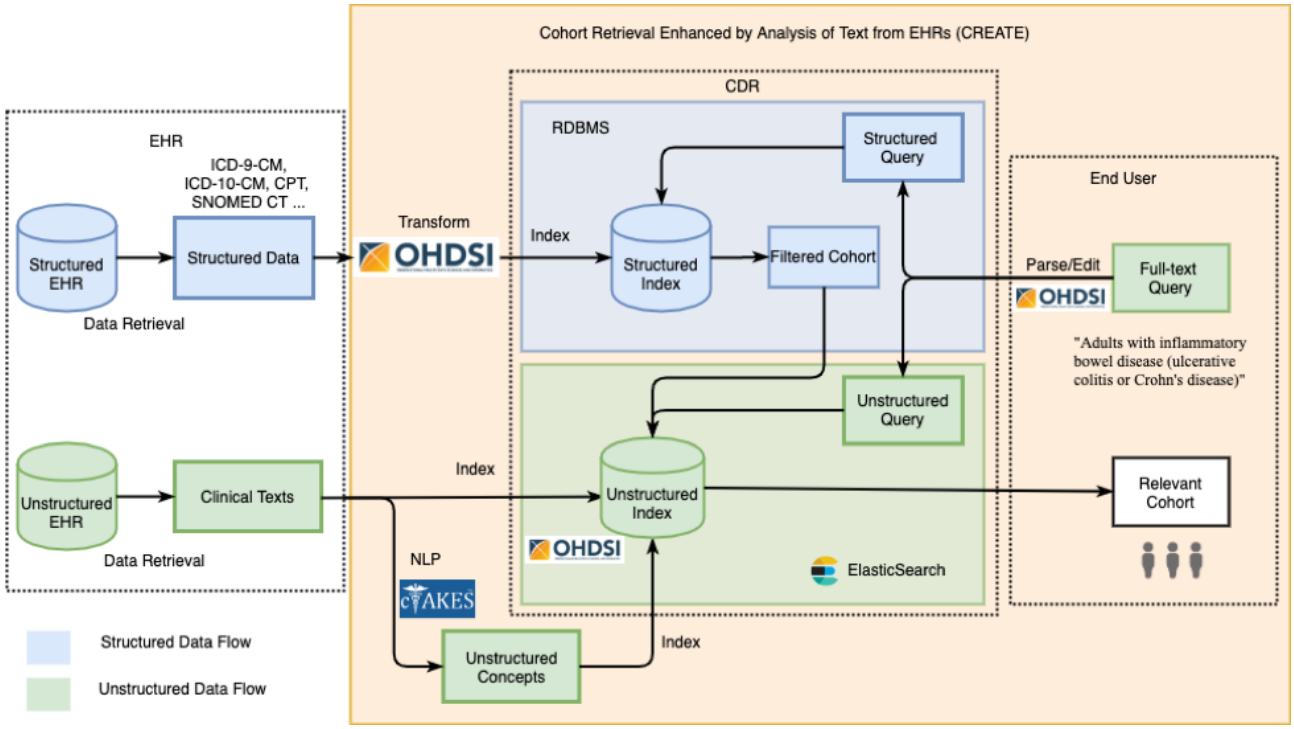
data for multi-institutional research, (2) the flexibility and ability to apply state-of-the-art information retrieval methods in the retrieval system, (3) the incorporation of relevance judgment for downstream machine learning–based cohort selection methods, and (4) the generation of semantic annotations during the indexing phrase to provide a real-time semantic search experience.

Methods

Overview of System Architecture

An overview of our cohort retrieval system for clinical data repositories is shown in Figure 1. Specifically, a textual query is expanded and divided, either automatically or manually, into structured and unstructured data fields according to specific clinical data repository implementations. The query fulfillment for structured data and unstructured text data are managed differently: structured electronic health record data can be retrieved from the corresponding clinical data repositories using Structured Query Language (SQL) on a relational database management system, and the unstructured electronic health record data can be preprocessed by natural language processing and retrieved by leveraging information retrieval techniques. Retrieved results can then be combined and aggregated for clinical research applications, such as clinical trial feasibility assessment or cohort identification. For cohort identification, the retrieved and screened cohort can be treated as a weakly labeled data set. Human relevance judgment is a potential subsequent step to manually validate the results through chart review.

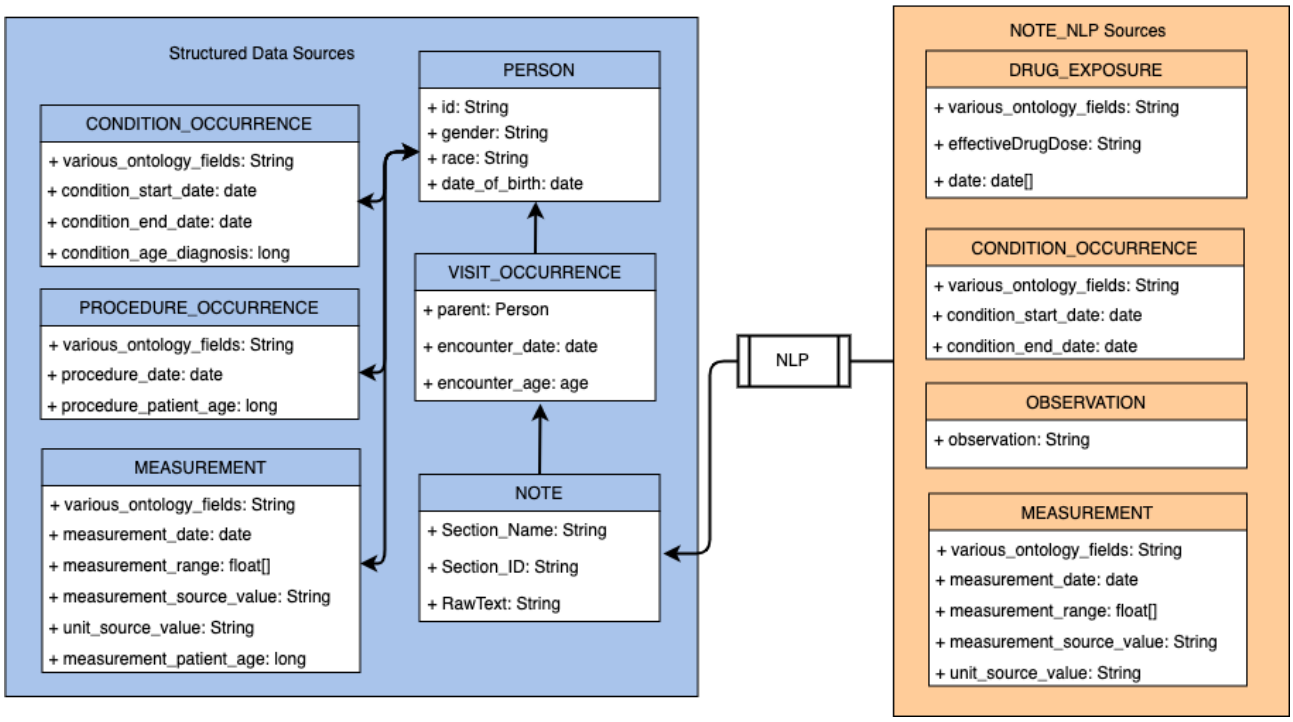
Figure 1. Overview of CREATE's workflow. CDR: clinical data repository; CM: clinical modification; CPT: current procedures terminology; EHR: electronic health record; ICD: International Classification of Diseases; NLP: natural language processing; OHDSI: Observational Health Data Sciences and Informatics; RDBMS: relational database management system.



Adopting OMOP Common Data Model for Patient Retrieval

To improve the interoperability and portability of our system (use with disparate data sources), we adopted the OMOP Common Data Model (version 5.3.1) [32] to index electronic health record data. The hierarchical index structure of clinical data repositories using OMOP Common Data Model for cohort retrieval is shown in Figure 2. The indexed tables include data from both unstructured and structured sources, consisting of extracted OMOP Common Data Model artifacts from unstructured clinical notes and encounter information, demographic information (represented as a common data model person), and diagnoses, procedures, and lab tests from structured data. The distinction between structured and unstructured data varies between different electronic health record systems. The specifics of implementation in adopters may, therefore, differ from those implemented in this study.

Figure 2. Hierarchical index structure using the OMOP Common Data Model. NLP: natural language processing.



Structured data such as procedures, diagnoses, lab tests, and demographics are directly queried from relational databases and loaded into the index through an extract-transform-load process. We map structured data to Unified Medical Language System (UMLS) concept unique identifiers either through the usage of mapping definitions already in the UMLS Metathesaurus [33] (eg, ICD-9-CM or ICD-10-CM, Current Procedural Terminology 4, and SNOMED Clinical Terms) or through the use of natural language processing (eg, local lab test codes). The concepts are subsequently mapped to equivalent OHDSI- or OMOP-compliant vocabulary codes via Athena (version 1.10.0; OHDSI) standardized vocabularies [34].

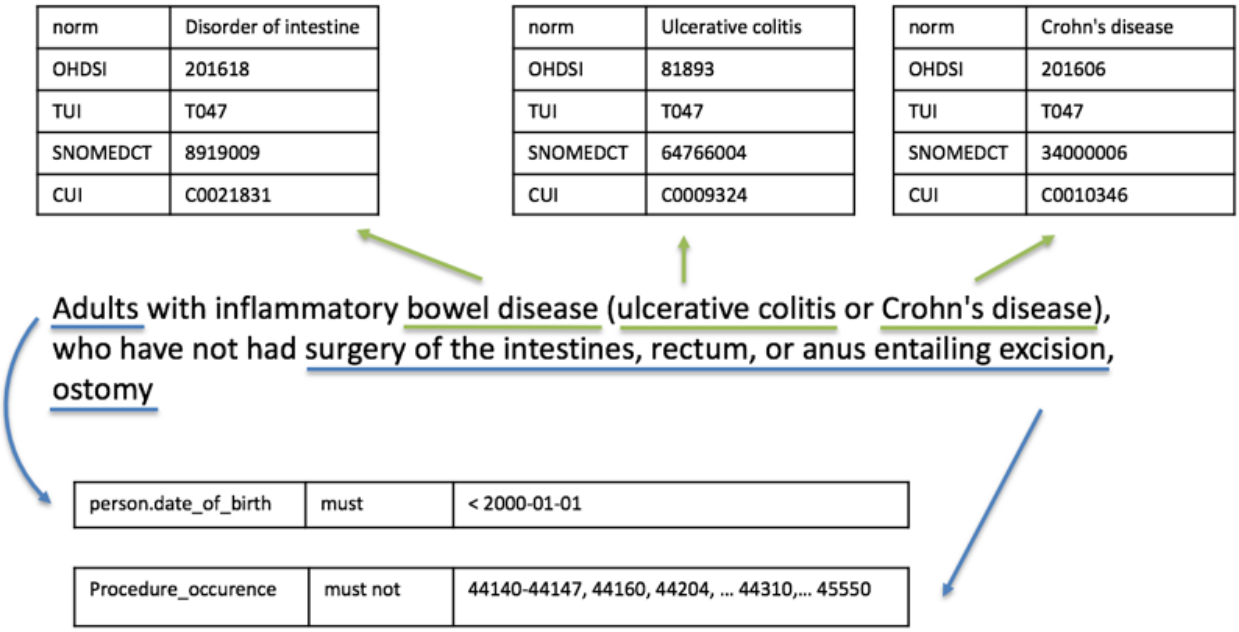
The clinical texts from Mayo Clinic electronic health records consisted of existing sections that provide brief descriptions of a specific perspective from a patient encounter, such as social history, diagnosis, and chief complaints. We chose to use the document sections to index clinical text for cohort retrieval based on the observation that while retrieval at a sentence level is insufficient for relevance judging relevance in the topic collections that we investigated, document-level retrieval may provide mostly irrelevant information.

Various common data model concepts were extracted via the dictionary lookup component as entity mentions such as *Drug*, *Procedure*, and *SignSymptom* with their concept identifiers (eg, UMLS concept unique identifiers) by cTAKES (Apache Software Foundation) [12] from clinical documents and subsequently indexed into Elasticsearch (Elasticsearch BV). In addition, the entity mention attributes such as *negation*, *certainty*, and *family history* are stored in the field *term_modifiers*.

Textual Query Formation

Natural language textual queries are fed into the same concept extraction pipeline used for indexing. Similarly, the normalized concepts and their associated attributes (eg, negation, certainty, experiencer, or status) are extracted from the textual query. Logical concepts such as *must* and *must not* are also used when generating queries from the text for further parsing and interpretation in the query backend. An example of the textual query modeling process is illustrated in Figure 3. In the query "Adults with inflammatory bowel disease (ulcerative colitis or Crohn's disease), who have not had surgery of the intestines, rectum, or anus entailing excision, ostomy," the natural language processing component can detect and normalize the raw mentions of "bowel disease," "ulcerative colitis," and "Crohn's disease" into various coding systems including OHDSI IDs, while the demographic information of "adults" and the list of surgeries can be manually added as structured data filters based on the date of birth and Current Procedural Terminology 4 codes.

Figure 3. Textual query modeling example. CUI: concept unique identifier; OHDSI: Observational Health Data Sciences and Informatics; SNOMEDCT: Systematized Nomenclature of Medicine—Clinical Terms; TUI: (semantic) type unique identifier.



User Interface

We developed a web-based user interface for CREATE, the details of which are described in Multimedia Appendix 1. All the information extracted is shown to the users by subject for potential insertion, modification, and deletion before query execution. Since the natural language processing component will suggest parsing results and map them into common data model concepts, the users are expected to focus on configuring the logistics between the extracted concepts and removing generic concepts (eg, UMLS concepts of *Drug* or *Treatment*), which do not require searching concepts among standard vocabularies from various sources and are not time consuming.

Retrieval Methods

CREATE uses Elasticsearch [35] as the search engine of the backend information retrieval component. Since Elasticsearch includes support for hierarchical queries of parent–child relations, the hierarchical index architecture shown in Figure 2 allows for significant flexibility in query strategies. Patients with a certain set of common data model concepts can be retrieved and filtered during the query execution by one of the structured fields (eg, encounter age), one of the unstructured fields (eg, whether the patient has sections containing common data model concepts from unstructured data), or both.

Given a document d and a textual query q, the set of common data model concepts extracted from q can be represented as $O = \{o_1,...,o_n\}$ where o is a common data model concept. The similarity score between d and o can then be represented as $s(d,o)$. The total score of each document for each query would then be defined as:

The first term on the right-hand side of the equation is the mean similarity of all common data model concepts in the query. The second term is the similarity between the document and the full-text query. In extreme use cases, the 2 terms can be weighted to place more emphasis on the contribution of either structured or unstructured data to the query. The patient-level similarity score is the mean of the top 100 document scores. The top rank threshold of 100 was selected based on our experiments on top 10, 20, 50, and 100 from test query results and may be subject to further tuning.

Functionality Assessment of CREATE

There are 2 aspects of the system design that require feasibility assessment by real-world implementation for clinical data repository.

First, the data mapping needs to be created specifically for each clinical data repository architecture. A site-dependent correlation between clinical data repository representation, OMOP Common Data Model tables, and extracted natural language processing concepts has to be established before the data can be indexed into CREATE. Second, retrieved results need to be assessed to validate that the proposed query modeling and retrieval methods can generate meaningful retrieval results.

The performance was measured using the mean *precision at 5* of 5 queries. As an evaluation of CREATE functionality, we randomly sampled 5 queries from a previously curated query collection [21,36] to evaluate CREATE through manual chart review. The structured query used manually transformed ICD-9-CM or ICD-10-CM codes. There was no ranking of relevance for the retrieved patients from structured electronic health record data, thus we randomly selected 5 patients from the relevant patients to be used as the top 5 in calculating the precision at 5. The top 5 patients from unstructured text queries and CREATE results were retrieved based on best match 25 [30]. A medical expert performed complete chart review on the top 5 patients for each retrieval cohort. The patient relevancy was scored into the 3 categories, *definitely relevant*, *partially*

relevant, and *not relevant*, by the medical expert. Definitely relevant, partially relevant, and not relevant were assigned scores of 1, 0.5, and 0, respectively, for precision at 5 calculations.

Results

We implemented CREATE as a feasibility assessment tool for the Mayo Clinic Biobank Rochester cohort, which is a large-scale institutionally funded research resource initiated in 2009 with blood, electronic health record, and patient-provided data on 45,613 Mayo Clinic Rochester patients who had consented to participate. This resource has been used in a wide array of over 250 health-related research and clinical studies [22]. In our experiments, we limited inclusion to patients with at least one clinical note in their electronic health record and extracted the corresponding structured data.

After data extraction, we investigated and compared the electronic health record system implementation at the Mayo Clinic to OMOP Common Data Model tables. During the data exploration stage, we found that the data elements under corresponding tables were generally straightforward to map; therefore, we show the mapping at the granularity of the table level. Table 1 shows our mapping of several OMOP Common Data Model tables to Mayo Clinic electronic health record tables. The mapping used to transform named entity mention types of the cTAKES-type system to common data model tables is also listed in Table 1.

Table 1. Table-level mapping between OMOP Common Data Model and Mayo Clinic electronic health records.

OMOP[a] Common Data Model and Mayo Clinic clinical data repository	Number of records	Vocabulary	Natural language processing cTAKES-type system
Person			
Demographics	45,613	—[b]	—
Condition			
Diagnosis	9,712,736	• ICD-9-CM[c] • ICD-10-CM[d]	• SignSymptom • DiseaseDisorder
Procedures	13,014,264	CPT[e]	Procedure
Measurement			
Lab	15,719,203	Local code system	Lab
Vital Signs	—	—	VitalSigns
Drug_Exposure			
DrugExposure	—	UMLS[f]	Medication
Note			
Clinical notes	68,198,499	—	—

[a]OMOP: Observational Medical Outcomes Partnership.

[b]There is no equivalent or no system is used for the equivalent concept.

[c]ICD-9-CM: International Classification of Diseases, Ninth revision, Clinical Modification.

[d]ICD-10-CM: International Classification of Diseases, Tenth revision, Clinical Modification.

[e]CPT: Current Procedural Terminology.

[f]UMLS: Unified Medical Language System.

Table 2 lists the detailed description of the 5 queries and the corresponding keywords used in the manual chart review process for judging patient relevance. The queries were different from the single condition criteria used to evaluate systems in some of the related work with regard to the level of detail, logic, and semantic complexity involved. The complete parsing results of the structured part of the queries and the CREATE query format specification can be found in Multimedia Appendix 2 and Multimedia Appendix 3, respectively.

Precision at 5 results are shown in Table 3. The overall comparison shows that CREATE, as a combination of systems using structured and unstructured electronic health record data, outperformed the systems based on using only one of structured or unstructured electronic health record data for full-text queries. For each query, CREATE performs at least as well as the systems using only structured or unstructured electronic health record data.

Table 2. The list of tested queries.

Query	Description	Keywords
1	Adults with inflammatory bowel disease (ulcerative colitis or Crohn's disease), who have not had surgery of the intestines, rectum, or anus entailing excision, ostomy	Ulcerative colitis, Crohn's disease, excision, ostomy, rectal prolapse, anal fistula, stricturoplasty resection
2	Adults 18-100 years old who have a diagnosis of hereditary hemorrhagic telangiectasia (HHT), which is also called Osler-Weber-Rendu syndrome.	Osler-Weber-Rendu syndrome, hereditary hemorrhagic telangiectasia
3	Children with localization-related (focal) epilepsy with simple or complex partial seizures diagnosed before 4 years old who have had an outpatient neurology visit.	Epilepsy, partial seizure, neurology
4	Adults 18-70 years old with rheumatoid arthritis currently treated with methotrexate who have never used a biologic disease-modifying antirheumatic drug (DMARD).	Rheumatoid arthritis biologic methotrexate abatacept, adalimumab, anakinra, certolizumab, etanercept, golimumab, infliximab, rituximab, tocilizumab, tofacitinib
5	Adults who have been treated with an angiotensin-converting-enzyme (ACE) inhibitor and developed an associated cough, consistent with ACE inhibitor–induced cough as an adverse effect of the medication.	Benazepril, Lotensin, Captopril, Enalapril, Vasotec, Fosinopril, Lisinopril, Prinivil, Zestril, Moexipril, Perindopril, Aceon, Quinapril, Accupril, Ramipril, Altace, Trandolapril, Mavik, cough, angiotensin-converting-enzyme (ACE) inhibitor

Table 3. Precision at 5 of sampled queries for electronic health record text.

Query	Structured	Unstructured	CREATE[a] (unstructured and structured combined)
1	0.8	0.6	0.8
2	0.7	1.0	1.0
3	0.3	0.5	0.8
4	0.7	0.7	1.0
5	0.2	0.9	0.9
Mean	0.54	0.74	0.90

[a]CREATE: Cohort Retrieval Enhanced by Analysis of Text from Electronic Health Records.

Discussion

Principal Findings

CREATE is a proof-of-concept for leveraging the combination of structured queries and information retrieval techniques to improve cohort retrieval performance while adopting the OMOP Common Data Model to enhance model portability. The evaluation of the implementation using sample queries supports our hypothesis that using a combination of structured and unstructured electronic health record data outperforms a single-source system in determining the relevance, from an input query, of any given patient electronic health record data for a particular clinical application. CREATE was designed to improve the efficiency of judging patient relevance, by shifting from human-query judgment (pull) to system-feed judgment (push).

Intuitively, the nature of the queries and how the query-related data are presented in the clinical data repository significantly impact the performance of the data source queried (ie, structured, unstructured, and combined). For instance, one of the major concepts in query 5 is treatment with angiotensin-converting enzyme (ACE) inhibitors. It is effective for information retrieval methods on unstructured text to select patients with ACE inhibitor–related cough, as the keywords *ACE inhibitor* and *cough* usually co-occur in clinical text contexts as adverse drug events. In contrast, it is challenging for structured data queries in this experiment. In our clinical data repository, the medication information is present only as semistructured text generated by computerized provider order entry without normalization into structured data. Therefore, there is no reliable way to obtain the relevant cohort purely on structured data, which leads to very low relevancy of the retrieved cohort. Such a limitation is usually not critical when unstructured text are queried, because most of the clinical data are either presented or summarized in clinical notes.

However, when querying on a cohort with age or gender criteria, querying solely on unstructured data cannot work effectively. For example, even when the age is mentioned in query 3, all the retrieved patients are adult patients rather than the expected pediatric patients. This is caused by the lack of the extraction the dates and ages from narrative texts, which is not a trivial information extraction task. To build a reliable query system for unstructured texts without providing metadata, such as date of birth or age at encounter, usually requires corpus-dependent engineering efforts to extract the dates and ages from narrative text.

Limitations

This study has multiple limitations that may offer directions for our future work. Our current functionality test is based on 5-query precision at 5 by one annotator, which is not sufficient

to cover all cohort retrieval cases and longitudinal patient condition scenarios. Though we acknowledge that a larger number of queries on a fully-annotated patient cohort from annotators and adjudicators would be very helpful in evaluating the performance of the system, it is time consuming to judge complete patient history, especially for negated conditions and treatments (eg, to check if the patient does not have a certain disorder or procedure). With the system in production, the feedback of each study leveraging the system can be then retained and analyzed for more comprehensive statistics of the performance of the system.

When processing concepts without a global coding system, concept mapping, such as that used in our solution, relies on the output of natural language processing algorithms. Although it is a fast and straightforward solution, current natural language processing tools cannot achieve the same level of accuracy as human assigned codes. Complete mapping from a local vocabulary requires extensive human efforts with data quality assurance [37], thus it was not feasible within the scope of this study. A solution for this issue is to utilize value set repositories to manage the concepts. Though a one-to-one mapping may not be found in all semantic spaces, value set repositories can provide a systematic way to manage the concept sets in collections or aggregations [38].

There are also several potential approaches to further improve the information retrieval component in this system's framework. We only used the out-of-box query algorithms to measure the patient similarity and rank the relevancy in this study. More advanced information retrieval methods can be applied to the queries such as case-based reasoning [39-41], pseudo relevance feedback [42], and different ranking models [43,44]. Though the equal weights of common data model concepts and raw text provide information from both sides, the weights can be tuned to meet different retrieval perspectives and demands.

Conclusion

We developed CREATE, an end-to-end patient-level information retrieval system, with the ability to query both structured and unstructured data by leveraging the OMOP Common Data Model. Implementation and functionality assessment on Mayo Clinic Biobank demonstrated that CREATE outperforms cohort retrieval systems that use only one of either structured or unstructured data in complex textual cohort queries. The source code of the CREATE can be found at Multimedia Appendix 4.

In the future, we will refine the evaluation process by adding more query topics and larger cohort of manual chart reviews. An active learning component will be added to the system to enable human-in-the-loop analysis on the system-screened cohort to further improve the efficiency of relevance judgment. In doing so, both machine learning–based or rule-based cohort identification algorithms could be deployed and evaluated in real time. This could potentially then be extended to an active-learning cohort-identification framework [45].

Acknowledgments

We sincerely thank Donna Ihrke who annotated the query corpus. The work was supported by the National Institutes of Health (grants R01LM011934, R01EB19403, R01LM11829, and U01TR02062). The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Web-based graphical user interface of CREATE.
[DOCX File , 658 KB - medinform_v8i10e17376_app1.docx]

Multimedia Appendix 2
Parsing results of CREATE from textual queries.
[DOCX File , 18 KB - medinform_v8i10e17376_app2.docx]

Multimedia Appendix 3
Format specification for CREATE queries.
[DOCX File , 16 KB - medinform_v8i10e17376_app3.docx]

Multimedia Appendix 4
Source code of CREATE.
[DOCX File , 12 KB - medinform_v8i10e17376_app4.docx]

References

1. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association 2013 Jan 01;20(1):144-151. [doi: 10.1136/amiajnl-2011-000681]
2. Accrual to Clinical Trials (ACT) Network. Clinical and Translational Science Institute. URL: https://www.ctsi.umn.edu/consultations-and-services/multi-site-study-support/accrual-clinical-trials-act-network [accessed 2020-08-20]
3. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 2013 Jun 6;15(10):761-771. [doi: 10.1038/gim.2013.72]
4. PCORnet: the National Patient-Centered Clinical Research Network. URL: https://pcornet.org/clinical-research-network/ [accessed 2020-08-20]
5. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform 2017 Aug;26(1):38-52 [FREE Full text] [doi: 10.15265/IY-2017-007] [Medline: 28480475]
6. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018 Oct 10;562(7726):203-209. [doi: 10.1038/s41586-018-0579-z]
7. Liu S, Wang Y, Hong N, Shen F, Wu S, Hersh W, et al. On Mapping Textual Queries to a Common Data Model. 2017 Presented at: 2017 IEEE International Conference on Healthcare Informatics (ICHI); 23-26 Aug. 2017; Park City, UT, USA p. 21-25 URL: https://doi.org/10.1109/ICHI.2017.63
8. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. Journal of Biomedical Informatics 2018 Jan;77:34-49. [doi: 10.1016/j.jbi.2017.11.011]
9. Wu PY, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. -Omic and Electronic Health Record Big Data Analytics for Precision Medicine. IEEE Trans Biomed Eng 2017 Feb;64(2):263-273 [FREE Full text] [doi: 10.1109/TBME.2016.2573285] [Medline: 27740470]
10. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011 Aug 24;306(8):848-855. [doi: 10.1001/jama.2011.1204] [Medline: 21862746]
11. Maddox TM, Albert NM, Borden WB, Curtis LH, Ferguson TB, Kao DP, et al. The Learning Healthcare System and Cardiovascular Care: A Scientific Statement From the American Heart Association. Circulation 2017 Apr 04;135(14). [doi: 10.1161/cir.0000000000000480]
12. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010 Sep 01;17(5):507-513. [doi: 10.1136/jamia.2009.001560]
13. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010 May 01;17(3):229-236. [doi: 10.1136/jamia.2009.002733]
14. Pradhan S, Elhadad N, Chapman W, Manandhar S, Savova G. SemEval-2014 Task 7: Analysis of Clinical Text. 2014 Presented at: 8th International Workshop on Semantic Evaluation (SemEval 2014); August 23-24, 2014; Dublin, Ireland p. 54-62. [doi: 10.3115/v1/s14-2007]
15. Pradhan S, Elhadad N, South B, Martinez D, Christensen L, Vogel A, et al. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. In: CLEF (Working Notes). 2013 Presented at: ShARe/CLEF eHealth Evaluation Lab; September 23-26, 2013; Valencia, Spain.
16. Carroll R, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc 2012 Jun;19(e1):e162-e169 [FREE Full text] [doi: 10.1136/amiajnl-2011-000583] [Medline: 22374935]
17. Mehrabi S, Krishnan A, Roch AM, Schmidt H, Li D, Kesterson J, et al. Identification of Patients with Family History of Pancreatic Cancer--Investigation of an NLP System Portability. Stud Health Technol Inform 2015;216:604-608 [FREE Full text] [Medline: 26262122]
18. Goodwin TR, Harabagiu SM. Multi-modal Patient Cohort Identification from EEG Report and Signal Data. In: Annual Symposium proceedings. 2016 Presented at: American Medical Informatics Association; Nov 12-16, 2016; Chicago, IL, USA p. 1794-1803.
19. Hanauer DA, Mei Q, Law J, Khanna R, Zheng K. Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). Journal of Biomedical Informatics 2015 Jun;55:290-300. [doi: 10.1016/j.jbi.2015.05.003]
20. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016 Feb 05;23(5):1007-1015. [doi: 10.1093/jamia/ocv180]
21. Wu S, Liu S, Wang Y, Timmons T, Uppili H, Bedrick S, et al. Intrainstitutional EHR collections for patient-level information retrieval. Journal of the Association for Information Science and Technology 2017 Sep 18;68(11):2636-2648. [doi: 10.1002/asi.23884]
22. Olson JE, Ryu E, Johnson KJ, Koenig BA, Maschke KJ, Morrisette JA, et al. The Mayo Clinic Biobank: A Building Block for Individualized Medicine. Mayo Clinic Proceedings 2013 Sep;88(9):952-962. [doi: 10.1016/j.mayocp.2013.06.006]

23. Biron P, Metzger M, Pezet C, Sebban C, Barthuet E, Durand T. An information retrieval system for computerized patient records in the context of a daily hospital practice: the example of the Léon Bérard Cancer Center (France). Appl Clin Inform 2017 Dec 20;05(01):191-205. [doi: 10.4338/aci-2013-08-cr-0065]
24. Hanauer DA, Mei Q, Law J, Khanna R, Zheng K. Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). Journal of Biomedical Informatics 2015 Jun;55:290-300. [doi: 10.1016/j.jbi.2015.05.003]
25. Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, et al. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. Journal of Biomedical Informatics 2018 Apr;80:52-63. [doi: 10.1016/j.jbi.2018.02.019]
26. Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc 2018 May 01;25(5):530-537 [FREE Full text] [doi: 10.1093/jamia/ocx160] [Medline: 29361077]
27. FHIR (Fast Healthcare Interoperability Resources). FHIR Overview. URL: https://www.hl7.org/fhir/overview.html
28. Stubbs A, Filannino M, Soysal E, Henry S, Uzuner Ö. Cohort selection for clinical trials: n2c2 2018 shared task track 1. J Am Med Inform Assoc 2019 Nov 01;26(11):1163-1171. [doi: 10.1093/jamia/ocz163] [Medline: 31562516]
29. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: A literature review. Journal of Biomedical Informatics 2010 Jun;43(3):451-467. [doi: 10.1016/j.jbi.2009.12.004]
30. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. Journal of the American Medical Informatics Association 2011 Dec 01;18(Supplement 1):i116-i124. [doi: 10.1136/amiajnl-2011-000321]
31. Kang T, Zhang S, Tang Y, Hruby GW, Rusanov A, Elhadad N, et al. EliIE: An open-source information extraction system for clinical trial eligibility criteria. Journal of the American Medical Informatics Association 2017;24(6):1062-1071. [doi: 10.1093/jamia/ocx019]
32. OMOP Common Data Model v5.3.1. GitHub. URL: https://github.com/OHDSI/CommonDataModel/tree/v5.3.1 [accessed 2020-08-20]
33. Metathesaurus. UMLS. URL: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html [accessed 2020-08-20]
34. ATHENA standardized vocabularies - OHDSI. URL: https://www.ohdsi.org/analytic-tools/athena-standardized-vocabularies/ [accessed 2020-08-20]
35. Elasticsearch: The Official Distributed Search & Analytics Engine. URL: https://www.elastic.co/elasticsearch/ [accessed 2020-08-20]
36. Wang Y, Wen A, Liu S, Hersh W, Bedrick S, Liu H. Test collections for electronic health record-based clinical information retrieval. JAMIA Open 2019 Oct;2(3):360-368 [FREE Full text] [doi: 10.1093/jamiaopen/ooz016] [Medline: 31709390]
37. Huser V, DeFalco FJ, Schuemie M, Ryan PB, Shang N, Velez M, et al. Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets. eGEMs 2016 Nov 30;4(1):24. [doi: 10.13063/2327-9214.1239]
38. Peterson KJ. Mining Hierarchies Similarity Clusters from Value Set Repositories. In: AMIA Annu Symp Proc. 2017 Presented at: AMIA Annual Symposium; Nov 4-8, 2017; Washington, DC, USA.
39. Marling C, Whitehouse P. Case-Based Reasoning in the Care of Alzheimer's Disease Patients. In: Aha DW, Watson I, editors. International Conference on Case-Based Reasoning. Lecture Notes in Computer Science. Berlin: Springer; 2001:702-715.
40. van den Branden M, Wiratunga N, Burton D, Craw S. Integrating case-based reasoning with an electronic patient record system. Artificial Intelligence in Medicine 2011 Feb;51(2):117-123. [doi: 10.1016/j.artmed.2010.12.004]
41. Miotto R, Weng C. Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. Journal of the American Medical Informatics Association 2015 Mar 13;22(e1):e141-e150. [doi: 10.1093/jamia/ocu050]
42. Buckley C, Singhal A, Mitra M, Salton G. New Retrieval Approaches Using SMART: TREC 4. In: The Fourth Text REtrieval Conference (TREC-4).: NIST Special Publication 500-236; 1995 Presented at: Fourth text retrieval conference (TREC-4); November 1-3; Gaithersburg, Maryland p. 25-48 URL: https://trec.nist.gov/pubs/trec4/papers/Cornell_trec4.ps.gz
43. Cao Z, Wei F, Dong L, Li S, Zhou M. Ranking with recursive neural networks and its application to multi-document summarization. 2015 Presented at: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence; 2015; Austin, TX, USA p. 2153-2159.
44. Chen J, Yu H. Unsupervised ensemble ranking of terms in electronic health record notes based on their importance to patients. Journal of Biomedical Informatics 2017 Apr;68:121-131. [doi: 10.1016/j.jbi.2017.02.016]
45. Chen Y, Lasko TA, Mei Q, Denny JC, Xu H. A study of active learning methods for named entity recognition in clinical text. Journal of Biomedical Informatics 2015 Dec;58:11-18. [doi: 10.1016/j.jbi.2015.09.010]

Abbreviations

CREATE: Cohort Retrieval Enhanced by Analysis of Text from Electronic Health Records

EMERSE: Electronic Medical Record Search Engine
ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification
ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification
OHDSI: Observational Health Data Sciences and Informatics
OMOP: Observational Medical Outcomes Partnership
SNOMED: Systematized Nomenclature of Medicine
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 10.12.19; peer-reviewed by T Goodwin, J Lalor, S Meystre; comments to author 29.02.20; revised version received 04.06.20; accepted 28.07.20; published 06.10.20.

Please cite as:
Liu S, Wang Y, Wen A, Wang L, Hong N, Shen F, Bedrick S, Hersh W, Liu H
Implementation of a Cohort Retrieval System for Clinical Data Repositories Using the Observational Medical Outcomes Partnership Common Data Model: Proof-of-Concept System Validation
JMIR Med Inform 2020;8(10):e17376
URL: http://medinform.jmir.org/2020/10/e17376/
doi:10.2196/17376
PMID:33021486

Original Paper

Automated Cluster Detection of Health Care–Associated Infection Based on the Multisource Surveillance of Process Data in the Area Network: Retrospective Study of Algorithm Development and Validation

Yunzhou Fan[1*], PhD; Yanyan Wu[1*], PhD; Xiongjing Cao[1], MSc; Junning Zou[1], BSc; Ming Zhu[1], MSc; Di Dai[1], BA; Lin Lu[1], BA; Xiaoxv Yin[2*], PhD; Lijuan Xiong[1], PhD

[1]Department of Nosocomial Infection Management, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[2]School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[*]these authors contributed equally

Corresponding Author:
Lijuan Xiong, PhD
Department of Nosocomial Infection Management, Union Hospital
Tongji Medical College
Huazhong University of Science and Technology
1277 JieFang Avenue
Wuhan, 430022
China
Phone: 1 86 02785726293
Email: lijuanxiong2016@126.com

Abstract

Background: The cluster detection of health care–associated infections (HAIs) is crucial for identifying HAI outbreaks in the early stages.

Objective: We aimed to verify whether multisource surveillance based on the process data in an area network can be effective in detecting HAI clusters.

Methods: We retrospectively analyzed the incidence of HAIs and 3 indicators of process data relative to infection, namely, antibiotic utilization rate in combination, inspection rate of bacterial specimens, and positive rate of bacterial specimens, from 4 independent high-risk units in a tertiary hospital in China. We utilized the Shewhart warning model to detect the peaks of the time-series data. Subsequently, we designed 5 surveillance strategies based on the process data for the HAI cluster detection: (1) antibiotic utilization rate in combination only, (2) inspection rate of bacterial specimens only, (3) positive rate of bacterial specimens only, (4) antibiotic utilization rate in combination + inspection rate of bacterial specimens + positive rate of bacterial specimens in parallel, and (5) antibiotic utilization rate in combination + inspection rate of bacterial specimens + positive rate of bacterial specimens in series. We used the receiver operating characteristic (ROC) curve and Youden index to evaluate the warning performance of these surveillance strategies for the detection of HAI clusters.

Results: The ROC curves of the 5 surveillance strategies were located above the standard line, and the area under the curve of the ROC was larger in the parallel strategy than in the series strategy and the single-indicator strategies. The optimal Youden indexes were 0.48 (95% CI 0.29-0.67) at a threshold of 1.5 in the antibiotic utilization rate in combination–only strategy, 0.49 (95% CI 0.45-0.53) at a threshold of 0.5 in the inspection rate of bacterial specimens–only strategy, 0.50 (95% CI 0.28-0.71) at a threshold of 1.1 in the positive rate of bacterial specimens–only strategy, 0.63 (95% CI 0.49-0.77) at a threshold of 2.6 in the parallel strategy, and 0.32 (95% CI 0.00-0.65) at a threshold of 0.0 in the series strategy. The warning performance of the parallel strategy was greater than that of the single-indicator strategies when the threshold exceeded 1.5.

Conclusions: The multisource surveillance of process data in the area network is an effective method for the early detection of HAI clusters. The combination of multisource data and the threshold of the warning model are 2 important factors that influence the performance of the model.

(JMIR Med Inform 2020;8(10):e16901) doi:10.2196/16901

KEYWORDS

health care–associated infection; cluster detection; early warning; multi sources surveillance; process data

Introduction

Health care–associated infections (HAIs) are a socially sensitive and important public health issue that threatens patient safety, prolongs hospital stays, and increases economic burden. The incidence of HAIs in developed countries is 2%-6%, and in developing countries it is 12.6%-18.9% [1]. In China, the extra medical expenses per HAI patient varied from 9725 to 18,909 RMB (US $1427 to 2775) [2], and the total medical costs due to HAI have increased by nearly 70% [3]. Outbreaks are the main manifestation of the risk of HAIs, as HAIs are contagious, and approximately 2%-10% of HAI cases occur in the form of outbreaks [4]. In the past 40 years, there have been 465 major HAI outbreak events in China, with an average of 11.6 outbreak events annually reported by the media [5,6]. Because a significant number of HAI outbreaks have not been detected or reported in a timely manner, the severity of HAI outbreaks in China is likely to be seriously underestimated.

The key to establishing a methodology for HAI prevention and control is to develop a reliable outbreak warning system based on surveillance. To identify HAI outbreaks, HAI clusters must first be detected and then confirmed through epidemiological investigations. Therefore, detecting aggregated HAI cases is crucial to establishing a sound early warning system for HAI outbreaks. Traditional HAI surveillance is a form of passive monitoring, which relies on case reports by clinicians. However, owing to the compliance of clinicians with case reporting and the delay in HAI diagnosis, the timeliness of surveillance and warning for HAI outbreaks is limited.

In this paper, process data refer to the continuous, traceable, and basic information on patients who are admitted to hospitals; these data can be collected automatically by a search engine based on the local area network of the hospital. The proposed process data surveillance would be a form of active monitoring, which would not rely on delayed case reports. Therefore, the use of infection-related process data to detect the aggregation of HAI cases is likely to be a reliable method of early warning for HAI outbreaks. In recent years, the rapid development of information technology has led to a noticeable improvement in process data collection. Consequently, automated surveillance using process data related to infections has become a widely researched topic among the researchers of early warning systems for HAI outbreaks.

Recent studies have used a large amount of process data related to infections to identify HAI clusters [7-12]. However, surveillance that relied on a single indicator of process data limited the accuracy of HAI cluster detection because a solo process indicator was not sufficiently specific to reflect the occurrence and progress of infections. Some studies have confirmed that multisource surveillance for health-related data could improve the accuracy and timeliness of outbreak warning for infectious diseases [13,14]. Therefore, we considered that if a variety of process indicators related to infections could be combined for surveillance, the accuracy of the detection of HAI clusters could also be improved.

In a previous study [15], we assessed the performance and feasibility of automated cluster detection of multidrug-resistant organism–related HAIs using data on antibiotic use. In this study, we conducted an integrated surveillance of 3 process indicators using an electronic records information system based on the local area network of the hospital, including the antibiotic utilization rate in combination, inspection rate of bacterial specimens, and positive rate of bacterial specimens. We then analyzed the different combinations of the warning signals of these multisource process surveillance data to verify their early warning capability for HAI cluster detection.

Methods

Study Design and Setting

This was a retrospective observational study. The time series data of HAI incidences and the 3 indicators of process data were collected from 4 HAI high-risk units in Wuhan Union Hospital (WHUH). WHUH is a tertiary hospital in Wuhan, China, with a 5000-bed capacity. The process data, in this study, included the antibiotic utilization rate in combination, inspection rate of bacterial specimens, and positive rate of bacterial specimens from the 4 units with the highest HAI incidences. All data presented are from January 1, 2017, to June 28, 2019. Indicators were collected weekly at the unit level.

Surveillance and demographic data are available in the Real-Time Nosocomial Infection Surveillance System (RT-NISS) database. Briefly, the RT-NISS is seamlessly connected with several electronic information systems, including the hospital information system, laboratory information system, and other information systems in the local area network of the hospital. The infection-related process data are extracted and stored in real time in the database. The details of the RT-NISS database have been previously described [16].

Indicators of Process Data

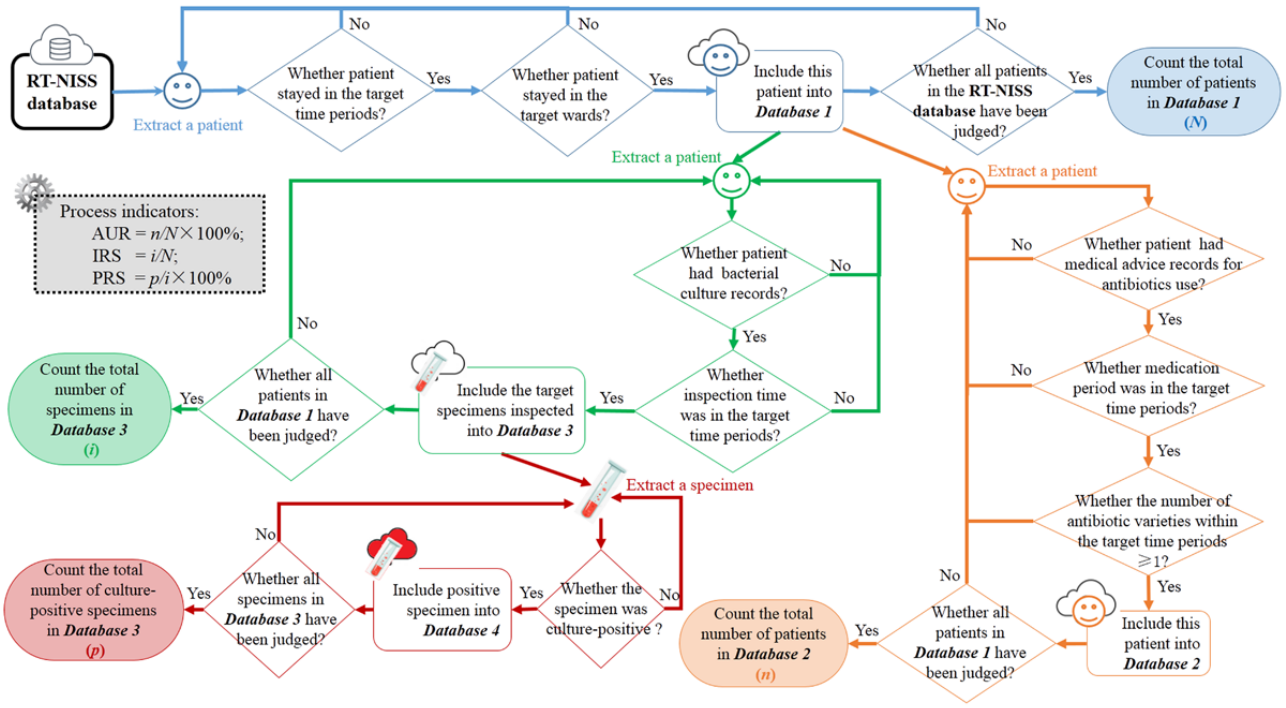
All indicators in this study were obtained from the RT-NISS database. The process data associated with antibiotic use and bacterial culture were automatically extracted from data sets containing doctor's advice and nursing records by the RT-NISS using web mining and web crawler technology. The 3 process data indicators in this study were calculated weekly within each unit.

The antibiotic utilization rate in combination was determined to be the proportion of the number of admitted patients who used more than 1 antibiotic (n) divided by the total number of admitted patients (N), that is, antibiotic utilization rate in combination = n/N × 100%; the inspection rate of bacterial specimens was calculated as the number of specimens that were collected for bacterial testing (i) divided by the number of

admitted patients (N), that is, inspection rate of bacterial specimens = i/N; the positive rate of bacterial specimens was calculated as the number of positive specimens with cultured bacteria (p) divided by the number of specimens collected for bacterial testing (i), that is, positive rate of bacterial specimens = p/i × 100%.

Data on the prescribed oral and intravenous antibiotics were collected, while topical antibiotics were excluded from the data collection. The sputum of bacterial culture included throat secretion, urine, blood, stool, pleural effusion, cerebrospinal fluid, ascites, and venous catheter, among others. Repeated samples from each individual were excluded. The data extraction process of the variables (N, n, i, and p) used to calculate the process indicators is shown in Figure 1.

Figure 1. The flow diagram of data extraction process of the variables used to calculate the process indicators. RT-NISS: Real-time nosocomial infection surveillance system; AUR: Antibiotic utilization rate in combination; IRS: Inspection rate of bacterial specimens; PRS: Positive rate of bacterial specimens.



Identification of HAI Cases

HAI cases were identified according to the diagnostic criteria for HAIs, which were issued by the Ministry of Health of China in 2001 [17]. The HAI case findings were documented weekly by a hospital infection management team. The hospital infection management team comprised clinicians, nurses, and full-time infection control practitioners. All members within the hospital infection management team independently reviewed the clinical records of the patients to include reports of illness, microbiology data, antibiotic data, imaging reports, and results of clinical laboratory tests, and HAI cases were identified after the hospital infection management team members reached a consensus. The weekly HAI incidence was measured as the number of new HAI cases in a week divided by the total number of inpatients in that week.

Warning Detection Model

In this study, the time series data sets of each surveillance indicator were analyzed using the Shewhart warning model, which is a common statistical process control for detecting clusters. We used a 4-week moving average of time series data in the Shewhart model, considering the inpatient's average length of hospitalization and the epidemiologic characteristics of infected patients. We then used the data from the nearest 4 weeks before the current week as the dynamic warning baseline of the Shewhart model. Finally, the Shewhart warning statistics (S_t) for each week were calculated using the mean and SD of the dynamic baseline data sets according to the following formula:

$$S_t=(X_t - \mu_t)/\sigma_t$$

where X_t is the observation value at week t; μ_t and σ_t are the mean and SD of the observation values for the warning baseline from week t–4 to week t–1, respectively. The warning signal at week t was generated when S_t exceeded the threshold.

An HAI cluster is considered to exist when a group of HAIs occurs closely together in a health care unit, so the previous warning threshold of an HAI cluster was based on the statistical variations in the frequency. The Shewhart model with a threshold of 2.0 was used for detecting HAI clusters in WHUH according to the Guideline of Control of Health Care-Associated Infection Outbreak [18]. This implies that a warning signal for an HAI cluster was generated when the 4-week moving average of HAI incidence at the current week exceeded the mean plus 2 SDs of the past 4 weeks. We used 51 thresholds (0.0-5.0, steps of 0.1) to detect process data clusters to explore the optimal

threshold of the Shewhart warning model for process data warning.

Warning Strategies for Process Data

We designed 5 warning strategies of process indicators based on the combination of 3 single-indicator warning strategies: (1) antibiotic utilization rate in combination only, (2) inspection rate of bacterial specimens only, and (3) positive rate of bacterial specimens only, and 2 multi-indicator warning strategies, (4) antibiotic utilization rate in combination + inspection rate of bacterial specimens + positive rate of bacterial specimens in parallel, and (5) antibiotic utilization rate in combination + inspection rate of bacterial specimens + positive rate of bacterial specimens in series. The parallel warning signal is generated once any subindicator generates a signal, and the series warning signal is generated only when all subindicators generate signals during the same period.

Comparison of Warning Signals of Process Data With HAI Incidence

We used the consistency of warning signals between the HAI incidence and process data to evaluate the warning performance for HAI cluster detection. The warning signals of the process data were considered as the test and those of the HAI incidences as references. The early warning signal was defined as the signal of process data generated earlier than the signal of HAI incidence within the 4-week period. Accordingly, we calculated the sensitivity, specificity, and Youden index under each threshold of process data for the early detection of HAI clusters. Furthermore, the receiver operating characteristic (ROC) curve of the process data for the early detection of the signals of HAI clusters was plotted using sensitivity and 1–specificity under 51 thresholds (0.0 to 5.0, steps of 0.1). Youden index was used to evaluate the comprehensive warning performance for HAI cluster detection under each threshold.

Sensitivity = Number of HAI cluster signals detected by the early warning signals/Total number of HAI cluster signals

Specificity = Number of weeks that signal generated neither in HAI incidence nor in process indicators/Number of weeks that no signal generated in the HAI incidence

Youden index = Sensitivity + Specificity–1

Statistical Analysis

The one-way analysis of variance was used to compare the differences between the mean values, and a chi-square test was used to compare the differences between the proportions among the 4 independent units. A statistical evaluation of Youden index among the warning strategies in each threshold was performed using the paired samples *t* test. A *P*-value of .05 or less was considered statistically significant in all analyses.

Results

Demographic Characteristics

A total of 23,119 patients were admitted to the 4 HAI high-risk units in WHUH during the study period. The hospital infection management team diagnosed 1503 HAI cases. The HAI incidence in these high-risk units ranged from 5.36% (462/8618 patients) to 9.06% (316/3489 patients). Statistically significant differences were observed in all demographic characteristics of patients among the 4 HAI high-risk units (Table 1).

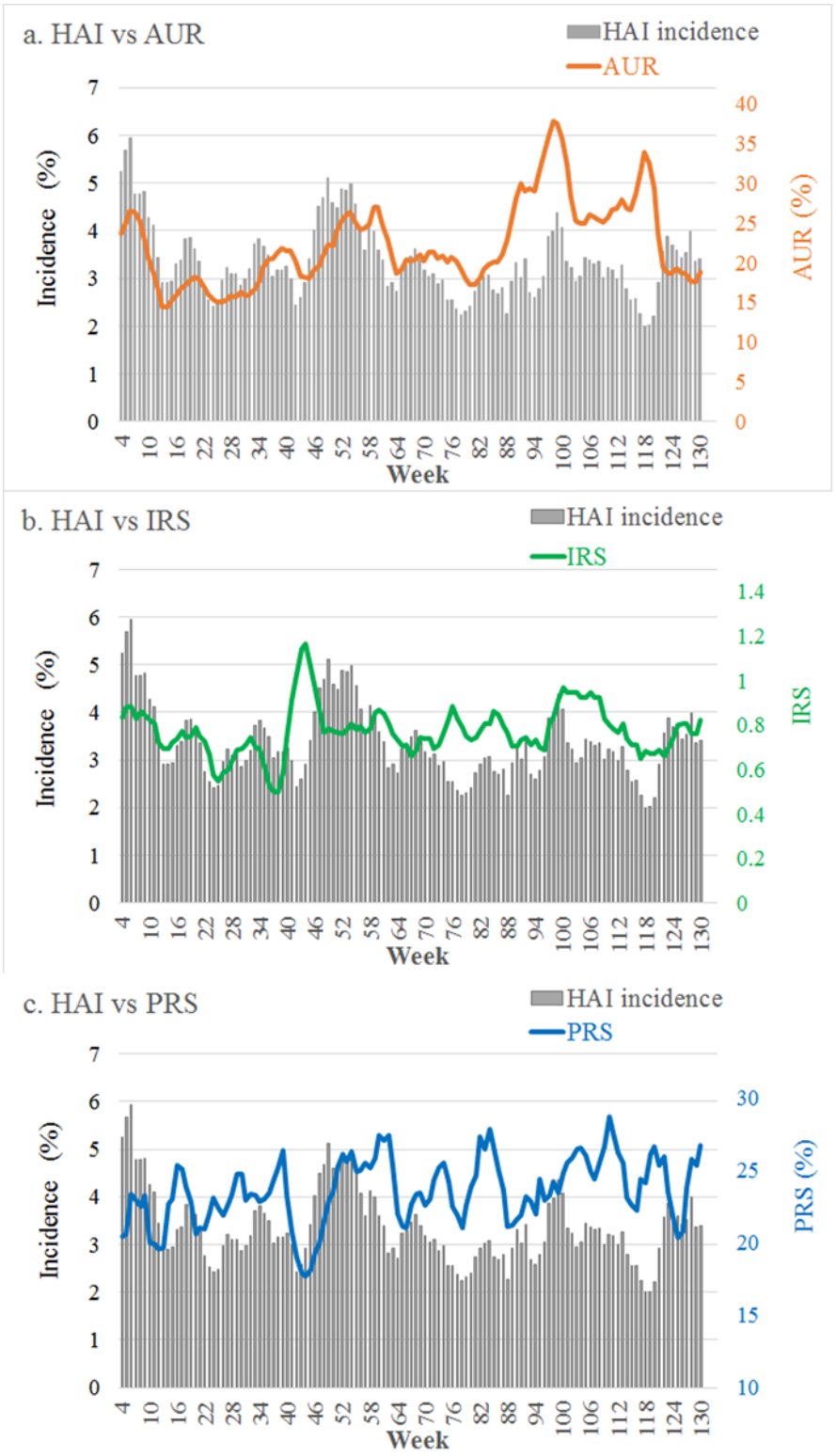
Table 1. Demographic characteristics of inpatients in the 4 high-risk units in Wuhan Union Hospital during the surveillance period.

Characteristics	Total	High-risk unit				*P* value
		Unit 1	Unit 2	Unit 3	Unit 4	
Participants (N)	23,119	8618	7414	3598	3489	
Male, n (%)	13,153 (56.9)	4679 (54.3)	4284 (57.8)	2117 (58.8)	2073 (59.4)	<.001
Age in years, mean (SD)	44.9 (21.7)	47.3 (16.3)	34.1 (25.9)	51.1 (15.1)	55.6 (19.5)	<.001
Hospitalization days, mean (SD)	19.6 (26.2)	15.2 (32.3)	24.1 (15.5)	16.1 (13.1)	24.6 (34.1)	<.001
Surgical procedure, n (%)	13,747 (59.5)	3459 (40.1)	6386 (86.1)	1503 (41.8)	2399 (68.8)	<.001
Mechanical ventilation, n (%)	10,496 (45.4)	371 (4.3)	6828 (92.1)	227 (6.3)	3077 (88.2)	<.001
Central venous catheter, n (%)	8485 (36.7)	353 (4.1)	6643 (89.6)	450 (12.5)	1026 (29.4)	<.001
Urinary catheter, n (%)	17,779 (76.9)	5378 (62.4)	7051 (95.1)	1979 (55.0)	3370 (96.5)	<.001
Health care–associated infection, n (%)	1503 (6.5)	462 (5.4)	418 (5.6)	307 (8.5)	316 (9.1)	<.001
Antibiotics used, n (%)	18,124 (78.4)	4736 (55.0)	7214 (97.3)	2749 (76.4)	3425 (98.2)	<.001
Antibiotic days, mean (SD)	10.7 (11.3)	5.6 (8.7)	13.2 (8.6)	10.9 (12.1)	17.6 (15.1)	<.001
Antibiotics used in combination, n (%)	6356 (27.5)	1010 (11.7)	1895 (25.6)	1166 (32.4)	2285 (65.5)	<.001
Antibiotic days in combination used, mean (SD)	2.7 (6.6)	1.0 (3.5)	2.2 (5.2)	3.6 (8.1)	7.2 (10.3)	<.001
Microbiological test, n (%)	6040 (26.1)	1415 (16.4)	1596 (21.5)	1262 (35.1)	1767 (50.6)	<.001
Microbiological test with positive result, n (%)	3129 (13.5)	728 (8.4)	677 (9.1)	632 (17.6)	1092 (31.3)	<.001
Microbiological specimens	43,070	11,785	8685	5647	16,953	
Positive, n (%)	10,086 (23.4)	2600 (22.1)	1551 (17.9)	1927 (34.1)	4008 (23.6)	<.001
Isolated strains	11,808	3070	1679	2326	4733	
Acinetobacter baumannii, n (%)	3430 (29.0)	769 (25.0)	456 (27.2)	319 (13.7)	1886 (39.8)	<.001
Staphylococcus aureus, n (%)	1683 (14.3)	581 (18.9)	88 (5.2)	535 (23.0)	479 (10.1)	<.001
Pseudomonas aeruginosa, n (%)	1214 (10.3)	362 (11.8)	219 (13.0)	126 (5.4)	507 (10.7)	<.001
Klebsiella pneumonia, n (%)	1076 (9.1)	326 (10.6)	166 (9.9)	322 (13.8)	262 (5.5)	<.001
Saccharomyces albicans, n (%)	792 (6.7)	148 (4.8)	159 (9.5)	176 (7.6)	309 (6.5)	<.001
Escherichia coli, n (%)	624 (5.3)	119 (3.9)	78 (4.6)	223 (9.6)	204 (4.3)	<.001
Other, n (%)	2989 (25.3)	765 (24.9)	513 (30.6)	625 (26.9)	1086 (22.9)	<.001

Surveillance and Cluster Detection

The time series charts of the 3 process indicators and HAI incidences for all units are shown in Figure 2, as well as in Multimedia Appendix 1. The fluctuations of the time series in the process data are generally synchronous with those in HAI incidence. For the HAI cluster detection using the Shewhart warning model in each unit, there were 20 signals generated in unit 1, 16 signals in unit 2, 18 signals in unit 3, and 16 signals in unit 4. These HAI cluster signals were compared with those of the process data warning at each threshold. An example of signal comparison at the threshold of 2.0 is shown in Multimedia Appendix 1.

Figure 2. The time-series charts comparison of process data with HAI incidence in all surveillance units. AUR: Antibiotic utilization rate in combination; IRS: Inspection rate of bacterial specimens; PRS: Positive rate of bacterial specimens.



Warning Detection Evaluation

According to the definition of early warning signals, the ROC curves of 5 warning strategies for early detected HAI cluster signals were plotted using scattered points of 51 thresholds. Figure 3 depicts the overall ROC curves of process data warning for detecting HAI cluster signals across the 4 units. Generally, all ROC curves are located above the standard line, and the area under the ROC curve is larger in the parallel warning strategy than in the single-indicator warning strategies and the series warning strategy.

The optimal Youden index for the early detection of HAI cluster signals was higher in the parallel warning strategy than in any other warning strategies. Specifically, the optimal Youden indexes were 0.48 (95% CI 0.29-0.67) at a threshold of 1.5 for antibiotic utilization rate in combination only, 0.49 (95% CI

0.45-0.53) at a threshold of 0.5 for inspection rate of bacterial specimens only, 0.50 (95% CI 0.28-0.71) at a threshold of 1.1 for positive rate of bacterial specimens only, 0.63 (95% CI 0.49-0.77) at a threshold of 2.6 in the parallel strategy, and 0.32 (95% CI 0.00-0.65) at a threshold of 0.0 in the series strategy.

Figure 4 illustrates the overall curves of the Youden index variation with the warning thresholds across the 4 units. A threshold of 1.5 was the demarcation point of Youden index for judging the superiority between the parallel warning strategy and the single-indicator warning strategies.

Figure 3. The ROCs of five warning strategies of process data for identifying signals of HAI clusters. Fifty-one thresholds (0.0 to 5.0 step by 0.1) were used for detecting clusters of process data. Dots indicate the sensitivities and 1-specificities for each threshold. AUR: Antibiotic utilization rate in combination; IRS: Inspection rate of bacterial specimens; PRS: Positive rate of bacterial specimens.

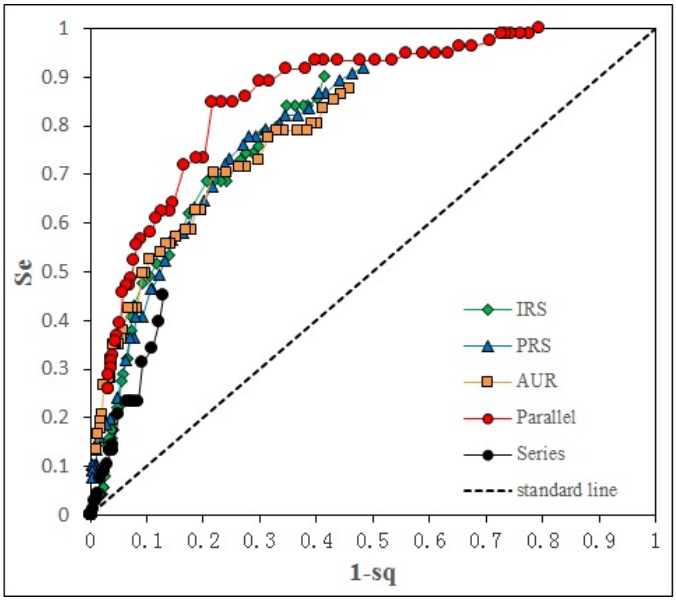


Figure 4. The curves of Youden index varied with thresholds of Shewhart detection model. AUR: Antibiotic utilization rate in combination; IRS: Inspection rate of bacterial specimens; PRS: Positive rate of bacterial specimens.

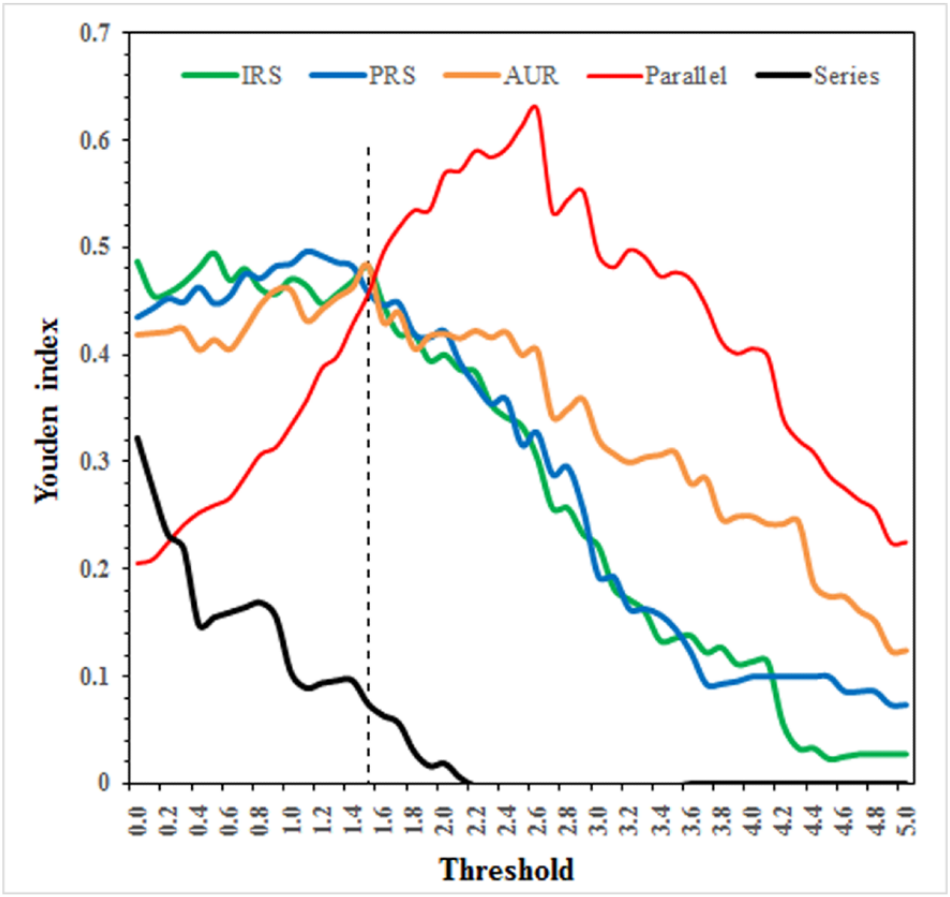


Table 2 shows the mean difference in Youden index between the warning strategies. When the threshold of the Shewhart model was less than or equal to 1.5, Youden indexes in the single-indicator warning strategies were higher than those in the parallel warning strategy, and those of inspection rate of bacterial specimens only and positive rate of bacterial specimens only were better than that of antibiotic utilization rate in combination only; however, when the threshold was greater than 1.5, Youden indexes in the parallel warning strategy were higher than that in the single-indicator warning strategies, and Youden index of antibiotic utilization rate in combination only was better than those of inspection rate of bacterial specimens only and positive rate of bacterial specimens only. In addition, under most thresholds, Youden indexes in the series warning strategy were lower than those in the single-indicator warning strategies and parallel warning strategy.

Table 2. Threshold-matched comparison of Youden index of early warning detection for health care–associated infection clusters.

Threshold and comparison	Mean difference of Youden index (95% CI)	*t*	*df* (n–1)	*P* value
Overall (from 0.0 to 5.0)				
IRS[a] – PRS[b]	–0.011 (–0.023 to 0.001)	–1.877	203	.062
IRS – AUR[c]	–0.062 (–0.085 to –0.038)	–5.206	203	<.001
PRS – AUR	–0.051 (–0.072 to –0.030)	–4.797	203	<.001
IRS – Parallel	–0.124 (–0.155 to –0.093)	–7.856	203	<.001
PRS – Parallel	–0.112 (–0.142 to –0.083)	–7.450	203	<.001
AUR – Parallel	–0.062 (–0.085 to –0.038)	–5.234	203	<.001
IRS – Series	0.230 (0.206 to 0.254)	18.643	203	<.001
PRS – Series	0.241 (0.218 to 0.264)	20.701	203	<.001
AUR – Series	0.292 (0.273 to 0.311)	29.654	203	<.001
Threshold ≤ 1.5 (from 0.0 to 1.5)				
IRS – PRS	0.002 (–0.019 to 0.023)	0.155	63	.878
IRS – AUR	0.033 (0.001 to 0.065)	2.033	63	.046
PRS – AUR	0.031 (0.008 to 0.054)	2.711	63	.009
IRS – Parallel	0.161 (0.131 to 0.191)	10.646	63	<.001
PRS – Parallel	0.159 (0.131 to 0.187)	11.217	63	<.001
AUR – Parallel	0.128 (0.102 to 0.153)	10.037	63	<.001
IRS – Series	0.309 (0.270 to 0.348)	15.851	63	<.001
PRS – Series	0.308 (0.266 to 0.349)	14.750	63	<.001
AUR – Series	0.276 (0.238 to 0.314)	14.489	63	<.001
Threshold > 1.5 (from 1.6 to 5.0)				
IRS – PRS	–0.017 (–0.031 to –0.003)	–2.368	139	.019
IRS – AUR	–0.105 (–0.133 to –0.077)	–7.388	139	<.001
PRS – AUR	–0.088 (–0.115 to –0.062)	–6.614	139	<.001
IRS – Parallel	–0.254 (–0.272 to –0.235)	–26.475	139	<.001
PRS – Parallel	–0.237 (–0.255 to –0.218)	–25.011	139	<.001
AUR – Parallel	–0.148 (–0.167 to –0.130)	–15.637	139	<.001
IRS – Series	0.194 (0.165 to 0.223)	13.217	139	<.001
PRS – Series	0.211 (0.184 to 0.237)	15.818	139	<.001
AUR – Series	0.299 (0.277 to 0.322)	26.259	139	<.001

[a]IRS: inspection rate of bacterial specimens.

[b]PRS: positive rate of bacterial specimens.

[c]AUR: antibiotic utilization rate in combination.

Discussion

Principal Findings

In this study, we retrospectively analyzed the time series surveillance data in 4 HAI high-risk units in WHUH to evaluate the early warning performance of 3 process indicators (antibiotic utilization rate in combination, inspection rate of bacterial specimens, and positive rate of bacterial specimens) for detecting HAI clusters under different warning strategies. The ROC curves of all warning strategies are located above the standard line, indicating that surveillance based on process data was able to detect HAI clusters. Unit-specific results manifested similar outcomes in the 4 independent high-risk units, suggesting a universal warning capability of process data surveillance for HAI cluster detection. However, the accuracy of warnings varied in different units, mainly owing to the differences in population characteristics, antimicrobial utilization behaviors, and pathogenic spectrum.

Based on the correlation between process indicators and infections, process indicators have been used to detect HAI cases and outbreaks. In Freeman's review of research progress in electronic HAI surveillance [19], 77% (34/44) of studies used electronic medical records to detect HAI cases. In another review of the automated detection of HAI outbreaks, 62% (18/29) of studies used microbiological data to detect HAI outbreaks [9]. For example, Fournier et al [10] demonstrated that the consumption of antibiotics for *Pseudomonas aeruginosa* infection could identify 3 epidemics of *P. aeruginosa* infections in a burn center [10]. Carron et al [20] suggested that the prospective electronic surveillance of drug consumption could identify the outbreaks of *P. aeruginosa* infections in the absence of routine traditional surveillance. Moreover, a recent retrospective study in the United States revealed that 9 HAI outbreaks between 2011 and 2016 were successfully detected via data mining of the electronic medical records database, and the earliest warning signal in one of the outbreaks could be generated when the second HAI patient was diagnosed [12]. In a study conducted in 2 hospitals in France, researchers used a space-time permutation scan statistics model to analyze the microbial data in the WHONET system and successfully detected several HAI outbreaks [11].

Combining multiple independent indicators together to detect HAI clusters would be a new research direction for the early warning of HAI outbreaks. Informatization technology provides a convenient tool for the real-time surveillance of multisource process data. Because process indicators are nonspecific for infections, monitoring a single indicator alone cannot fully reflect the occurrence and progression of an HAI, which may limit the accuracy and timeliness of HAI detection. To overcome this problem, a combination of multiple nonspecific indicators provides more infection-related information, which could be expected to improve the early warning performance of HAI detection. This hypothesis was confirmed in our study. The area under the ROC curve was higher for the multi-indicator parallel warning strategy than all other single-indicator warning strategies, indicating that the combined monitoring of multiple process indicators improves the performance of HAI cluster detection. Furthermore, other researchers have proposed similar views. Spolaore et al [21] suggested that the combination of multiple surveillance indicators improved the accuracy of surgical site infection detection. In their study, the positive predictive values for detecting surgical site infections using discharge codes alone or microbiology reports alone were only 70%, but the positive predictive value increased to 97% when these 2 indicators were used in combination.

It is worth mentioning that the combination of multiple indicators is an important factor that affects the accuracy of HAI cluster detection. In our study, compared with the single-indicator warning strategies, the area under the ROC curve was increased when using the parallel warning strategy but decreased when using the series warning strategy. The results of a Youden index comparison exhibited the same situation: the average value of Youden index under each threshold in the parallel warning strategy was greater than those in the single-indicator warning strategies, but the average value of Youden index under each threshold in the series warning strategy was lower than those in the single-indicator warning strategies. In general, the combination of multiple indicators in parallel could improve the sensitivity of warnings but decrease their specificity. Conversely, the combination of multiple indicators in series could improve the specificity of warnings but reduce their sensitivity. This situation was also examined by Bouzbid et al [22]. The sensitivity and specificity for HAI identification using the indicator of a drug prescriptions algorithm alone were 82.3% and 66.7%, respectively, and those using the indicator of the microbiological algorithm alone were 94.0% and 77.3%, respectively. Furthermore, when these 2 indicators were combined in parallel, the sensitivity increased to 99.3%, but the specificity decreased to 58.6%. When these 2 indicators were combined in series, the sensitivity reduced to 77.0%, and the specificity increased to 87.3%.

The threshold of the warning model is another important factor affecting the performance of HAI cluster detection. In prospective surveillance and warning, it was necessary to consider the risk severity and preventive costs of HAI clusters. The threshold of the warning model should be set according to the demand for warning sensitivity and the costs for responding to warning signals. From our results of the Youden index variation with the thresholds of the warning model in Figure 4, we found that when the threshold of the Shewhart model was 1.5 or less, the performance of the parallel warnings for HAI clusters was lower than that of the single-indicator warnings. Only when the threshold was greater than 1.5, the performance of the parallel warnings overtook the single-indicator warnings. Theoretically, a low threshold is prone to higher sensitivity and lower specificity for warnings, whereas a high threshold is prone to lower sensitivity and higher specificity. Owing to the opposing relationship between sensitivity and specificity, the maximum value of Youden index, which comprehensively considers sensitivity and specificity, could be regarded as an alternative criterion for determining the optimal threshold. Our results indicated that Youden index of parallel warnings was optimal at a threshold of 2.6. In addition, the optimal Youden index of parallel warnings exceeded that of single-indicator warnings; furthermore, the optimal Youden indexes of

single-indicator warnings were higher than those of the series warnings. This result again proves that the parallel warning strategy could improve the performance of HAI cluster detection, while the series warning strategy reduced it.

Previous studies have reported some available novel methods for HAI outbreak detection, mainly including (1) exploration of new monitoring objects, (2) innovation of statistical models, and (3) application of intelligent algorithms.

A French project consortium confirmed the feasibility of natural language processing for automatic HAI detection in hospital facilities by developing a natural language processing solution for detecting HAI events in electronic medical records. The overall sensitivity and specificity of the automatic detection of HAIs were 83.9% and 84.2%, respectively [23]. This detection efficiency is similar to that of the multisource surveillance of process data in our study. Another study reported a novel statistical process control chart using Twitter's anomaly and breakout algorithm to detect anomalous HAI surveillance data. It appeared to work better than the statistical process control charts in the context of seasonality and autocorrelation, showing an available algorithm for anomalous HAI detection [24]. In addition, Adhikari et al [25] introduced an efficient data- and model-driven algorithm to detect HAI outbreaks. They designed a near-optimal algorithm to obtain the monitoring data sets and simulated the spread of *Clostridiumdifficile* infection in hospitals. Their algorithm displayed a high sensitivity of 95% for HAI outbreak detection according to data simulation, better than many natural heuristics. In addition, researchers in the Ourense University Hospital Complex (Spain) developed the InNoCBR system for HAI surveillance based on the implementation of intelligent diagnosis for HAIs. Similar to our RT-NISS, the InNoCBR was established using databases of microbiology and pharmacy, but the difference is that it integrates an intelligent diagnostic module into the acquisition process module. The InNoCBR achieved a sensitivity of 70.83% and a specificity of 97.76%, displaying an acceptable detection performance for HAI surveillance [26]. In general, exploring high-quality monitoring data and an intelligent detection model would be the main direction of HAI detection in future research.

Some limitations regarding the generalizability of the findings in this study must be addressed. First, a false correlation likely exists in the warning signals between process data and HAI incidence. This study was a retrospective analysis based on historical surveillance data; thus, the correlation of warning signals between the process data and HAI incidence was judged according to the signal's time and place, lacking epidemiological investigation. Therefore, the applicability of our results requires further research in prospective surveillance.

Second, the process indicators used in our study were a type of nonspecific data, which could provide limited information regarding the occurrence and progress of infections, so it is susceptible to generating negative signals when these nonspecific indicators are used to detect HAI clusters. Although the multiple indicators combined in parallel could improve the warning performance for detecting HAI clusters, they also increased the number of negative signals, resulting in excessive costs for responding to these false warning signals. Consequently, multisource surveillance based on process data could not completely replace the traditional case surveillance at present, and it would be an auxiliary method for detecting disease cases or clusters.

Finally, surveillance noise is an inevitable problem in the automatic surveillance systems based on process data. In fact, automated monitoring is a process of automatically retrieving, identifying, and collecting the formatted data from databases using computer technology. Although automatization improved surveillance efficiency, it was inevitable that some confounding information would be mixed into surveillance data. Because these confounding data, which add noise to surveillance, were usually stored in an unstructured form, it was difficult to automatically wash and refine them in our RT-NISS system. For example, the data on prophylactic medication and therapeutic medication for community infections were mixed into the indicator of antibiotic utilization rate in combination. In addition, some repeated cultures of blood specimens were mixed into the indicators of inspection rate of bacterial specimens and positive rate of bacterial specimens because blood specimens from adults were collected in 2-3 sets each time from different puncture points in WHUH, according to the Operating Procedures of Blood Culture for Clinical Microbiology Laboratory, as issued by the National Health Commission of China. Although these confounding noises could affect the performance of HAI cluster detection, we considered that manually washing and refining them was time-/labor-consuming, and this is contrary to the intention of automatic early warning. In fact, considering that infection control practitioners could investigate warnings more easily in the hospital than in the community, we suggest that it is acceptable to raise the timeliness of warnings at the expense of surveillance noises. We also believe that an automatic washing and refining function for these surveillance noises in HAI cluster detection will be achieved by artificial intelligence technology in the future.

Conclusion

The multisource surveillance of process data in the area network could detect HAI clusters without relying on case reports; moreover, it has advantages in terms of timeliness and automation compared with traditional HAI case surveillance. In this study, we demonstrated that the automated monitoring of the process data of antibiotic utilization rate in combination, inspection rate of bacterial specimens, and positive rate of bacterial specimens could provide early warnings of HAI clusters. The combination of multiple indicators and the threshold of the detection model are 2 important factors affecting warning performance. Multiple data combined in parallel can improve the warning performance, whereas when combined in series, these data can reduce performance. A low threshold of the detection model is more suitable for the single-indicator warning strategies, whereas a high threshold is more suitable for multi-indicator warning strategies. Further prospective research is required to confirm the warning theory of multisource surveillance based on process data.

Acknowledgments

We deeply appreciate the contribution to this thesis made in various ways by all members in the Department of Nosocomial Infection Management, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology. This work was supported by the National Natural Science Foundation of China [NSFC, 72004068]. The funding agreements ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. We thank Editage (www.editage.cn) for English language editing.

Authors' Contributions

YF and YW were equal contributors to this article. YF conceived and designed the study. JZ, MZ, DD, and LL conducted the data collection. XC, XY, and LX, reviewed the articles, YF and YW conducted the statistical analyses, and drafted the manuscript. XY and LX made substantial contributions to reviewing the articles, interpreting data, and drafting or critically revising the manuscript. XY and LX were equal corresponding contributors to this article. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The unit-specific time-series charts of surveillance data and their warning signals (red dots) generated by Shewhart detection model at a threshold of 2.0.
[PNG File , 928 KB - medinform_v8i10e16901_app1.png]

References

1. Allegranzi B, Bagheri Nejad S, Combescure C, Graafmans W, Attar H, Donaldson L, et al. Burden of endemic health-care-associated infection in developing countries: systematic review and meta-analysis. Lancet 2011 Jan 15;377(9761):228-241. [doi: 10.1016/S0140-6736(10)61458-4] [Medline: 21146207]
2. Huixue J, Tieying H, Weiguang L. Economic loss due to Health care-associated infection in 68 general hospitals in China. Chinese Journal of Infection Control 2016;15(9):637-641.
3. Yinghong W, Jie C, Rong L. Marginal analysis of economic burden of hospital acquired infections. China Preventive Medicine 2012;13(4):320-322.
4. Vonberg R, Weitzel-Kage D, Behnke M, Gastmeier P. Worldwide Outbreak Database: the largest collection of nosocomial outbreaks. Infection 2011 Feb;39(1):29-34 [FREE Full text] [doi: 10.1007/s15010-010-0064-6] [Medline: 21153042]
5. Ping C, Ding L. Epidemiological characteristics and preventive strategies of nosocomial infection outbreak incidents in China in recent 30 years. Chinese Journal of Infection Control 2010;9(6):99-92.
6. Shasha W, Yunxi L, Yuqing B. Epidemiological characteristics of nosocomial infection outbreaks in China in recent 13 years. Chinese Journal of Nosocomiology 2018;28(18):2786-2788.
7. Leal J, Laupland KB. Validity of electronic surveillance systems: a systematic review. J Hosp Infect 2008 Jul;69(3):220-229. [doi: 10.1016/j.jhin.2008.04.030] [Medline: 18550211]
8. Leclère B, Lasserre C, Bourigault C, Juvin M, Chaillet M, Mauduit N, et al. Matching bacteriological and medico-administrative databases is efficient for a computer-enhanced surveillance of surgical site infections: retrospective analysis of 4,400 surgical procedures in a French university hospital. Infect Control Hosp Epidemiol 2014 Nov;35(11):1330-1335. [doi: 10.1086/678422] [Medline: 25333426]
9. Leclère B, Buckeridge DL, Boëlle P, Astagneau P, Lepelletier D. Automated detection of hospital outbreaks: A systematic review of methods. PLoS One 2017;12(4):e0176438 [FREE Full text] [doi: 10.1371/journal.pone.0176438] [Medline: 28441422]
10. Fournier A, Voirol P, Krähenbühl M, Bonnemain C, Fournier C, Pantet O, et al. Antibiotic consumption to detect epidemics of Pseudomonas aeruginosa in a burn centre: A paradigm shift in the epidemiological surveillance of Pseudomonas aeruginosa nosocomial infections. Burns 2016 May;42(3):564-570. [doi: 10.1016/j.burns.2015.10.030] [Medline: 26708236]
11. Lefebvre A, Bertrand X, Vanhems P, Lucet J, Chavanet P, Astruc K, et al. Detection of Temporal Clusters of Healthcare-Associated Infections or Colonizations with Pseudomonas aeruginosa in Two Hospitals: Comparison of SaTScan and WHONET Software Packages. PLoS One 2015;10(10):e0139920 [FREE Full text] [doi: 10.1371/journal.pone.0139920] [Medline: 26448036]
12. Sundermann AJ, Miller JK, Marsh JW, Saul MI, Shutt KA, Pacey M, et al. Automated data mining of the electronic health record for investigation of healthcare-associated outbreaks. Infect Control Hosp Epidemiol 2019 Mar;40(3):314-319. [doi: 10.1017/ice.2018.343] [Medline: 30773168]
13. Fan Y, Wang Y, Jiang H, Yang W, Yu M, Yan W, et al. Evaluation of outbreak detection performance using multi-stream syndromic surveillance for influenza-like illness in rural Hubei Province, China: a temporal simulation model based on

healthcare-seeking behaviors. PLoS One 2014;9(11):e112255 [FREE Full text] [doi: 10.1371/journal.pone.0112255] [Medline: 25409025]
14. Fan Y, Yang M, Jiang H, Wang Y, Yang W, Zhang Z, et al. Estimating the effectiveness of early control measures through school absenteeism surveillance in observed outbreaks at rural schools in Hubei, China. PLoS One 2014;9(9):e106856 [FREE Full text] [doi: 10.1371/journal.pone.0106856] [Medline: 25250786]
15. Fan Y, Zou J, Cao X, Wu Y, Gao F, Xiong L. Data on antibiotic use for detecting clusters of healthcare-associated infection caused by multidrug-resistant organisms in a hospital in China, 2014 to 2017. J Hosp Infect 2019 Mar;101(3):305-312 [FREE Full text] [doi: 10.1016/j.jhin.2018.06.011] [Medline: 29935193]
16. Jijiang S, Yubin X, Mingmei D, Junwen L, Wanguo X, Rui H, et al. Hospital communicable disease real time monitoring and early warning system function design and practice. Chinese Hospitals 2013;17(3):11-17. [doi: 10.1080/21548331.1976.11706967] [Medline: 1026634]
17. Zhang Y, Zhang J, Wei D, Yang Z, Wang Y, Yao Z. Annual surveys for point-prevalence of healthcare-associated infection in a tertiary hospital in Beijing, China, 2012-2014. BMC Infect Dis 2016 Apr 18;16:161 [FREE Full text] [doi: 10.1186/s12879-016-1504-4] [Medline: 27091177]
18. Tieying H, Yu Z, Zhenfeng Z. Interpretation of Guideline of control of Health care associated infection outbreak and its practice in clinical nursing. Chinese Nursing Management 2017;20(6):721-724.
19. Freeman R, Moore LSP, García Álvarez L, Charlett A, Holmes A. Advances in electronic surveillance for healthcare-associated infections in the 21st Century: a systematic review. J Hosp Infect 2013 Jun;84(2):106-119. [doi: 10.1016/j.jhin.2012.11.031] [Medline: 23648216]
20. Carron C, Voirol P, Eggimann P, Pannatier A, Chioléro R, Wasserfallen J. Five-year evolution of drug prescribing in a university adult intensive care unit. Appl Health Econ Health Policy 2012 Sep 01;10(5):355-358. [doi: 10.1007/BF03261869] [Medline: 22809277]
21. Spolaore P, Pellizzer G, Fedeli U, Schievano E, Mantoan P, Timillero L, et al. Linkage of microbiology reports and hospital discharge diagnoses for surveillance of surgical site infections. J Hosp Infect 2005 Aug;60(4):317-320. [doi: 10.1016/j.jhin.2005.01.005] [Medline: 16002016]
22. Bouzbid S, Gicquel Q, Gerbier S, Chomarat M, Pradat E, Fabry J, et al. Automated detection of nosocomial infections: evaluation of different strategies in an intensive care unit 2000-2006. J Hosp Infect 2011 Sep;79(1):38-43. [doi: 10.1016/j.jhin.2011.05.006] [Medline: 21742413]
23. Tvardik N, Kergourlay I, Bittar A, Segond F, Darmoni S, Metzger M. Accuracy of using natural language processing methods for identifying healthcare-associated infections. Int J Med Inform 2018 Sep;117:96-102. [doi: 10.1016/j.ijmedinf.2018.06.002] [Medline: 30032970]
24. Wiemken TL, Furmanek SP, Mattingly WA, Wright M, Persaud AK, Guinn BE, et al. Methods for computational disease surveillance in infection prevention and control: Statistical process control versus Twitter's anomaly and breakout detection algorithms. Am J Infect Control 2018 Feb;46(2):124-132. [doi: 10.1016/j.ajic.2017.08.005] [Medline: 28916373]
25. Adhikari B, Lewis B, Vullikanti A, Jiménez JM, Prakash BA. Fast and near-optimal monitoring for healthcare acquired infection outbreaks. PLoS Comput Biol 2019 Sep;15(9):e1007284 [FREE Full text] [doi: 10.1371/journal.pcbi.1007284] [Medline: 31525183]
26. Villamarín-Bello B, Uriel-Latorre B, Fdez-Riverola F, Sande-Meijide M, Glez-Peña D. Gold Standard Evaluation of an Automatic HAIs Surveillance System. Biomed Res Int 2019;2019:1049575 [FREE Full text] [doi: 10.1155/2019/1049575] [Medline: 31662963]

Abbreviations

HAI: Health care–associated infection
ROC: receiver operating characteristic
RT-NISS: Real-Time Nosocomial Infection Surveillance System
WHUH: Wuhan Union Hospital

Edited by G Eysenbach; submitted 05.11.19; peer-reviewed by A Aminbeidokhti, M Herdeiro; comments to author 09.03.20; revised version received 13.07.20; accepted 02.08.20; published 23.10.20.

Please cite as:
Fan Y, Wu Y, Cao X, Zou J, Zhu M, Dai D, Lu L, Yin X, Xiong L
Automated Cluster Detection of Health Care–Associated Infection Based on the Multisource Surveillance of Process Data in the Area Network: Retrospective Study of Algorithm Development and Validation
JMIR Med Inform 2020;8(10):e16901
URL: http://medinform.jmir.org/2020/10/e16901/
doi:10.2196/16901
PMID:32965228

Original Paper

Investigating the Acceptance of Video Consultation by Patients in Rural Primary Care: Empirical Comparison of Preusers and Actual Users

Marius Mueller[1], MSc; Michael Knop[1], MSc; Bjoern Niehaves[1], PhD; Charles Christian Adarkwah[2,3], MD, PhD

[1]Chair of Information Systems, University of Siegen, Siegen, Germany

[2]Department of General Practice and Family Medicine, Philipps-University, Marburg, Germany

[3]Department of Health Services Research, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, Netherlands

Corresponding Author:
Marius Mueller, MSc
Chair of Information Systems
University of Siegen
Kohlbettstraße 15
Siegen, 57072
Germany
Phone: 49 2717402289
Email: marius.mueller@uni-siegen.de

Abstract

Background: The ongoing digitalization in health care is enabling patients to receive treatment via telemedical technologies, such as video consultation (VC), which are increasingly being used by general practitioners. Rural areas in particular exhibit a rapidly aging population, with an increase in associated health issues, whereas the level of attraction for working in those regions is decreasing for young physicians. Integrating telemedical approaches in treating patients can help lessen the professional workload and counteract the trend toward the spatial undersupply in many countries. As a result, an increasing number of patients are being confronted with digital treatment and new forms of care delivery. These novel ways of care engender interactions with patients and their private lives in unprecedented ways, calling for studies that incorporate patient needs, expectations, and behavior into the design and application of telemedical technology within the field of primary care.

Objective: This study aims to unveil and compare the acceptance-promoting factors of patients without (preusers) and with experiences (actual users) in using VC in a primary care setting and to provide implications for the design, theory, and use of VC.

Methods: In total, 20 semistructured interviews were conducted with patients in 2 rural primary care practices to identify and analyze patient needs, perceptions, and experiences that facilitate the acceptance of VC technology and adoption behavior. Both preusers and actual users of VC were engaged, allowing for an empirical comparison. For data analysis, a procedure was followed based on open, axial, and selective coding.

Results: The study delivers factors and respective subdimensions that foster the perceptions of patients toward VC in rural primary care. Factors cover attitudes and expectations toward the use of VC, the patient-physician relationship and its impact on technology assessment and use, patients' rights and obligations that emerge with the introduction of VC in primary care, and the influence of social norms on the use of VC and vice versa. With regard to these factors, the results indicate differences between preusers and actual users of VC, which imply ways of designing and implementing VC concerning the respective user group. Actual users attach higher importance to the perceived benefits of VC and their responsibility to use it appropriately, which might be rooted in the technological intervention they experienced. On the contrary, preusers valued the opinions and expectations of their peers.

Conclusions: The way the limitations and potential of VC are perceived varies across patients. When practicing VC in primary care, different aspects should be considered when dealing with preusers, such as maintaining a physical interaction with the physician or incorporating social cues. Once the digital intervention takes place, patients tend to value benefits such as flexibility and effectiveness over potential concerns.

(JMIR Med Inform 2020;8(10):e20813) doi:10.2196/20813

KEYWORDS

video consultation; technology acceptance; digital health care technology; primary care; rural areas; telemedicine; behavioral intention; eHealth; teleconsultation; electronic consultation; general practitioners

Introduction

Background

In many countries, health care systems are facing increasing challenges that are obliging care providers as well as consumers to adapt. In many rural regions, a shortage of physicians, especially general practitioners (GPs), is obvious and will dramatically increase in the near future [1-3]. A smaller number of GPs will have to take care of a larger number of patients because of demographic changes and an aging population, and catchment areas will increase [4]. Furthermore, GPs—especially in rural areas—have problems finding successors for their practices [5,6]. As a result, imbalances, disparities, and inequitable distributions of care occur, which threaten the comprehensive provision of care and the maintenance of population-wide health [7,8]. The short-term availability of care and medical expertise to which patients are accustomed is at risk. Accordingly, the patients' readiness to change the process of care delivery represents a major governmental as well as scientific issue.

The digitalization of health care processes and treatments over the last two decades represents a promising measure to counteract these issues. A large number of digital technologies have been applied within different medical domains to bridge the emergent gaps in patient treatment, ranging from preventive tools to rehabilitation support systems [9]. For instance, technological advancements in care occur in the form of digitalized patient-physician communication and consultation via web-based video consultation (VC) [10], which enables, for example, remote examinations [11,12], virtual visits at patients' homes [13], and the involvement of relatives and caregivers [14]. Further applications cover the remote collection of patient data through user input or body-worn sensors measuring vital parameters [15,16]; digital prescriptions [17] and web-based scheduling [18]; web-based provision of information on diseases, symptoms, and treatments [19]; and telemonitoring of patients [20].

Clearly, the beneficial implementation, evaluation, and continuous use of health care technologies are vital [21]. A crucial factor for this is the users' acceptance of the technology in play [22,23]. Accordingly, the investigation of factors determining the acceptance of telemedical technology by patients in rural areas represents a major scientific task. Technology acceptance by patients has been subject to several studies [24-26], using models such as the Technology Acceptance Model (TAM) [27,28], the Unified Theory of Acceptance and Use of Technology (UTAUT) [29], or models based on the Theory of Planned Behavior (TPB) [30,31]. However, in the case of telemedicine, these models deliver varying results [23,32], using a wide spectrum of variables without preselection [33,34]. Furthermore, the proposed models often neglect contextual factors and have a narrow view of complex phenomena [35]. Prior models might deliver results that have low explanatory power with regard to primary care settings. Consequently, this study takes an exploratory approach to shed further light on the acceptance of VC as a prominent representative of telemedicine in primary care.

Objectives

Previous studies have looked at telemedical support, for example, in the form of mobile apps, in the case of specialized care and for specific indications, such as palliative medicine [36] or stroke care [13,37]. When looking at primary care, VC represents a telemedical solution that has already been used widely by GPs and specialists to offer innovative ways of patient care and to cope with increasing challenges. A few studies have investigated how patients and medical professionals experience the use of VC systems in primary care [38-42]. Although these studies delivered first insights into the use and acceptance of VC by patients, the focus was predominantly on the convenience and benefits of VC [39,40], and an in-depth study seeking to unveil the social, personal, technical, environmental, and organizational factors affecting the use of VC in primary care remains to be done. In addition, the samples involved do not account for the vast majority of patients who have not yet encountered VC for treatment and are thus still in the process of forming behavioral intentions and attitudes toward VC in primary care.

Thus, the objectives of this study are (1) to empirically identify factors that drive patient evaluation, acceptance, and utilization of VC technologies, using research on patients with and without experience in using such a system within rural primary care; and (2) to contrast these 2 populations to expose the differences and commonalities that are potentially rooted in digital interventions. On the basis of these findings, implications can be drawn for design, application, and theory. This paper contributes to understanding what is important to patients in their roles as preusers as well as actual users of VC. The paper focuses on primary care setting because it affects a majority of citizens, from chronic patients who are obliged to interact frequently with their physicians to patients with nonsevere and acute diseases that render visits and consultations occasional. As the supply situation mentioned earlier reveals rural areas are threatened by a shortage of GPs, this study investigates patients and practices in a representative rural area in Germany.

Methods

Study Design

We conducted a qualitative study, as part of a regional project agenda, empirically investigating the digitization of primary care practices and health care processes in the German setting, focusing in particular on the specific conditions in rural areas. With regard to our study design, we seek to empirically explore and identify factors that shape patients' perceptions, evaluations, adoption, and continuous use of VC in the primary care setting, focusing on the patient perspective. We conducted

semistructured interviews with patients with and without experience in using VC, which allows for a comparison of these 2 patient cohorts and unveil differences and commonalities in what is important to the patients. We draw upon the notion of *preusers*, who are "[...] individuals and groups who do not have well-developed notions of how digital technologies fit into or affect their lives" [31]. As telemedical solutions such as VC are not widely used in health care [43,44], this user group represents a majority of patients. On the other hand, some primary care practices have already adopted VC systems for patient treatment. Accordingly, the population of patients who have actually encountered telemedicine is growing, forming a group of actual users who potentially pursue different norms, beliefs, and behaviors. Hence, we engage both preusers and actual users for 3 reasons. First, the intention to use a system is a major predictor of actual use [27] and is formed beforehand based on expectations and, in many cases, lack of actual experience. Therefore, it is useful to interrogate preusers to shed light on the emergence of behavioral intentions. Second, to establish fair and equitable access to care, it is important to include all patients who have already used or potentially will use VC for treatment. This includes patients with a lack of technical affinity or willingness to participate in VC; hence, they might remain preusers. It is important to determine what drives or hinders these patients from using VC. Finally, from a provider's perspective, the economic success of implementing telemedicine seems important. Here, achieving a critical mass of users is crucial, calling for comprehensive technology acceptance to transform preusers into actual users.

Data Collection

All 158 GP practices in the region of Siegen-Wittgenstein, Germany, were contacted and asked for their experience with VCs. Of these practices, only 2 GPs stated that they had intensive experience with this method of treatment. These 2 GP practices included VCs in their regular office hours, that is, patients can opt for a VC or a face-to-face consultation (FTFC). For a VC, patients have to register and book an appointment through the website of the GP practice. Afterward, a link is sent to the patient via a text message and email, specifying the date and time of appointment. Finally, the patient needs to click on the link to enter the conference room. The GP immediately gets a notification when a patient is online and can start the VC. Patients were offered VC use instead of FTFC. All patients who registered for a VC were consecutively asked to participate in the study to have a representative sample of practice attendees. Owing to the COVID-19 pandemic, there was great interest in VCs among patients, and only one patient in each practice refused to take part. In rural GP practices in Germany, the number of patients registered is higher on average than in practices in urban regions. Furthermore, the proportion of older patients is somewhat greater. This is also the case for the practices participating in this study.

We conducted 20 semistructured interviews, drawing samples from these 2 primary care practices. Interviews were carried out in 2 phases. In phase 1, we conducted 10 interview sessions (sample A) at the first site with patients who did not have any experience with VCs. Thus, sample A represents the preuser group. In phase 2, we conducted 10 additional interviews with participants from the second practice (sample B) who had already used a VC system to consult their physician. As 4 of these interviews took place digitally because of the COVID-19 pandemic, interviewees were asked to evaluate their VC experience despite the acute circumstances (eg, restrictions on personal contact) if possible. In doing so, we aimed to collect coherent data. Sample B forms the actual user group. Both samples were recruited via 2 GP practices, as mentioned earlier, who reached out to suitable patients willing to participate in the project, thus allowing a convenient (sample A) and purposeful (sample B) sampling approach [45]. The sample yielded a total of 22 interviewees, 9 women and 13 men, with an average age of 51.2 years (SD 19.2). Interviews took 26 min on average and were conducted from August 2019 to April 2020. The comprehensive sample and interview process characteristics are illustrated in Multimedia Appendix 1. The samples thus consisted of patients with different education levels, age, gender, and health status. We tried to recruit samples that (1) adequately represent the common client base of rural primary care practices in the investigated region and (2), in the case of sample B, can be seen as recurrent users of telemedicine according to the physicians and self-disclosure.

The interview guideline covered 5 questions groups seeking to unveil different factors explaining the patients' attitudes toward VC and adjacent telemedical solutions. Questions covered patient, social, environmental, and organizational as well as technical and interaction factors, building upon the classification by Or and Karsh [33]. The guideline was adapted between the interview phases to reflect on the varying level of experience with VC between samples. However, we did not change the guideline in between interviews of the same sample, thus avoiding the possibility that the interviewees' statements could influence each other. In doing so, we aimed for unbiased data because the attitudes and perceptions under investigation are highly individual. Both interview guidelines are presented in Multimedia Appendix 2. In the case of sample A, a technical scenario was presented to the interviewees at the beginning of each interview to allow for a common understanding of telemedical treatments. The scenario involved 2 components: first, a live VC with the GP about distance, for example, from home; and second, mobile sensory equipment that enables patients to measure and transfer vital parameters (eg, blood pressure) on their own. In the case of sample B, for the sake of comparability, the application of sensor equipment besides the experienced VC was introduced to the interviewees at the end of each interview session. Here, we additionally asked for the patients' perception of the usefulness and applicability of the sensory equipment in future treatments.

The interviews were conducted in German by 2 members of the research group, audio recorded, and transcribed nonverbatim, leaving out pauses and off-topic exchanges of words. Before each interview, the interviewees signed an informed consent, inter alia briefing them about voluntariness, the anonymization and partial publication of data, and their right to withdraw their study participation. For the sake of analysis, comparability with literature, and reporting, the transcripts were translated into English. The study was approved by the data protection commissioner of the University of Siegen.

Data Analysis

To reflect our data collection procedure, we followed a two-phased data analysis approach consisting of an inductive phase, analyzing data from sample A, and a subsequent deductive phase, analyzing data from sample B. Here, the results from phase 1 are used for analysis in phase 2. This procedure allows factors to persist, but also to change, complement, or substitute each other or be canceled out entirely. In this way, differences as well as commonalities between preusers and actual users become visible.

In the first phase, we analyzed the data gathered from sample A inductively to identify and comprehensibly define the first set of relevant factors associated with preusers. We followed a three-step approach [46]. First, 2 authors coded the interview data independently. The approach proposed by Strauss and Corbin [47], consisting of open and axial coding, was followed. Selective codes were formed by subsuming redundant and/or related codes into superordinate categories that represent factors with regard to the patients' attitudes toward and their acceptance of the technologies involved. Second, the 2 coding authors discussed their individual categorizations, merging codes with similar reasoning and formulation, and resolved disagreements. Consequently, some of the standalone codes were subsumed under others because they represented a particular facet of the resulting factor. This procedure led to a first categorization scheme consisting of 3 groups, which involve 7 subsumed factors, and 1 standalone group, which forms a factor by itself. Finally, based on the elaborate scheme, each involved researcher recoded the data, assigning the 8 factors to the interviewees' statements. After that, a final discussion on categories, their dimensions and facets, and factor-to-data assignments was carried out.

In the second phase, we analyzed the data collected from sample B in a deductive manner. Here, the final coding scheme from the first phase was applied as the initial template to code the remaining data. Again, the data were coded in 3 steps as in phase 1. First, the authors independently assigned identified codes to the data, allowing new codes to emerge and existing codes to be redefined. Statements that did not fit in the coding scheme were again coded following the inductive approach described earlier (open, axial, and selective coding). This led to a new factor dealing with patient responsibilities and obligations, which included novel insights with regard to the actual user group. Second, step 2 was carried out analogous to the first phase, leading to a new merged categorization that comprised the extended 4 factor groups, followed by, finally, a recoding of the data by both members of the research group. Before recoding, the raters agreed upon the data segments to which codes were assigned. To check for interrater reliability of the coding performed, we calculated Cohen kappa [48] after the final recoding of all the data was done (see step 3 during data analysis). The resulting value of 0.75 indicates a substantial agreement in coding and, thus, sufficient reliability [49].

Multimedia Appendix 3 shows the quantity of interview coding that relates to the factors after recoding of the data and the number of interviews that involve the respective factor. Both samples are presented individually and complemented by total numbers.

Results

Overview

In total, 4 different design and application relevant factor groups (attitudes and expectations, human interaction, rights and obligations, and social factors), each with their respective subdimensions, emerged from samples A and B. Although the context and connotations of specific factors varied between our 2 samples, we explored interesting commonalities and differences. The presented findings come from the experiences of patients with no experience in digital or video-based treatment (sample A) and those who have already experienced VCs with their GP (sample B). To preserve the anonymity of interviewees and to avoid the potential delineation of interviews (eg, by their order), we assigned a random number (from A1/B1 to A10/B10) to each interview [46].

Attitudes and Expectations Toward Telemedicine

Usefulness of VC

In general, participants linked the use of VCs to perceived benefits. Although participants from sample A focused on 3 specific, positive aspects, participants from sample B mentioned several more factors they considered useful.

Of the 10 participants from sample A, 8 assumed that VC could be useful in saving their trip to the physician's practice, as did the majority of interviewees from sample B. Participants associated the travel-saving effect of VC use with further benefits, that is, saving time and not being exposed to potential sources of infection from other patients:

> *Via Skype or the like, I would be able to talk to my doctor, tell him my problems. And if he could solve my problems right away, I wouldn't have to go to the practice. That would be something I appreciate.* [Interview A6]

Participants from sample B found further aspects of VC beneficial, including higher flexibility to integrate an appointment into their daily routine and the prompt setting of a virtual appointment as opposed to an office appointment. Participants from sample B especially emphasized its practicality with regard to their own professional or informal obligations:

> *Well, concerning organization, it was quite easy, and of course quite practical, because I hadn't to leave work. I had my appointment at 10 am, I just went into another room, where I was undisturbed. That's just very comfortable.* [Interview B9]

Furthermore, half of all participants from sample B mentioned that a video appointment appeared to be more focused because of its transparent time limit. When using a web-based application form to receive an appointment for VC, participants were able to choose between different time slots, each comprising 10 min. Therefore, some interviewees argued that the scope of a specific appointment appeared to be clearer and more narrowed through digitization, as the timeframes of the appointments were

displayed by the program they used to connect with their physician. In addition, they distinguished between appointments where their physical presence was necessary and those where their digital presence was sufficient. Overall, participants from sample B differentiated the usefulness of telemedicine systems to a higher degree and acknowledged more perceived benefits from digital appointments than did participants from sample A.

Security Aspects

Although interviewees were asked about the potential and actual disadvantages of VC, participants from both samples emphasized the need for data security. Participants were generally aware of the sensitivity of their medical data and expressed their concerns about the possibility of misuse. Foremost, interviewees described their personal medical data as vulnerable and transparent:

> *I already said it, the past shows how little you can trust the whole thing. I am as transparent as [...] this window.* [Interview A7]

> *[...] Technology certainly has, the definition of it certainly is to support humans and to be helpful, but every coin has got two sides, therefore every technology used by bad people holds the possibility to be misused.* [Interview B8]

Although the majority of interviewees from both samples A and B considered data security an important issue and a fundamental precondition for fully trusting a telemedicine system, participants from sample B put such statements into another perspective by stating that they risked the possibility of data insecurity to enjoy the benefits of VC:

> *But I don't necessarily look at it that way, you might say, well data security, but I'm not attaching too much value on such things. See, we've so much data to disclose every day, you just have to be alert.* [Interview B4]

> *Well, it's [digital appointment] working with video, internet, whatever. Who knows if it's been recorded or what. In the beginning, I thought that way, but in the end, it's nonsense. If it happens, it happens.* [Interview B8]

Accordingly, interviewees were aware of the importance of personal data in relation to the use of digital appointments. Participants who actually used VC compared the possibility of a breach of data with the normality of the potential misuse of data they could experience in comparable situations. In the end, the threat of data interception by third parties did not seem to outweigh the perceived advantages of digital appointments.

Operability of VC

As an antecedent to using digital technology, participants discussed the benefits of a preferably easy operation of a VC system. Although interviewees from sample A talked about prospective barriers they might have to face to use VC, participants from sample B emphasized the actual operability of the system they used for digital appointments:

> *My wife, she had to work with computers. Nowadays, she's just like me, overstrained. Because she didn't use it anymore.* [Interview A8]

> *Well, it was really easy. When you're booking an appointment online, you have to register. Afterwards you just choose a time slot and you get an e-mail with a PIN, and within the e-mail there's a link. And you even didn't need to enter the PIN.* [Interview B3]

In addition, participants from sample B discussed possible features for extending the operability or functionality of the system they had experienced, such as the simultaneous transfer of personal medical data they collected by themselves (eg, blood pressure or coagulation level), better feedback functions while waiting for the physician to join the digital appointment, and an app to use instead of a website.

Human Interaction and Its Impact on the Use of VC

Human Contact

Participants emphasized their need for personal and direct interactions. Although participants from both samples mentioned their concerns about technological changes leading to the replacement of direct physical contact between them and their physician, only participants from sample A expressed their wish for personal assistance regarding the use of VC at home. Overall, interviewees from sample A used the uniqueness of direct, personal human interaction as an argument to reject VC, whereas interviewees from sample B described situations in which they considered adequate digital appointments.

Nonetheless, for participants from both samples, personal contact with their physician remained highly important. Participants from sample A insisted on office visits and tended to exclude the possibility of audiovisual treatment from their own scope of action. Of the 10 participants from sample A, 8 mentioned the importance of a personal relationship with their physician, even if that meant accepting specific disadvantages. Similarly, participants from sample B also emphasized their need for office appointments as well:

> *Even if you have to wait a long time, the personal contact, you have to keep it upright.* [Interview A1]

> *It [video consultation] won't work for every situation, logically. You need a personal talk. You need that.* [Interview B4]

Participants from sample A clearly distinguished between a physical meeting with their physician and contact mediated by VC. They seemed to assume that through personal contact, physicians are able to provide them with better care. VC was seen to restrict the senses of the physician and therefore limit the scope of examining a patient:

> *I don't want that; I like to have personal contact. I think just from the way a patient behaves, as a doctor you're able to recognize certain things [...] that cannot be transmitted through video.* [Interview A5]

In contrast, participants from sample B often assessed the appropriateness of a digital appointment through their actual interaction with their physician. They clearly perceived the specific limitations of a digital appointment, for example, the

inability of their physician to examine them physically, to discuss severe diagnostic results, or deal appropriately in situations of high emotional stress:

> *A digital appointment, it's limited by definition. You can't, like when you're actually in your physician's practice, get a sonography, for example.* [Interview B6]

> *[...] when you get a bad diagnosis in a hospital and have to discuss it with your physician. When it's really serious, and you like to talk about it, I'd rather do it face to face.* [Interview B1]

> *When I'm mentally unstable [...] when I face specific problems, I prefer to speak with someone in person. It's maybe, I don't know, it's a matter of trust […].* [Interview B4]

Overall, participants from sample B differentiated the occasions for the use of telemedicine, whereas participants from sample A expressed their concern regarding a potential lack of physical and personal contact with their physician. Therefore, participants from sample B were able to provide specific situations that they preferred not to be digitally mediated.

Trust in Physician

Regarding the acceptance of digital appointments, participants from both samples discussed the relationship between them and their physician as a determining factor. Although even skeptical participants from sample A agreed to use VC when they were told to do so by their physician, interviewees from sample B emphasized the importance of trusting their physician to find the best medical solution for their problem, even without being physically present:

> *If my doctor says "Hey look, I've got a cool thing here, we're able to communicate regularly. I am always there for you. If anything happens, you come to my practice, otherwise let's try it that way," I think if he says it that way, if the doctor I trust means it, it's more likely I'll do it.* [Interview A3]

> *And I think there has to be a trusting relationship to your doctor. To really want to test it [video appointment], to try something new, and to have trust in your physician, that everything will be ok, when you're treated via video talk.* [Interview B3]

Although the role of the physician as a mediator between technology and the patient seemed to be essential to all participants, most interviewees from sample B indicated that nonetheless, some medical indications might justify digital treatment from an unfamiliar physician. Without being explicitly asked about it, some participants came up with the idea of being treated by unknown physicians for minor physical complaints (eg, a cold), a discussion of objective medical data (eg, test results), or highly urgent and acute symptoms (eg, an emergency):

> *When it's just about a cold, or a cough, or whatever, it doesn't really matter who's treating me. As long as I've got the feeling of being taken seriously to some degree.* [Interview B5]

In summary, a trusting relationship between participants and physicians fostered a positive attitude toward the use of VC and might be considered an important condition for effective digital treatment. Furthermore, interviewees from sample B appeared to be partially willing to receive care from unfamiliar professionals to receive the perceived benefits from digital appointments.

Rights and Obligations

Voluntariness of Use

Participants from both samples liked the idea of video appointments being an optional extension of the already existing primary care services and emphasized that using it needed to be a voluntary choice. Participants from sample B in particular recognized that choosing between a digital or an office treatment involved a bilateral process of negotiation between them and their physician:

> *It would be nice if my doctor doesn't tell me to use it, but if he makes me an offer with specific advantages.* [Interview A2]

> *I think, I would appreciate having a voice. It's one thing to say, well, when my doctor asks me "could we talk about it digitally?" [...] But you have to have a choice to say "no, I'd like to speak to you in person."* [Interview B5]

Although participants would clearly like to choose a specific type of medical service voluntarily, interviewees also realized that their health status sometimes indicated a specific kind of medical service (digital or office) and agreed to follow the advice of their physician:

> *If you've got minor questions, concerning your medication or high blood pressure or anything else. Then you don't have to come here, just get such a long distance consultation.* [Interview A4]

> *[...] and some appointments can be digitalized, you might ask your patient, what can be done digitally and when do you need an actual [office] appointment.* [Interview B4]

In general, participants from both samples seemed to express their wish to participate in the decision-making process regarding whether a digital appointment appeared to be adequate in a specific situation. Acknowledging the primary care physician's professional assessment of their health status and indication for a specific service (digital or office treatment), participants emphasized the importance of the voluntary use of VC.

Availability of Care and VC

Participants from both samples were concerned about a present or future shortcoming of medical service in general because of a lack of professionals. Interviewees gave examples of long waiting times to get office appointments and severe problems in reaching their physician's medical assistants via telephone:

> *Nobody answers the phone, when you've got something important to tell. Nobody's answering it.* [Interview A1]

> *When I look at it, well, members of my own family were diagnosed with cancer tentatively, they needed an MRI really quick. They had to wait six months, every day they died of worry.* [Interview B8]

Broaching the issue of VC, participants from both samples generally described digital appointments as an opportunity to increase reachability and shorten waiting time:

> *In the morning I asked myself if I should go to the [physician's] practice. Then I remembered he's offering that service [digital appointment]. I logged in and had a look at it. When I had a closer look, I realized there was a slot vacant at 11 am. Wouldn't have got a real [office] appointment that quick.* [Interview B4]

Although participants from sample B emphasized the benefit of digital appointments in increasing the availability of medical services and as an opportunity for primary care physicians to increase the number of patients they are able to care for, participants discussed potential disadvantages from their physician's perspective, for example, an increased workload and unnecessary appointments because of the simplicity and availability of digital appointments. However, overall, participants from sample B suggested that VC might be able to counter the present challenges regarding the provision of care, which were mentioned by nearly all participants from both samples.

Patient Responsibilities

Only participants from sample B discussed the matter of self-responsibility regarding digital appointments. They mentioned that their own technological competence fostered the smooth processing of a digital appointment and that their own preparations were necessary beforehand:

> *Someday you'll use it [video consultation] the first time and then you may realize that the camera won't work or something. Surely, patients have to prepare. I've got the feeling, such an [digital] appointment, I have to write it in my calendar, it's easy to forget, rather than actually going to the practice.* [Interview B10]

Necessary preparations were not reduced to technological issues. Participants mentioned that they had to focus on a specific issue rather than portray their pathogenetic history extensively. Furthermore, participants from sample B emphasized the importance of the competence to interpret one's own symptoms and decide on an office or digital appointment:

> *Well, you've got a certain period of time, and when I've got my appointment, I know I can't tell the whole story around it, for a quarter of an hour, but there are specific things [...]* [Interview B9]

> *But I think everyone's able to judge, depending on your symptoms or pre-existing conditions, whether or not you have to go to the physician's practice or if it's suitable to use digital appointments.* [Interview B3]

In this regard, participants suggested that patients should carefully assess their health status, potential issues, and appropriate ways of dealing with them. Overall, interviewees from sample B reflected on the conditions for a satisfactory use of VC regarding their own possibilities of shaping an interaction between themselves and their physician.

Social Factors

In general, several social factors influencing the use of technology can be found in our data. Unconsidered habitual attitudes toward digital technologies were often expressed in nonspecific, generic terms. Responses from both samples can be divided into statements concerning the social expectations of technology use in general and individual, private social interaction related to one's own experiences with VC.

Interestingly, the majority of interviewees from sample A tended to express their readiness in a more passive way, assuring that they would not stand in the way of technological innovation, whereas participants from sample B stated their willingness to actively promote VC as an innovative technology. To explain user-specific readiness to use, participants from both samples often draw on stereotypes related to age:

> *So, I believe the willingness of older people to learn something new isn't there. If I want to deal with it [new technologies], I have to be competent. Otherwise, when a problem occurs, something won't work, and when the problem occurs a second time, they just throw it away. That's how I see it.* [Interview A6]

> *Well, my mother, she was born in 1943, she'll have trouble using it [video consultation], because she doesn't know how to handle a pc, how to use a video chat function.* [Interview B8]

Participants from both samples reported the importance of talking to family members, friends, and colleagues about VC. Participants from sample A related their statements to relatively close family members and described their behavior as reactive, whereas interviewees from sample B considered themselves as being one of the first among their peers to use such innovative technology:

> *They always try to motivate us. "Daddy do this, do that," they know I always decline, but their father complies with it.* [Interview A4]

> *Well, when I talk about it [use of video consultation] with my former wife, I have to add, we've got a good connection [...] she said, she'll give it a try, because you're just more flexible when you're an employed person.* [Interview B9]

Overall, participants from sample B appeared to see themselves as pioneers when using VC. They actively discussed their experiences of digital appointments with peers and seemed to influence others rather than be influenced. Nonetheless, social interaction and the impact of social expectations and norms, including stereotypes, remain an essential factor in the use of technology.

Discussion

The results shed light on factors that influence the attitudes, acceptance, and behavior of patients regarding the application of VCs as a representative of telemedicine in rural primary care. Studying preusers and actual users of telemedical solutions enables the empirical comparison of these 2 populations. The main findings are discussed against the background of technology design, application, and theory, thus delivering implications for practitioners, developers, and researchers.

Differences in the Perception of Benefits and Security Issues

With regard to the participants' expectations and perceptions toward the application of telemedicine in primary care, they showed high levels of perceived usefulness and beneficial effects of the technology. The literature on technology acceptance and adoption behavior has a vast corpus of studies that incorporate the perceived usefulness (TAM) and expected performance (UTAUT) of a technology as an antecedent to its use, together with associated intentions and attitudes [29]. A recent meta-analysis of research on the acceptance of consumer health technologies has shown that perceived usefulness can explain use behavior on a significant level [32]. Our study delivers further insights by considering both preusers and actual users of VC. As our findings indicate, preusers seem to attach less importance to the potential benefits of VC while focusing on other considerations for and against VC. Therefore, the inclusion of the patient's role (preuser vs actual user) as a factor within research models on the acceptance of VC in primary care appears promising.

The preuser group mentioned only a few benefits they could think of, such as avoiding long and repeated travel to the practice. In contrast, the actual user group cited more examples of profitable outcomes. They experienced VCs to be more focused, efficient, and flexible. Literature has shown that there is no significant difference between text-based, information technology–mediated consultations and FTFC with regard to effectiveness as perceived by patients [50]. Our findings complement prior research on the use of VC in primary care, which deemed VC as a more thorough and convenient treatment method compared with FTFC [40,41,51] and telephone consultations [39], and indicate that video-based consultations are perceived as more effective and targeted than FTFC. Interestingly, while perceiving VC as a thorough approach [41,52], patients comply with the time limits of concise video meetings. Despite the limited time given, patients are satisfied with the experienced VC. From a practical standpoint, this makes it easier for GPs to schedule and adhere to appointments. On the contrary, preusers lack the experience of VC being a sufficient and satisfactory way of treatment. Here, the temporal limitation of virtual sessions can hinder patients from opting for VC. As research shows, patients are oftentimes skeptical about their health issues being addressed via VC depending on their condition, which renders VC inapplicable in certain situations [39,40,51].

Accordingly, from a practical standpoint, to increase the acceptance and use intentions of preusers, telemedical solutions such as VC systems should be promoted in more detail, clarifying what a VC can and cannot accomplish. It can be assumed that a higher awareness of benefits can lead to increased intentional and proactive use. In this regard, the benefits of VC have become apparent during the ongoing COVID-19 pandemic, which has put restrictions on the physical contact between GPs and their patients [53]. In times where access to care is limited, the potential of VC to bridge spatial gaps between GPs and patients has led to an uptake in VC implementation and use [54]. This is particularly true in rural areas that often lack comprehensive access to care [7]. Telemedicine, and VC in particular, enables GPs and clinicians to cope with given restrictions, maintain care of infected patients as well as those not related to COVID-19, and decrease infection rates [54].

As research shows [39], although the operability, usability, and ease of use of VC as well as the process of familiarizing oneself with the system are important to both user groups, the prevalence of security concerns and associated behavior varies. Although research on VC in primary care has focused primarily on the patient's security in the sense of reducing physical harm and achieving health progress, our findings represent a novel aspect that contrasts preusers and actual users of VC. The preuser group indicates great concerns regarding the security of telemedicine and the potential of data misuse and leakage. In addition, preusers tend to affiliate these concerns with the intention of not using telemedicine. Actual users, while still aware of security issues, seem to be more willing to take risks in light of experienced benefits and convenience. The actual use of and exposure to telemedicine seems to alleviate patients' concerns regarding technology-associated security. Literature has shown that the perceived benefits of technology use can outweigh perceived risks [55]. Accordingly, technology design should focus on alleviating the risks and threats perceived by preusers. From a design standpoint, to increase patients' trust in telemedicine, technologies should present their privacy policies in an accessible and understandable way [56]. It appears to be important to incorporate ways of displaying technical security measures to the patient while not requiring high levels of technical skills, for instance, in the form of protection-ensuring labels [57] or a lucid and manageable list of people and institutions having access to the data [58]. This information can also be delivered to preuser patients by GPs to alleviate potential concerns that might not be perpetuated once the VC is experienced.

Impacts of VC on the Patient-Physician Relationship

In the context of human interaction and its impact on the use of VC, the results indicate the importance of maintaining physical contact with the physician. The preuser group in our study expects fewer positive outcomes for virtual treatments and tends to reject the technology because in-office treatment by the physician is perceived to be superior. This finding is in line with prior studies on VC primary care, which indicate that the lack of physical contact potentially impedes adequate examination and proper treatment [59,60]. VC was deemed useful in nonurgent or routine situations [51]. In addition to this occasion-based opting for VC or FTFC, as our findings show, several patients requested aid by a competent person (eg,

medical staff or peers) in case they had to use VC. This finding closely relates to the facilitating conditions that form an antecedent of the intention to use as well as the actual use of a technology in the UTAUT model. In particular, the model states that the degree of guidance and support experienced by the user when opting for a technology has an effect on their willingness to (continuously) use it [24,29]. Concerning our findings, this relation seems to be particularly relevant when dealing with preusers of VC in primary care. In the meantime, actual users seem to be able to fathom the feasibility and applicability of VC, enabling them to identify occasions and health issues that a digital treatment can address while placing less importance on guidance or external support. Apparently, patients are more able to differentiate occasions for office or digital treatments once they have conducted a VC with the physician. This finding concurs with studies that have shown that patients gain deeper knowledge about the occasions that are suitable for VC in comparison with FTFC when actively using the system [51].

Closely related to the relationship between patients and physicians, participants from both samples indicated that trust in the respective physician and an existing relationship are major drivers of technology acceptance and willingness to use it. Looking at investigations on technology acceptance and adoption behavior, trust in the opposite party (here, GPs offering VC) and their actions represents an important factor in the users' technology assessment [32]. The findings indicate that a trusting patient-physician relationship increases the belief in an effective, beneficial, and safe treatment via VC, which is in line with prior studies on VC in primary care [39]. In the case of preusers, the data suggest that even obligatory technology use becomes more acceptable once interpersonal trust is achieved. Although actual users of VC have concrete experiences and specific benefits at their disposal, preusers tend to use trust as a heuristic input to decision making, making it easier for them to form an attitude [61]. Accordingly, the physician's proactive invitation to arrange a digital appointment can potentially achieve higher use intentions once the relationship is considered trustworthy. By offering VC to the patient as an alternative to FTFC, the GP conveys that the virtual treatment is deemed suitable and beneficial, which could mitigate a patient's potential concerns.

Revealing another interesting finding that complements the literature on VC in primary care, our study suggests that patients are partially willing to be treated via VC by a physician who is not their regular GP. Apparently, there are health-related occasions that go beyond the choice between VC and FTFC, which has been subject to prior studies [40,62-64] and further subdivide the feasibility of VC based on the need for trust. Both our findings and the literature show that there are suitable issues that can be addressed via VC but that call for different degrees of trust in the treating physician, such as receiving a severe diagnosis. There is still ambiguity on whether patients prefer a comforting environment (eg, at home) or an FTFC when talking about issues that are perceived to be sensitive or serious [60]. However, our findings reveal that there are health issues (such as a cold) that do not call for an already existing relationship with the physician. Accordingly, bringing together such patients with GPs who offer VC and are available for consultation represents a promising treatment model that further alleviates disparities in access to care. This concept can increase the number of patients who are suitable for treatment via VC while reducing the workload for the GPs responsible. This is particularly relevant in today's health care because patients who can be treated virtually represent only a fraction of the clinical workload [63]. Therefore, based on the patient's indication, perceived severity, and need for a trustworthy relationship, connecting patients with available physicians other than their own GP via VC promises a flexible and cost-effective way of delivering treatment [65].

Emerging Tasks and Freedoms for Patients in a Virtual Setting

Looking at the patients' rights and obligations that come along with the introduction of VC in primary care, the results show emerging freedoms, tasks, and behavioral patterns that patients should be aware of. Both samples wished for a voluntary and autonomous use of VC that enables them to adopt or reject the technology without disadvantages. The literature on technology acceptance has already identified the degree of voluntariness when choosing a technology as an important factor that influences users in their decision making [28,29]. Further research in the domain of health care technologies identified the patients' freedom and preferences when choosing between VC and FTFC as an important factor that fosters their adoption or rejection of VC [40,51,59,60]. Our findings complement these studies by shedding light on the scenario in which using VC in primary care becomes obligatory and free of alternatives, for instance, in remote areas with detrimental access to care or in times of viral outbreaks such as the COVID-19 pandemic. Although the preuser group mentioned that they were willing to use an obligatory VC system if their physician suggested it, the actual user group indicated that they would obey telemedical obligations if they deemed the treatment occasion appropriate. That is, once a patient experiences VC and is able to fathom its applicability, the need for freedom of choice seems to decrease. Instead, actual users of VC tend to agree with obligatory digital appointments because they have gained the know-how regarding the way VC is applied in primary care. Theoretically speaking, they might have achieved higher levels of health literacy, which enables them to assess and understand health issues and necessary treatments [66] and computer self-efficacy, making them more competent in adequately choosing and using VC [67]. This is in line with previous research indicating that illiteracy with regard to proper technology use is a barrier to opting for VC instead of FTFC [68].

In addition to the degree of voluntariness in the use of VC, digital primary care also comes with obligations for the patient. Looking at prior research on the use of VC in primary care, our findings complement the first insights on the patient's role in achieving an effective and satisfactory experience and use of technology. One of the first studies on VC in primary care indicated that patients perceive "[…] that they had responsibilities in ensuring the VC happened in an appropriate way, for example, conducting the VC in an appropriate setting […]" [39]. Further research raised the need for patients to prepare for a VC session, for example, by finding a private room and using headphones to secure privacy, as a novel consideration that is unique to telehealth [60]. Our data complement these

findings and suggest that patients become aware of their roles and responsibilities through the actual use of the technology. Although the first sample did not mention this issue, the actual user group described the need to assess the feasibility of digital treatment as opposed to a physical visit. The participants stressed that prevalent health issues and potential treatments should be considered before making an appointment. When the participants considered a treatment via VC inappropriate for solving the health issues, they emphasized that a patient should be able to reject a digital appointment. Again, this requires patients to achieve higher levels of health literacy, so they are able to understand their condition, possible treatments, and the potential of telemedicine. Physicians might actually need to increase the effort of patient empowerment to ensure a degree of health literacy, which enables patients to decide what kind of treatment is appropriate in a specific situation [69,70]. With regard to technology design, the VC system can provide information about potentially prevalent diseases, feasible treatments, and contacts to specialized care to increase the patients' health literacy and degree of empowerment. This information and potentially resulting measures by the patient can also be used to inform upcoming VCs, enriching patient-physician communication and mutual understanding.

Social Impact on the Use and Design of VC

The data show different views on social factors in using VC. Apparently, the preuser group incorporated social cues and external norms into their attitude toward VC. The data suggest a subconscious trend toward social conformity when talking about technology in primary care. Interestingly, both groups gave credence to social stereotypes, claiming telemedicine to be more appropriate for younger generations. Preusers therefore seem to act according to what they think is the social norm, as suggested by prior studies on technology acceptance behavior [29,71,72]. In contrast, actual users talk about their influence on their peers. They appear (to themselves) to be innovative pioneers and inform their social cues about their mostly positive experiences. This is closely related to the image of the user (which the UTAUT model incorporates) coined as "[…] the degree to which use of an innovation is perceived to enhance one's image or status in one's social system" [29]. Although our findings show no support for actual users intentionally seeking to improve their image, their positive influence on their peers' assessment of VC for treatment can still be identified. Thus, the patients' self-perception as the first adopter of VC within their social system holds the potential to further explain why patients opt for VC in primary care and stick with it.

In addition, prior research on the use of VC in primary care has already shown that specific patient groups, such as older adults and the housebound, are perceived by GPs as not having the degree of technical skill to use VC effectively, although they would benefit from it the most [38]. Interestingly, the demographics of patients who opt for VC and those who do not differ significantly [68], indicating a social bias in the form of stereotyping [73]. Our study enhances these findings by indicating that lack of skill is also perceived among patients. As a result, to profit from the social dissemination of VC and its benefits, the resolution of these perceived gaps between patient groups by practitioners and policy makers seems necessary. GPs, for instance, are potentially able to achieve mutual understanding between patients and thus increase the intention to use VC by being transparent about the actual use of VC by different populations, including older adults. Furthermore, identified *pioneers* of VC can serve the GP as gatekeepers who influence their peers in a positive way.

At the design level, incorporating social cues and the adoption behavior of peers into telemedicine, and VC in particular, can potentially increase a patient's willingness to (continuously) use it. Preusers, in particular, seem to highly value opinions and assessments by their peers. With regard to actual users, research shows that experienced users of virtual consultation increasingly form negative attitudes toward the use of the system [51]. From a theoretical standpoint, establishing and maintaining the use of VC can be achieved by finding ways to present behaviors of others to the patient, following the concept of *nudging* [74]. The idea of nudging is to gently encourage people to behave in a certain way at a subconscious level [75]. Nudges in the form of messages presented to the patient (eg, "Most of your friends have used VC before to contact their physician.") can potentially lead to higher use intentions. Our findings expand prior research that shows that digital nudges can positively influence the willingness to use novel technologies in hospitals [76]. In turn, our findings contribute to the theoretical concept of nudging by indicating that the use of social norms as a nudging option [74] holds the potential to increase the acceptance rate of VC in primary care.

Limitations

This study has some limitations. First, the sampling procedure is prone to selection bias because we did not strictly regulate participant characteristics and demographics. Thus, the sample yields varying degrees of technical affinity and age, which could frame the results in a certain direction. People opting for telemedicine (representing sample B) might exhibit particular characteristics such as dispositional innovativeness that could partially explain patient perceptions and behavior. In addition, the interviews were partially conducted during the COVID-19 pandemic, which could have influenced the responses of the participants. As VC is the only way for many patients to consult their GP, at least during the acute times of the pandemic, interviewees might have formed stronger intentions and more positive reactions to the technology. Second, it is difficult to discuss identified factors in comparison with patients living in urban areas because the data are limited to the chosen context. The urban patients' experiences of VC and their intention to participate might differ with regard to the varying structural circumstances and quantity of practitioners. Third, participants were recruited in a limited region. Nevertheless, this area is representative of rural regions in Germany according to size and demographic characteristics. Further studies should be conducted to shed light on urban environments and enable rural-urban comparisons in a reliable and insightful way.

Conclusions and Outlook

This study investigates factors that constitute patients' attitudes, perceptions, and technology acceptance behavior regarding the use of VC in the rural primary care setting. To account for different levels of experience with technology use, this study

involves the perspectives of preusers as well as actual users of VC. The empirical data enable the comparison of these 2 perspectives and the provision of implications for the design, application, and theory of VC. The study delivers an in-depth description and discussion of patients' experiences and attitudes that complement our understanding of the use of VC in primary care by involving both preusers and actual users of VC. The findings can be of interest to researchers, medical practitioners, and designers of VC and telemedicine solutions, further enabling them to increase the behavioral intentions of preusers, maintain continuous use of VC by already experienced patients, and achieve a critical mass of patients participating in digital treatments.

With regard to the patients' behavioral intentions toward and actual use of VC in primary care, that is, their technology acceptance behavior, this study unveils several links to established models and includes antecedents of health care technology acceptance. Interestingly, when looking at models that have been put up and tested by researchers to investigate patients' acceptance of consumer health technologies, none of these models (TAM, TPB, or UTAUT) combines the factors of perceived usefulness, trust in GP, social norms and image, degree of voluntariness and obligatory use, patient responsibility and involvement, and need for physical contact, which our findings suggest [32]. Hence, proposing and testing a theoretical model that integrates these antecedents represents a promising avenue for technology acceptance researchers when investigating the use and acceptance of VC in primary care. In addition, the comparison of user groups shows that the priorities, needs, expectations, and attitudes toward using VC in primary care vary between preusers and actual users. Therefore, the inclusion of both patient groups appears to be feasible when testing new theoretical models of technology acceptance by patients. The role of the patient (preuser vs actual user) thus holds potential explanatory power when looking at antecedents of core constructs such as intention to use VC.

This paper opens up many further research opportunities for future work as well as for preceding studies. First, research can be conducted to further investigate the gap between different generations regarding their perceptions and opinions on telemedicine. The findings suggest that stereotyping takes place across all ages, that is, the association of older adults with a lack of technical skills or the perceived social pressure coming from younger generations. Second, to overcome the monomethod approach, studies engaging wider and more heterogeneous populations can be conducted, for instance, in the form of surveys conducted on the web or on the GP's site. In doing this, researchers can gather data with higher external validity and achieve further insights into how to implement digital technologies within the primary care setting, based on quantitative measures. Here, interventional studies appear to be feasible to shed light on the behavioral and attitudinal changes triggered by the use of digital technology. Third, to generate feasible and beneficial designs for technology, the involvement of technology experts and developers, working together in focus groups and workshops, can yield concrete technical features and innovations that further improve the comprehensive provision of primary care in rural areas.

Acknowledgments

This research was supported by the DIPRA project funded by Sparkasse Siegen.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample and interview characteristics.
[DOC File , 43 KB - medinform_v8i10e20813_app1.doc]

Multimedia Appendix 2
Interview guidelines.
[DOC File , 56 KB - medinform_v8i10e20813_app2.doc]

Multimedia Appendix 3
Code quantities.
[DOC File , 57 KB - medinform_v8i10e20813_app3.doc]

References

1. Adarkwah CC, Schwaffertz A, Labenz J, Becker A, Hirsch O. Burnout and work satisfaction in general practitioners practicing in rural areas: results from the HaMEdSi study. Psychol Res Behav Manag 2018;11:483-494 [FREE Full text] [doi: 10.2147/PRBM.S179503] [Medline: 30425595]
2. Adarkwah CC, Schwaffertz A, Labenz J, Becker A, Hirsch O. GPs' motivation for teaching medical students in a rural area-development of the motivation for medical education questionnaire (MoME-Q). PeerJ 2019;7:e6235 [FREE Full text] [doi: 10.7717/peerj.6235] [Medline: 30697479]

3. Broermann M, Wunder A, Messemaker A, Schnoor H, Gerlach FM, Sennekamp M. [Structuring and supporting specialist training in general practice: evaluation of a Hesse-wide mentoring program for doctors]. Z Evid Fortbild Qual Gesundhwes 2018 Nov;137-138:69-76. [doi: 10.1016/j.zefq.2018.08.001] [Medline: 30297261]
4. Demiris G, Hensel B. Technologies for an aging society: a systematic review of "smart home" applications. Yearb Med Inform 2008:33-40. [Medline: 18660873]
5. Adarkwah CC, Schwaffertz A, Labenz J, Becker A, Hirsch O. [Assessment of the occupational perspectives of general practitioners in a rural area. Results from the study HaMedSi (Hausärzte [GPs] for Medical education in Siegen-Wittgenstein)]. MMW Fortschr Med 2019 Oct;161(Suppl 6):9-14. [doi: 10.1007/s15006-019-0919-4] [Medline: 31587169]
6. Mueller M, Knop M, Reßing C, Freude H, Oschinsky FM, Klein HC, et al. Constituting Factors of a Digitally Influenced Relationship between Patients and Primary Care Physicians in Rural Areas. In: Proceedings of the 53rd Hawaii International Conference on System Sciences. 2020 Presented at: HICSS'20; January 7-10, 2020; Hawaii, USA. [doi: 10.24251/hicss.2020.447]
7. Wilson NW, Couper ID, De Vries E, Reid S, Fish T, Marais BJ. A critical review of interventions to redress the inequitable distribution of healthcare professionals to rural and remote areas. Rural Remote Health 2009;9(2):1060 [FREE Full text] [Medline: 19530891]
8. Politzer R, Yoon J, Shi L, Hughes R, Regan J, Gaston M. Inequality in America: the contribution of health centers in reducing and eliminating disparities in access to care. Med Care Res Rev 2001 Jun;58(2):234-248. [doi: 10.1177/107755870105800205] [Medline: 11398647]
9. Kvedar J, Coye MJ, Everett W. Connected health: a review of technologies and strategies to improve patient care with telemedicine and telehealth. Health Aff (Millwood) 2014 Feb;33(2):194-199. [doi: 10.1377/hlthaff.2013.0992] [Medline: 24493760]
10. Almathami HK, Win KT, Vlahu-Gjorgievska E. Barriers and facilitators that influence telemedicine-based, real-time, online consultation at patients' homes: systematic literature review. J Med Internet Res 2020 Feb 20;22(2):e16407 [FREE Full text] [doi: 10.2196/16407] [Medline: 32130131]
11. Seuren LM, Wherton J, Greenhalgh T, Cameron D, A'Court C, Shaw SE. Physical examinations via video for patients with heart failure: qualitative study using conversation analysis. J Med Internet Res 2020 Feb 20;22(2):e16694 [FREE Full text] [doi: 10.2196/16694] [Medline: 32130133]
12. Host B, Turner A, Muir J. Real-time teleophthalmology video consultation: an analysis of patient satisfaction in rural Western Australia. Clin Exp Optom 2018 Jan;101(1):129-134 [FREE Full text] [doi: 10.1111/cxo.12535] [Medline: 28436157]
13. Appireddy R, Khan S, Leaver C, Martin C, Jin A, Durafourt BA, et al. Home virtual visits for outpatient follow-up stroke care: cross-sectional study. J Med Internet Res 2019 Oct 7;21(10):e13734 [FREE Full text] [doi: 10.2196/13734] [Medline: 31593536]
14. Funderskov K, Raunkiær M, Danbjørg DB, Zwisler A, Munk L, Jess M, et al. Experiences With Video Consultations in Specialized Palliative Home-Care: Qualitative Study of Patient and Relative Perspectives. J Med Internet Res 2019 Mar 21;21(3):e10208 [FREE Full text] [doi: 10.2196/10208] [Medline: 30896436]
15. Pantelopoulos A, Bourbakis N. A survey on wearable sensor-based systems for health monitoring and prognosis. IEEE Trans Syst Man Cybern C 2010 Jan;40(1):1-12. [doi: 10.1109/tsmcc.2009.2032660]
16. Vesnic-Alujevic L, Breitegger M, Guimarães Pereira A. 'Do-it-yourself' healthcare? Quality of health and healthcare through wearable sensors. Sci Eng Ethics 2018 Jun;24(3):887-904. [doi: 10.1007/s11948-016-9771-4] [Medline: 27029478]
17. Mayakul T, Ayuthaya SD. A digital prescription refill system based on healthcare standard in Thailand. Int J Appl Biomed Eng 2018;11(1):28-35.
18. Nazia S, Ekta S. Online Appointment Scheduling System for Hospitals–An Analytical Study. Int J Innov Res Sci Eng Technol 2014;4(1):21-27.
19. Ahmad F, Hudak PL, Bercovitz K, Hollenberg E, Levinson W. Are physicians ready for patients with internet-based health information? J Med Internet Res 2006 Sep 29;8(3):e22 [FREE Full text] [doi: 10.2196/jmir.8.3.e22] [Medline: 17032638]
20. Aamodt I, Lycholip E, Celutkiene J, Strömberg A, Atar D, Falk R, et al. Health Care Professionals' Perceptions of Home Telemonitoring in Heart Failure Care: Cross-Sectional Survey. J Med Internet Res 2019 Feb 06;21(2):e10362 [FREE Full text] [doi: 10.2196/10362] [Medline: 30724744]
21. Currie M, Philip LJ, Roberts A. Attitudes towards the use and acceptance of eHealth technologies: a case study of older adults living with chronic pain and implications for rural healthcare. BMC Health Serv Res 2015 Apr 16;15(1):162 [FREE Full text] [doi: 10.1186/s12913-015-0825-0] [Medline: 25888988]
22. Chau PY, Hu PJ. Investigating healthcare professionals' decisions to accept telemedicine technology: an empirical test of competing theories. Inf Manag 2002 Jan;39(4):297-311. [doi: 10.1016/S0378-7206(01)00098-2]
23. Harst L, Lantzsch H, Scheibe M. Theories predicting end-user acceptance of telemedicine use: systematic review. J Med Internet Res 2019 May 21;21(5):e13117 [FREE Full text] [doi: 10.2196/13117] [Medline: 31115340]
24. Kohnke A, Cole ML, Bush R. Incorporating UTAUT predictors for understanding home care patients' and clinician's acceptance of healthcare telemedicine equipment. J Technol Manag Innov 2014 Jul;9(2):29-41. [doi: 10.4067/S0718-27242014000200003]

25. Samhan B. Patients' Resistance Towards Health Information Technology a Perspective of the Dual Factor Model of IT Usage. In: Proceedings of the 50th Hawaii International Conference on System Sciences. 2017 Presented at: HICSS'17; January 4-7, 2017; Hawaii, USA. [doi: 10.24251/hicss.2017.412]
26. Rahman M. Does Personality Matter When We Are Sick? An Empirical Study of the Role of Personality Traits and Health Emotion in Healthcare Technology Adoption Decision. In: Proceedings of the 50th Hawaii International Conference on System Sciences. 2017 Presented at: HICSS'17; January 4-7, 2017; Hawaii, USA. [doi: 10.24251/hicss.2017.407]
27. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989 Sep;13(3):319-240. [doi: 10.2307/249008]
28. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manag Sci 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
29. Venkatesh, Morris, Davis, Davis. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425-278. [doi: 10.2307/30036540]
30. Yan M, Or C. A 12-week pilot study of acceptance of a computer-based chronic disease self-monitoring system among patients with type 2 diabetes mellitus and/or hypertension. Health Informatics J 2019 Sep;25(3):828-843. [doi: 10.1177/1460458217724580] [Medline: 28820007]
31. Deng Z, Mo X, Liu S. Comparison of the middle-aged and older users' adoption of mobile health services in China. Int J Med Inform 2014 Mar;83(3):210-224. [doi: 10.1016/j.ijmedinf.2013.12.002] [Medline: 24388129]
32. Tao D, Wang T, Wang T, Zhang T, Zhang X, Qu X. A systematic review and meta-analysis of user acceptance of consumer-oriented health information technologies. Comput Hum Behav 2020 Mar;104:106147. [doi: 10.1016/j.chb.2019.09.023]
33. Or CK, Karsh B. A systematic review of patient acceptance of consumer health information technology. J Am Med Inform Assoc 2009;16(4):550-560 [FREE Full text] [doi: 10.1197/jamia.M2888] [Medline: 19390112]
34. Bagozzi R. The legacy of the technology acceptance model and a proposal for a paradigm shift. J Assoc Inf Syst 2007 Apr;8(4):244-254. [doi: 10.17705/1jais.00122]
35. Salovaara A, Tamminen S. Acceptance or appropriation? A design-oriented critique of technology acceptance models. In: Isomäki H, Saariluoma P, editors. Future Interaction Design II. London, UK: Springer; 2009:157-173.
36. Steindal SA, Nes AA, Godskesen TE, Dihle A, Lind S, Winger A, et al. Patients' experiences of telehealth in palliative home care: scoping review. J Med Internet Res 2020 May 5;22(5):e16218 [FREE Full text] [doi: 10.2196/16218] [Medline: 32369037]
37. Kim DY, Kwon H, Nam K, Lee Y, Kwon H, Chung YS. Remote management of poststroke patients with a smartphone-based management system integrated in clinical care: prospective, nonrandomized, interventional study. J Med Internet Res 2020 Feb 27;22(2):e15377 [FREE Full text] [doi: 10.2196/15377] [Medline: 32130140]
38. Randhawa RS, Chandan JS, Thomas T, Singh S. An exploration of the attitudes and views of general practitioners on the use of video consultations in a primary healthcare setting: a qualitative pilot study. Prim Health Care Res Dev 2019 Jan;20:e5 [FREE Full text] [doi: 10.1017/S1463423618000361] [Medline: 29909798]
39. Donaghy E, Atherton H, Hammersley V, McNeilly H, Bikker A, Robbins L, et al. Acceptability, benefits, and challenges of video consulting: a qualitative study in primary care. Br J Gen Pract 2019 Sep;69(686):e586-e594. [doi: 10.3399/bjgp19X704141] [Medline: 31160368]
40. Hammersley V, Donaghy E, Parker R, McNeilly H, Atherton H, Bikker A, et al. Comparing the content and quality of video, telephone, and face-to-face consultations: a non-randomised, quasi-experimental, exploratory study in UK primary care. Br J Gen Pract 2019 Sep;69(686):e595-e604. [doi: 10.3399/bjgp19X704573] [Medline: 31262846]
41. Johansson AM, Lindberg I, Söderberg S. Patients' experiences with specialist care via video consultation in primary healthcare in rural areas. Int J Telemed Appl 2014;2014:143824 [FREE Full text] [doi: 10.1155/2014/143824] [Medline: 25243009]
42. Johansson AM, Lindberg I, Söderberg S. Healthcare personnel's experiences using video consultation in primary healthcare in rural areas. Prim Health Care Res Dev 2017 Jan;18(1):73-83. [doi: 10.1017/S1463423616000347] [Medline: 27640522]
43. Karsh B, Weinger M, Abbott P, Wears R. Health information technology: fallacies and sober realities. J Am Med Inform Assoc 2010;17(6):617-623 [FREE Full text] [doi: 10.1136/jamia.2010.005637] [Medline: 20962121]
44. Ranganathan C, Balaji S. Key factors affecting the adoption of telemedicine by ambulatory clinics: insights from a statewide survey. Telemed J E Health 2020 Feb;26(2):218-225. [doi: 10.1089/tmj.2018.0114] [Medline: 30874484]
45. Marshall MN. Sampling for qualitative research. Fam Pract 1996 Dec;13(6):522-526. [doi: 10.1093/fampra/13.6.522] [Medline: 9023528]
46. Mueller M, Heger O. Health at Any Cost? Investigating Ethical Dimensions and Potential Conflicts of an Ambulatory Therapeutic Assistance System through Value Sensitive Design. In: Proceedings of the 39th International Conference on Information Systems. 2018 Presented at: ICIS'18; September 16-18, 2018; San Francisco, CA, USA.
47. Strauss A, Corbin J. Basics of Qualitative Research. Thousand Oaks, CA: Sage Publications; 1998.
48. Brennan RL, Prediger DJ. Coefficient kappa: some uses, misuses, and alternatives. Educ Psychol Measure 2016 Jul 2;41(3):687-699. [doi: 10.1177/001316448104100307]

49. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
50. Mirzaei T, Kashian N. Revisiting Effective Communication Between Patients and Physicians: Cross-Sectional Questionnaire Study Comparing Text-Based Electronic Versus Face-to-Face Communication. J Med Internet Res 2020 May 13;22(5):e16965 [FREE Full text] [doi: 10.2196/16965] [Medline: 32401213]
51. Mold F, Hendy J, Lai Y, de Lusignan S. Electronic consultation in primary care between providers and patients: systematic review. JMIR Med Inform 2019 Dec 3;7(4):e13042 [FREE Full text] [doi: 10.2196/13042] [Medline: 31793888]
52. Mair F, Whitten P. Systematic review of studies of patient satisfaction with telemedicine. Br Med J 2000 Jun 3;320(7248):1517-1520 [FREE Full text] [doi: 10.1136/bmj.320.7248.1517] [Medline: 10834899]
53. Ohannessian R, Duong TA, Odone A. Global telemedicine implementation and integration within health systems to fight the covid-19 pandemic: a call to action. JMIR Public Health Surveill 2020 Apr 2;6(2):e18810 [FREE Full text] [doi: 10.2196/18810] [Medline: 32238336]
54. Jakhar D, Kaur I. Potential of chloroquine and hydroxychloroquine to treat COVID-19 causes fears of shortages among people with systemic lupus erythematosus. Nat Med 2020 May;26(5):632. [doi: 10.1038/s41591-020-0853-0] [Medline: 32269358]
55. Kehr F, Kowatsch T, Wentzel D, Fleisch E. Blissfully ignorant: the effects of general privacy concerns, general institutional trust, and affect in the privacy calculus. Info Systems J 2015 Mar 18;25(6):607-635. [doi: 10.1111/isj.12062]
56. Tsai J, Egelman S, Cranor L, Acquisti A. The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study. Information Systems Research 2011 Jun;22(2):254-268. [doi: 10.1287/isre.1090.0260]
57. Kelley PG, Bresee J, Cranor LF, Reeder RW. A 'Nutrition Label' for Privacy. In: Proceedings of the 5th Symposium on Usable Privacy and Security. 2009 Presented at: SOUPS'09; July 15-17, 2009; Mountain View, California, USA p. 1-12. [doi: 10.1145/1572532.1572538]
58. Benaloh J, Chase M, Horvitz E, Lauter K. Patient Controlled Encryption: Ensuring Privacy of Electronic Medical Records. In: Proceedings of the 2009 ACM workshop on Cloud computing security. 2009 Presented at: CCSW'09; Novermber 1-4, 2009; Chicago, Illinois, USA p. 103-114. [doi: 10.1145/1655008.1655024]
59. Fatehi F, Martin-Khan M, Smith AC, Russell AW, Gray LC. Patient satisfaction with video teleconsultation in a virtual diabetes outreach clinic. Diabetes Technol Ther 2015 Jan;17(1):43-48. [doi: 10.1089/dia.2014.0159] [Medline: 25296189]
60. Powell RE, Henstenburg JM, Cooper G, Hollander JE, Rising KL. Patient perceptions of telehealth primary care video visits. Ann Fam Med 2017 May;15(3):225-229. [doi: 10.1370/afm.2095] [Medline: 28483887]
61. Lewicki RJ, Brinsfield C. Framing trust: trust as a heuristic. In: Donohue WA, Rogan RG, Kaufman S, editors. Framing Matters: Perspectives on Negotiation Research and Practice in Communication. New York, USA: Peter Lang Publishing; 2011:110-135.
62. Tates K, Antheunis ML, Kanters S, Nieboer TE, Gerritse MB. The effect of screen-to-screen versus face-to-face consultation on doctor-patient communication: an experimental study with simulated patients. J Med Internet Res 2017 Dec 20;19(12):e421 [FREE Full text] [doi: 10.2196/jmir.8033] [Medline: 29263017]
63. Greenhalgh T, Shaw S, Wherton J, Vijayaraghavan S, Morris J, Bhattacharya S, et al. Real-world implementation of video outpatient consultations at macro, meso, and micro levels: mixed-method study. J Med Internet Res 2018 Apr 17;20(4):e150 [FREE Full text] [doi: 10.2196/jmir.9897] [Medline: 29625956]
64. Armfield NR, Bradford M, Bradford NK. The clinical use of Skype—for which patients, with which problems and in which settings? A snapshot review of the literature. Int J Med Inform 2015 Oct;84(10):737-742. [doi: 10.1016/j.ijmedinf.2015.06.006] [Medline: 26183642]
65. Sterling R, LeRouge C. On-demand telemedicine as a disruptive health technology: qualitative study exploring emerging business models and strategies among early adopter organizations in the United States. J Med Internet Res 2019 Nov 15;21(11):e14304 [FREE Full text] [doi: 10.2196/14304] [Medline: 31730038]
66. Nutbeam D. The evolving concept of health literacy. Soc Sci Med 2008 Dec;67(12):2072-2078. [doi: 10.1016/j.socscimed.2008.09.050] [Medline: 18952344]
67. Compeau DR, Higgins CA. Computer self-efficacy: development of a measure and initial test. MIS Quarterly 1995 Jun;19(2):189 [FREE Full text] [doi: 10.2307/249688]
68. Barsom EZ, Jansen M, Tanis PJ, van de Ven AW, Blussé van Oud-Alblas M, Buskens CJ, et al. Video consultation during follow up care: effect on quality of care and patient- and provider attitude in patients with colorectal cancer. Surg Endosc 2020 Mar 20 epub ahead of print. [doi: 10.1007/s00464-020-07499-3] [Medline: 32198552]
69. Osei-Frimpong K, Wilson A, Lemke F. Patient co-creation activities in healthcare service delivery at the micro level: the influence of online access to healthcare information. Technol Forecast Soc Change 2018 Jan;126:14-27. [doi: 10.1016/j.techfore.2016.04.009]
70. Castro EM, Van Regenmortel T, Vanhaecht K, Sermeus W, Van Hecke A. Patient empowerment, patient participation and patient-centeredness in hospital care: a concept analysis based on a literature review. Patient Educ Couns 2016 Dec;99(12):1923-1939. [doi: 10.1016/j.pec.2016.07.026] [Medline: 27450481]
71. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]

72. Taylor S, Todd PA. Understanding information technology usage: a test of competing models. Inf Syst Res 1995 Jun;6(2):144-176. [doi: 10.1287/isre.6.2.144]
73. Oakes P, Haslam S, Turner J. Stereotyping and Social Reality. Malden, UK: Blackwell Publishing; 1994.
74. Sunstein C. Nudging: A Very Short Guide. J Consum Policy 2014 Oct 16;37(4):583-588. [doi: 10.1007/s10603-014-9273-1]
75. Voyer B. 'Nudging' behaviours in healthcare: Insights from behavioural economics. British Journal of Healthcare Management 2015 Mar 02;21(3):130-135. [doi: 10.12968/bjhc.2015.21.3.130]
76. Meske C, Amojo I, Poncette A, Balzer F. The potential role of digital nudging in the digital transformation of the healthcare industry. In: Marcus A, Wang W, editors. Design, User Experience, and Usability. Application Domains. Cham, UK: Springer International Publishing; 2019:323-336.

Abbreviations

FTFC: face-to-face consultation
GP: general practitioner
TAM: Technology Acceptance Model
TPB: Theory of Planned Behavior
UTAUT: Unified Theory of Acceptance and Use of Technology
VC: video consultation

Edited by G Eysenbach; submitted 29.05.20; peer-reviewed by L Seuren, D Tao; comments to author 08.08.20; revised version received 31.08.20; accepted 22.09.20; published 22.10.20.

Please cite as:
Mueller M, Knop M, Niehaves B, Adarkwah CC
Investigating the Acceptance of Video Consultation by Patients in Rural Primary Care: Empirical Comparison of Preusers and Actual Users
JMIR Med Inform 2020;8(10):e20813
URL: http://medinform.jmir.org/2020/10/e20813/
doi:10.2196/20813
PMID:32969339

Original Paper

Factors Associated With Influential Health-Promoting Messages on Social Media: Content Analysis of Sina Weibo

Qingmao Rao[1,2], MA; Zuyue Zhang[1,2], MA; Yalan Lv[1], PhD; Yong Zhao[3], PhD; Li Bai[4], MA; Xiaorong Hou[1,2], MA

[1]College of Medical Informatics, Chongqing Medical University, Chongqing, China

[2]Medical Data Science Academy, Chongqing Medical University, Chongqing, China

[3]School of Public Health and Management, Chongqing Medical University, Chongqing, China

[4]Hospital of Zigong Mental Health Central, Zigong, China

Corresponding Author:
Xiaorong Hou, MA
College of Medical Informatics
Chongqing Medical University
No.1, Medical College Road
Yuzhong District
Chongqing, 400016
China
Phone: 86 138 8390 1680
Email: xiaoronghou@cqmu.edu.cn

Abstract

Background: Social media is a powerful tool for the dissemination of health messages. However, few studies have focused on the factors that improve the influence of health messages on social media.

Objective: To explore the influence of goal-framing effects, information organizing, and the use of pictures or videos in health-promoting messages, we conducted a case study of Sina Weibo, a popular social media platform in China.

Methods: Literature review and expert discussion were used to determine the health themes of childhood obesity, smoking, and cancer. Web crawler technology was employed to capture data on health-promoting messages. We used the number of retweets, comments, and likes to evaluate the influence of a message. Statistical analysis was then conducted after manual coding. Specifically, binary logistic regression was used for the data analyses.

Results: We crawled 20,799 Sina Weibo messages and selected 389 health-promoting messages for this study. Results indicated that the use of gain-framed messages could improve the influence of messages regarding childhood obesity (P<.001), smoking (P=.03), and cancer (P<.001). Statistical expressions could improve the influence of messages about childhood obesity (P=.02), smoking (P=.002), and cancer (P<.001). However, the use of videos significantly improved the influence of health-promoting messages only for the smoking-related messages (P=.009).

Conclusions: The findings suggested that gain-framed messages and statistical expressions can be successful strategies to improve the influence of messages. Moreover, appropriate pictures and videos should be added as much as possible when generating health-promoting messages.

(JMIR Med Inform 2020;8(10):e20558) doi:10.2196/20558

KEYWORDS

health-promoting messages; social media; Sina Weibo; influence; framing effects; health communication

Introduction

Overview

Television, newspapers, radio, magazines, and other traditional media have long been the communication tools relied upon for health communication. More recently, social media, such as Facebook, Twitter, and Sina Weibo (or Weibo), has gained explosive growth, especially in China [1]. As of June 2019, China has 854 million internet users, the vast majority of whom obtain information through social media [2]. An increasing number of scholars believe that social media has great potential as a tool in the field of health care [3] and health promotion [4,5].

Sina Weibo is one of the most popular social media platforms in China. In December 2018, this platform had 462 million active accounts, including more than 37,000 media organizations and 170,000 government agency accounts [6]. Yang et al [1] described Weibo as a mixture of features of Twitter and Facebook. Weibo also has some elements of a bulletin board system, blog, and social networking site. Social media has become a unique platform for health promotion due to its potential for viral messaging [7], its ability to challenge authority [8], and its diversity of users [9]. In China, Weibo has been widely used for health communication [10-12].

However, many health-promoting messages released on social media lack influence [13]. Health-promoting messages transmit health information through mass media to prevent diseases and promote health [14]. Van 't Riet et al [15] asserted that a health-promoting message should include health-related behaviors and the consequences of behaviors. As a result, health-promoting messages may contain terminologies and substantial expository text [16]. However, Chinese residents have low overall health information literacy [17]. Most people think that health messages on the internet are often too complex to understand [18]. The complex content in health-promoting messages hinders people's willingness to interact with them. Moreover, few studies have explored the effect of using specific communication strategies to enhance people's participation with health messages on social networking platforms [19].

Some strategies can improve the audience's acceptance of and participation with health messages. Myers [20] believed that health message–framing effects can be conducive to the spread of health-promoting messages and encourage people's health behaviors. Meppelink et al [21,22] used pictures and videos in a health-promoting message to change the communication effect. Allen and Preiss [23] found that a statistical type of information organization made information more persuasive. Sundar [24] suggested that audiences are more likely to recognize information provided by professionals than by nonprofessionals. Social media has broken through the limitations of traditional media and made these strategies easier to use. A previous health information survey on child obesity [25] verified that framing effects could significantly change the audience's attitude toward information. Whether these strategies, especially the framing effects, contribute to the impact of health-promoting messages on the Weibo platform is worth studying.

Weibo has become one of the most notable platforms for people in China to seek health-promoting messages [26]. Examining the factors that shape the degree of influence of health-promoting messages on the Weibo platform is crucial. Many studies on health information dissemination have been carried out by questionnaires, but this technique has the problem of subjective bias. Therefore, this work employed a web crawler and manual coding to collect data from the real-world platform of Weibo. We considered the message-framing types as the influencing factors and explored whether the message sources, expression types, and use of pictures or videos would affect the degree of influence of health-promoting messages. The results of this study can guide the communication of related health themes and provide experimental evidence for theoretical research related to the framing effects of health-promoting messages.

Background

Message-Framing Effects

Kahneman and Tversky [27] first proposed framing effects using the "Asian disease problem" example, thereby beginning the research on framing effects in the field of psychology. Message-framing effects for health-promoting messages have become a hot research topic. Prospect theory can explain framing effects. This theory holds that people can be acutely aware of whether a framing message emphasizes potential benefits or risks [27,28]. Health-promoting messages can be divided into gain-framed messages (which highlight the beneficial consequences of healthy behavior) or loss-framed messages (which underscore the detrimental counterpart) [15]. The gain- and loss-framed effects show that when health care messages emphasize the positive or negative results of an action or omission, the persuasiveness of the messages significantly differ. Previous studies have confirmed that a gain-framed message is effective in promoting the use of sunscreen and exercise activities [29]. Conversely, a loss-framed message is persuasive in promoting mammography, chest self-examination [30], and colorectal cancer detection [31]. Given the prior research [15] and the text-based message expression of Weibo [32], we posit that the gain-framed and loss-framed effects on health-promoting message dissemination on Weibo are similar to those of print media. A Weibo message that clearly expresses positive or negative consequences was regarded as framed.

Expression Type and Visuals

A statistical expression message refers to a missive that contains quantitative or numerical information [33]. Many studies have compared the persuasiveness of different types of information organization, especially the statistical and narrative evidence types, albeit with inconclusive results [34,35]. Allen and Preiss [23] believed that the use of statistical expressions in a highly technological world is crucial. Their meta-analysis also indicated that statistical expressions of proof are generally persuasive. A Weibo message that contains numerical evidence is considered a statistical expression message. Visuals [36] refer to adding descriptive pictures or videos to a Weibo message. Through literature review and the research completed by our group [25,37], we found that visuals [38,39] and statistics [39] were two important message features that can be combined with framing effects to affect the influence of health messages. It is easy for users to detect pictures, videos, or precise numbers in Weibo messages, and it is also easy for message creators to add this content to Weibo messages.

Evaluation Indicator of the Influence of Health-Promoting Messages in Weibo

In our study, influence was defined as the degree to which a Weibo health-promoting message attracts users to participate in the message interaction, which also evaluates the effectiveness of health communication [40]. Shiratuddin et al [41] and Hassan and Shiratuddin [42] found that retweets, comments, and likes can reflect a user group's participation and attention to the content of Weibo messages. Retweets constitute the crucial

mechanism of message diffusion on Weibo. Retweets are related to a variety of social motivations, such as spreading information to new audiences, pleasing specific audiences, publicly supporting someone, quoting others' views, and symbolizing friendship, loyalty, or respect. Starbird and Palen [43] believe that retweeting is a kind of information recommendation behavior. Comments and likes are the main ways for users to interact on Weibo. Comments on Weibo refer to users' personal opinions on a topic, according to their preferences and other subjective demands. A "like" button, represented by a thumbs up symbol, is present at the bottom of every post by the social network users. That symbol is clicked to express love and approval for a particular statement. The "like" option is easy to operate, thereby making the expression convenient and fast.

These 3 behaviors' costs mainly include time costs and credit costs. The time costs and credit costs for retweets, comments, and likes are different. Compared with the other two operations, the time costs and credit costs of the like operation are the lowest. The comment operation requires the highest time costs and some credit costs. The retweet operation requires few time costs and the highest credit costs. To summarize, we propose that when a Weibo message attracts users to participate in the interaction, the cost of commenting is the highest, retweeting is the second highest, and liking is the lowest. This means that the weights of likes (Lw), retweets (Rw), and comments (Cw) are different for the mathematical expression of the influence score. Therefore, the influence score (Is) of a Weibo message can be defined as the linear weighted sum of the number of retweets, comments, and likes. This can be expressed as:

$$Is = \alpha Lw + \beta Rw + \gamma Cw \quad \textbf{(1)}$$

$$\alpha + \beta + \gamma = 1 \quad \textbf{(2)}$$

$$\alpha < \beta < \gamma \quad \textbf{(3)}$$

Xiong et al [44] found that the number of retweets and comments had a positive correlation in a big data study on Weibo messages. The quantitative relationship between β and γ was obtained:

$$\beta:\gamma = 0.84 \quad \textbf{(4)}$$

Based on the data set of messages in the Weibo hot topic list, Wu [45] found the ratio of the number of likes to the sum of the number of retweets and comments in the Weibo messages in which users actively participated in the interaction. The quantitative relationship between α and (β + γ) was obtained:

$$\alpha:(\beta + \gamma) = 0.25 \quad \textbf{(5)}$$

The values of α, β, and γ can be obtained simultaneously with formulas 1, 4, and 5:

$$Is = 0.2Lw + 0.365Rw + 0.435Cw.$$

Methods

Health Message Themes

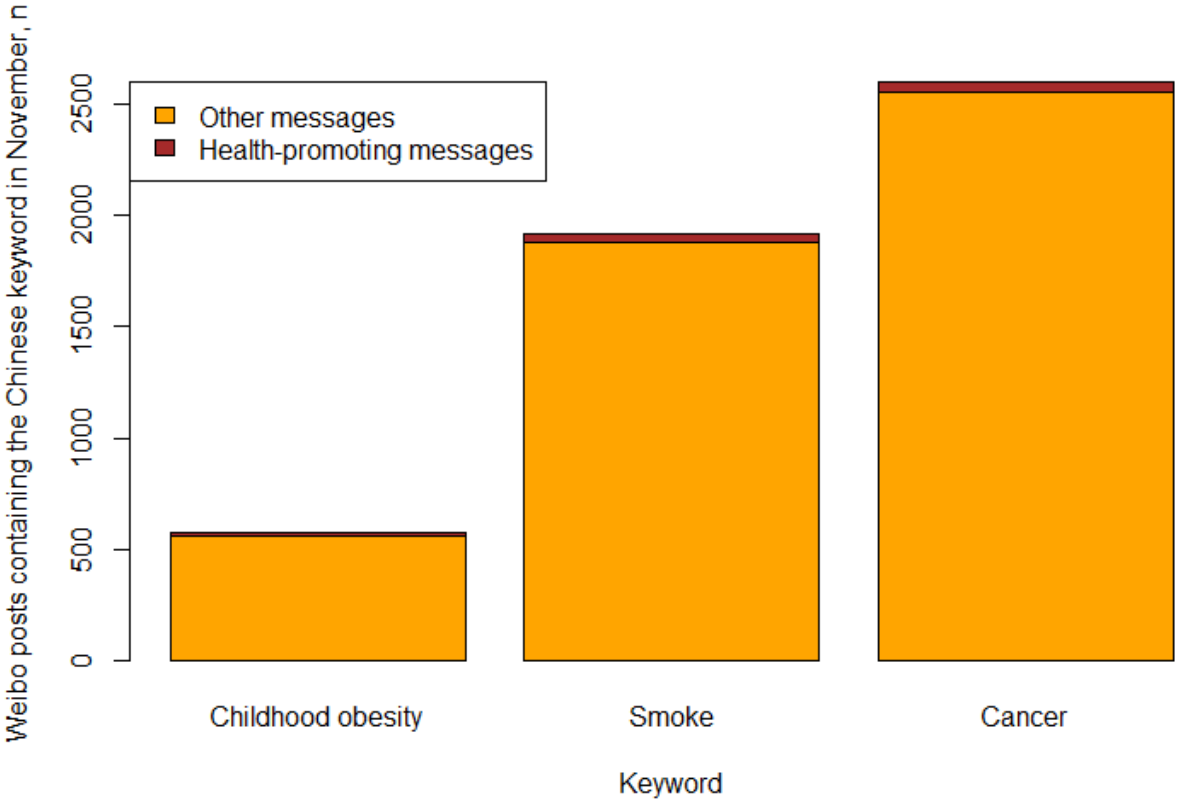
Because of the search method on Weibo, we first had to determine the keywords that could represent the health theme in order to search Weibo health messages. To identify health themes, we used the PubMed search engine, using "fram* effects" AND "health message*" as keywords, and obtained 229 papers. A panel of experts discussed health themes from these 229 papers. Through discussion, they found that obesity-related, smoking-related, and cancer-related health themes were mentioned more in the literature and that the public paid high attention to them. Combined with the previous research of our group [25], we ultimately chose "childhood obesity," "smoking," and "cancer" as keywords.

Using Python, we wrote a crawler program that could automatically obtain the fields in the Weibo platform according to the given keywords. The resulting fields included the message text, the publisher's name, and the number of retweets, comments, likes, pictures, and videos.

Time Ranges of Health-Promoting Messages

Weibo allows researchers to retrieve messages posted during a specified period through keyword searches. With childhood obesity, smoking, and cancer as the keywords, we retrieved posts for 1 month (November 1 to November 30, 2019). The 2 coders counted the number of health-promoting messages in the search posts and used the Cohen κ coefficient to ensure the consistency between the coders. The statistics of the 2 coders revealed that the proportion of health-promoting messages containing the keywords was 1:4.4:5.1 (childhood obesity:smoking:cancer) (Figure 1). In order to ensure that the number of health-promoting messages in the 3 themes remained similar, we used childhood obesity as the keyword and searched for posts from January 1, 2019, to January 31, 2020 (13 months). Moreover, we employed smoking and cancer as the keywords and searched posts from November 1, 2019, to January 31, 2020 (3 months).

Figure 1. Number of Weibo messages that contained the Chinese keyword in the Weibo database in November 2019. Health-promoting message content had to include health-related behaviors and the consequences of behaviors.

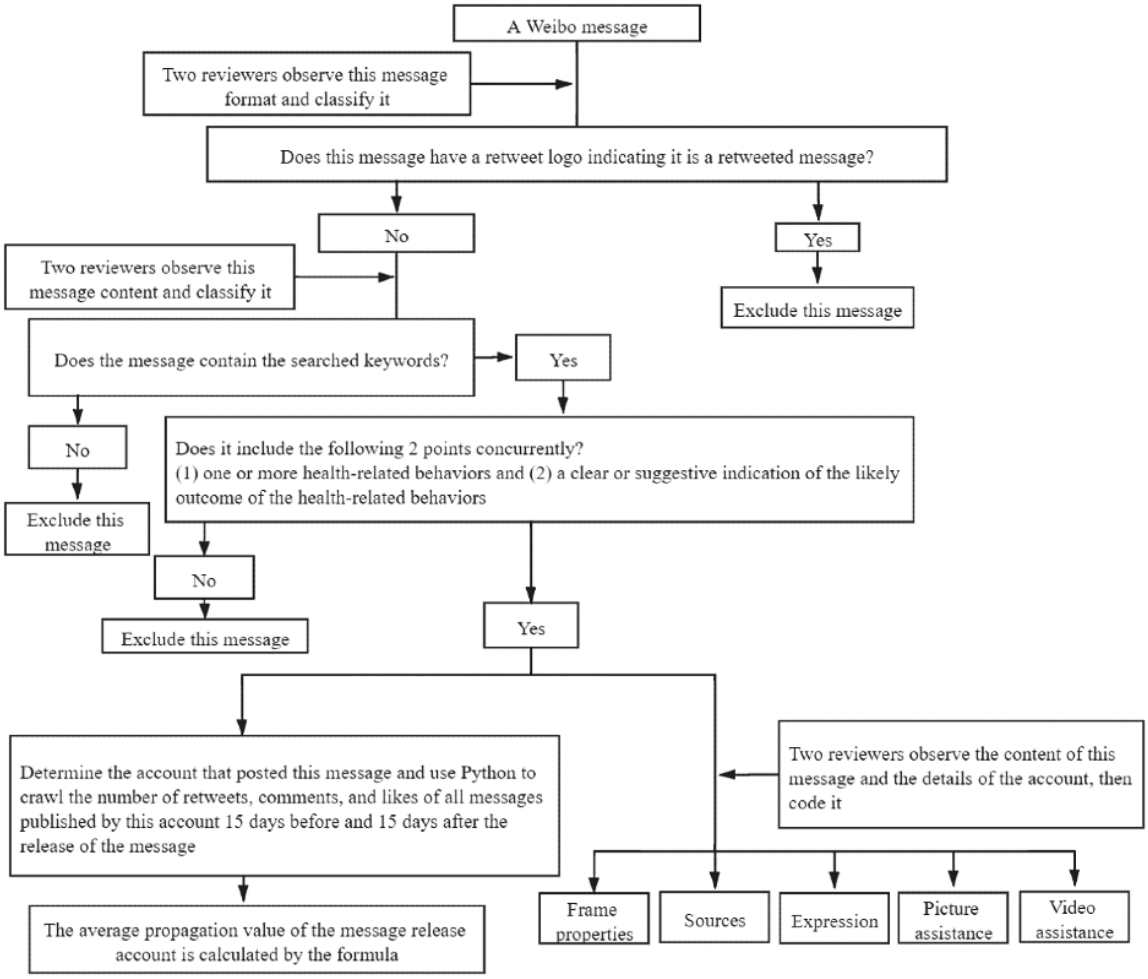


Coding of Health-Promoting Messages

We used Python's Selenium library to simulate users' log-ins to Weibo's webpage by employing a written crawler code and using Weibo's advanced collection mechanism to search for messages with the keywords. The crawler was designed to output the search results, including the messages; the account names of the messages; the number of retweets, comments, and likes on messages; and any pictures or videos.

Two reviewers then screened the health-promoting messages from the search results according to the coding process (Figure 2) and coded the degree of influence, frame properties, influence source type, expression type, and presence of pictures and videos, as described in Multimedia Appendix 1 [9,31,45]. The 2 coders studied the coding principles carefully, and pre-experiment coding was carried out. In the pre-experiment coding, the 2 coders communicated effectively. Results that were similarly coded by the 2 reviewers could be entered directly. Messages with different codes were re-examined by a reviewer. Then, after eliminating human error, the coding was submitted to an expert group for judgment if the outcome was inconsistent.

Figure 2. The coding process for the health-promoting messages.



Statistical Analysis

Data were processed with Excel (Microsoft Corp) software before entry into the database. Data analyses were performed using SPSS 20.0 software (IBM Corp). Frequency and percentage were used to describe the categorical variables of the message characteristics. Binary logistic regression analysis was implemented to evaluate the influencing factors related to the health-promoting messages' influence. *P* values below .05 were considered statistically significant.

Quality Control

Before the formal experiment, 2 coders randomly encoded 1800 Weibo messages. Cohen κ was used to measure the consistency in SPSS 20.0. The Cohen κ coefficient of the coders was 0.827 when judging whether a message was a health-promoting message, 0.962 when judging the influence of a health-promoting message, 0.859 when judging the frame properties of a health-promoting message, 0.943 when judging the expression properties of a health-promoting message, 0.977 when judging the sources of a health-promoting message, and 1.000 when judging whether a health-promoting message contained a picture or video. The two coders had good consistency and met the coding requirements of content analysis.

Results

Descriptive Statistics of the Characteristics of Health-Promoting Messages

A total of 389 health-promoting messages were included in this study (Table 1). Among these messages, 242 (62.1%) were lower than the average influence score of the disseminator. The chosen items used loss-framed (241/389, 61.8%), gain-framed (127/389, 32.6%), and neutral-framed (21/389, 5.4%) messages. The disseminators of these health-promoting messages were mostly individual users, 31.3% (122/389) of whom possessed professional accounts certified by the platform or were labelled as engaged in the health field and 26.7% (104/389) of whom were not certified by the platform nor labelled as engaged in the health field. Disseminators not from health institutions and organizations (86/389, 22.1%) outnumbered those from health institutions and organizations (77/389, 19.7%). A total of 44.1% (172/389) of the health-promoting messages contained statistical expressions. Furthermore, 51.3% (200/389) of the health-promoting messages added corresponding pictures in addition to the text description. However, only 20.0% (78/389) of the messages added corresponding videos.

Table 1. Descriptive statistics of characteristics of health-promoting messages.

Variables	Childhood obesity, n (%) (n=128)	Smoking, n (%) (n=114)	Cancer, n (%) (n=147)	Total, n (%) (N=389)
Influence degree				
High influence	45 (35.2)	36 (28.1)	66 (44.9)	147 (37.7)
Low influence	83 (64.8)	78 (60.9)	81 (55.1)	242 (62.1)
Frame properties				
Loss framed	74 (57.8)	94 (73.4)	73 (49.7)	241 (61.8)
Gain framed	48 (37.5)	17 (13.3)	62 (42.2)	127 (32.6)
Neutral framed	6 (4.7)	3 (2.3)	12 (8.2)	21 (5.4)
Sources				
Ordinary users (health field)	41 (32.0)	28 (21.9)	53 (36.1)	122 (31.3)
Ordinary users (nonhealth field)	35 (27.3)	29 (22.7)	40 (27.2)	104 (26.7)
Organizations (health field)	27 (21.1)	29 (22.7)	21 (14.3)	77 (19.7)
Organizations (nonhealth field)	25 (19.5)	28 (21.9)	33 (22.4)	86 (22.1)
Expression properties				
Statistical expression	54 (42.2)	67 (52.3)	51 (34.7)	172 (44.1)
Nonstatistical expression	74 (57.8)	47 (36.7)	96 (65.3)	217 (55.6)
Picture assistance				
Yes	77 (60.2)	48 (37.5)	75 (51.0)	200 (51.3)
No	51 (39.8)	66 (51.6)	72 (49.0)	189 (48.5)
Video assistance				
Yes	15 (11.7)	24 (29.7)	39 (26.5)	78 (20.0)
No	113 (88.3)	90 (70.3)	108 (73.5)	311 (79.7)

Binary Logistic Regression of the Information Characteristics and Degree of Dissemination

The influence of each health-promoting message was a dichotomous variable. In this study, we used binary logistic regression to evaluate the impact of framing effects, information sources, expression types, and pictures and videos on the degree of influence of health-promoting messages. The dependent variable of the binary logistic regression model was based on low influence.

We analyzed the 3 focal health message themes and found that the effect of the message characteristics on the message influence did not change with the alteration of the health message themes. For the messages on childhood obesity and cancer, the frame properties and whether the message was a statistical expression had an impact on the message influence. Compared with loss-framed messages, gain-framed messages had a higher degree of message influence ($P<.001$ in childhood obesity and cancer) and used statistical expressions with a higher degree of message influence ($P=.02$ in childhood obesity, $P<.001$ in cancer). In the messages about smoking, the frame properties (whether statistical expressions or otherwise) and the inclusion or exclusion of videos had an impact on the message influence. Compared with the loss-framed messages, the gain-framed messages had a higher degree of message influence ($P=.03$). Messages with statistical expressions had a higher degree of message influence than those with nonstatistical expressions ($P=.002$). Finally, messages with videos had a higher degree of message influence than those without videos ($P=.009$). The effect of the message characteristics on the influence of messages about childhood obesity, smoking, and cancer can be seen in Tables 2-4.

Table 2. Binary logistic regression for health-promoting messages regarding childhood obesity.

Parameter	β	SE	Wald chi-square (*df*)	OR[a] (95% CI)	*P* value
Frame properties					
Gain framed	–3.210	0.571	31.6 (1)	0.040 (0.013-0.124)	<.001
Neutral framed	0.280	1.227	0.1 (1)	1.323 (0.119-14.661)	.82
Loss framed (ref[b])	N/A[c]	N/A	N/A	N/A	N/A
Sources					
Ordinary users (health field)	–0.072	0.739	0.0 (1)	0.931 (0.219-3.962)	.92
Ordinary users (nonhealth field)	0.462	0.724	0.4 (1)	1.587 (0.384-6.562)	.52
Organizations (health field)	–0.958	0.785	1.5 (1)	0.384 (0.082-1.788)	.22
Organizations (nonhealth field) (ref)	N/A	N/A	N/A	N/A	N/A
Expression					
Statistical expression	–1.227	0.525	5.5 (1)	0.293 (0.105-0.820)	.02
Nonstatistical expression (ref)	N/A	N/A	N/A	N/A	N/A
Picture assistance					
Yes	–0.162	0.609	0.1 (1)	0.850 (0.258-2.806)	.79
No (ref)	N/A	N/A	N/A	N/A	N/A
Video assistance					
Yes	–0.334	0.890	0.1 (1)	0.716 (0.125-4.098)	.71
No (ref)	N/A	N/A	N/A	N/A	N/A

[a]OR: odds ratio.

[b]ref: reference category.

[c]N/A: not applicable.

Table 3. Binary logistic regression for health-promoting messages regarding smoking.

Parameter	β	SE	Wald chi-square (*df*)	OR[a] (95% CI)	*P* value
Frame properties					
Gain framed	–1.412	0.641	4.9 (1)	0.244 (0.069-0.856)	.03
Neutral framed	–0.163	1.389	0.0 (1)	0.850 (0.850-12.932)	.91
Loss framed (ref[b])	N/A[c]	N/A	N/A	N/A	N/A
Sources					
Ordinary users (health field)	0.538	0.687	0.6 (1)	1.713 (0.446-6.580)	.43
Ordinary users (nonhealth field)	1.209	0.747	2.6 (1)	3.351 (0.776-14.475)	.11
Organizations (health field)	0.300	0.722	0.2 (1)	1.350 (0.328-5.555)	.68
Organizations (nonhealth field) (ref)	N/A	N/A	N/A	N/A	N/A
Expression					
Statistical expression	–1.932	0.609	10.1 (1)	0.145 (0.044-0.478)	.002
Nonstatistical expression (ref)	N/A	N/A	N/A	N/A	N/A
Picture assistance					
Yes	–0.879	0.631	1.9 (1)	0.415 (0.121-1.429)	.16
No (ref)	N/A	N/A	N/A	N/A	N/A
Video assistance					
Yes	–2.016	0.767	6.9 (1)	0.133 (0.030-0.599)	.009
No (ref)	N/A	N/A	N/A	N/A	N/A

[a]OR: odds ratio.

[b]ref: reference category.

[c]N/A: not applicable.

Table 4. Binary logistic regression for health-promoting messages regarding cancer.

Parameter	β	SE	Wald chi-square (*df*)	OR[a] (95% CI)	*P* value
Frame properties					
Gain framed	–2.808	0.482	33.9 (1)	0.060 (0.023-0.155)	<.001
Neutral framed	–1.019	0.757	1.8 (1)	0.361 (0.082-1.590)	.18
Loss framed (ref[b])	N/A[c]	N/A	N/A	N/A	N/A
Sources					
Ordinary users (health field)	0.671	0.587	1.3 (1)	1.955 (0.618-6.183)	.25
Ordinary users (nonhealth field)	0.461	0.623	0.5 (1)	1.586 (0.468-5.375)	.46
Organizations (health field)	–0.226	0.745	0.1 (1)	0.798 (0.185-3.436)	.76
Organizations (nonhealth field) (ref)	N/A	N/A	N/A	N/A	N/A
Expression					
Statistical expression	–1.714	0.473	13.1 (1)	0.180 (0.071-0.455)	<.001
Nonstatistical expression (ref)	N/A	N/A	N/A	N/A	N/A
Picture assistance					
Yes	–0.286	0.577	0.2 (1)	0.751 (0.242-2.328)	.62
No (ref)	N/A	N/A	N/A	N/A	N/A
Video assistance					
Yes	–0.109	0.661	0.0 (1)	0.897 (0.245-3.277)	.87
No (ref)	N/A	N/A	N/A	N/A	N/A

[a]OR: odds ratio.

[b]ref: reference category.

[c]N/A: not applicable.

Discussion

Principal Results

To the best of our knowledge, this study has broken new ground in two aspects by (1) exploring the application of framing effects in social media and (2) providing ideas for drafting health-promoting messages with a high degree of influence.

First, the use of gain-framed messages in the health themes of childhood obesity (*P*<.001), smoking (*P*=.03), and cancer (*P*<.001) can significantly improve the influence of health-promoting messages (Table 2). Rothman and Salovey [46] and Rothman et al [47] divided health behaviors into prevention and detection behaviors according to the risk perception of individuals. Preventive behaviors include exercise, quitting smoking, eating healthy, and using sunscreen. They believed that gain-framed messages were more persuasive in promoting disease prevention behaviors. Goal-framing effects based on prospect theory have also revealed that factually equivalent messages have different levels of persuasiveness depending on the frame adopted by the messages [48]. Gallagher and Updegraff [49] believed that gain-framed health-promoting messages stimulated more information processing and better subsequent memory. Furthermore, many previous investigations support our conclusion. A cross-sectional study of 592 caregivers of preschool children found that gain-framed messages could significantly improve the acceptance of information by caregivers [25]. Romanowich and Lamb [50] posited that health education using gain-framed messages could be more useful for nonsmokers. A qualitative survey of African American adolescents by Satia et al [51] indicated greater consistency with gain-framed cancer prevention messages. Most research on the framing effects of health-promoting messages have been conducted by questionnaires or interviews [15]. However, the real world involves people observing information and making decisions in a complex environment [52]. This study further confirmed that gain-framed messages are a favorable strategy in the dissemination of health-promoting messages in everyday life.

Second, the use of statistical expression in the health themes of childhood obesity (*P*=.02), smoking (*P*=.002), and cancer (*P*<.001) can significantly improve the influence of health-promoting messages (Table 2). Statistical expressions refer to health-promoting messages with numerical content [53]. Nonstatistical expressions denote health-promoting messages without any precise numbers and are usually used in the description of examples and stories [53]. No conclusion has been reached about the persuasiveness of these two expression types [34]. A meta-analysis [23] and an investigation of 1270 participants [54] found that statistical messages were more convincing than narrative ones. A message that combines narrative and statistical expression is more convincing than one using either narrative or statistical expression alone. We hypothesized that adding statistical expressions to

health-promoting messages when describing health behaviors and consequences could create more active engagement toward those messages and earn them more retweets, comments, and likes. Another meta-analysis also revealed that statistical expression has a stronger impact on beliefs and attitudes than narrative expression and that statistical expressions, beliefs, and attitudes are mainly related to cognitive responses [55]. Wong et al [39] combined numerical framing effects and prospect theory and verified that precise numbers could more easily represent the probability of risk. We believe this finding may explain why people pay more attention to health-promoting messages that contain statistical expressions.

Third, the use of videos significantly improved the influence of health-promoting messages only for messages regarding smoking (P=.009). A review suggested that compared with text alone, adding pictures that are closely related to the written text can significantly improve the attention to and recall of health education information [56]. However, Houts et al [56] noted that great care should be taken when including picture materials in health messages so that the audiences can understand the key points of the message without being distracted by irrelevant details. Levie and Lentz [57] posited that pictures not closely related to the text have no beneficial effect on comprehension. Furthermore, the impact of using videos on health-promoting messages may be uncertain. Occa and Suggs [58] found that videos had a positive impact when communicating breast cancer information to 194 Italian women. Conversely, Xie [59] suggested that there was no significant difference in risk perception caused by words and sounds. We suggest that publishers add appropriate pictures and videos as much as possible when making health-promoting messages.

Fourth, we believe that accounts from organizations in health fields should release health-promoting messages more actively. People tend to trust universities and official institutions more than other types of organizations [60]. Furthermore, people consider private doctors, medical universities, and governments the most trusted sources of health messages [44,45]. As shown in Table 1, the proportion of accounts from individual users was higher than for organizational users, and users engaged in the health field outnumbered those in nonhealth fields. Specific audiences are more willing to believe that the most reliable information is provided by accounts from a health field [24]. Thus, ordinary users and organizations from health fields must participate in the dissemination of health-promoting messages.

Fifth, we found that health-promoting messages account for a very small proportion of the social media posts related to the 3 health themes. We thought this may be related to four main reasons. First, the definition of a health-promoting message was a message that included both health behaviors and health outcomes [15], so we excluded some messages, such as messages only referring to the cause of a disease. Second, in a health topic, the amount of social content was often much larger than the amount of health professional content [61]. Third, we found that the celebrity effect exists in Weibo health themes. If a celebrity died of cancer, there would soon be a lot of cancer-related messages on Weibo. However, there was a lack of clear health guidance in these messages. We excluded a lot of these kinds of eye-catching messages. Fourth, even under the theme of health, a lot of messages on social media were still related to advertising [62]. We excluded the messages that contained advertising. This result reflects reality. There were few health-promoting messages published on Weibo about childhood obesity, smoking, and cancer. In the health field, many researchers have confirmed that health-promoting messages using framing effects can stimulate people's health awareness [63] and improve their willingness to prevent and treat health conditions [20,49]. Health promotion messages including health behaviors and health outcomes should be widely used. Our research suggests that professionals in health fields should be more active in publishing health-promoting messages on social media.

Limitations

This study has several limitations. First, a small sample size was analyzed in this work. We needed to explore the impact of goal-framing effects on the influence of health-promoting messages. Accordingly, we included only health-promoting messages, such as those about behaviors and consequences [15], thereby limiting the writing template for such messages. This approach did not fully incorporate all health messages and may have produced errors. Moreover, this research only focused on certain health-promoting messages to fill the gaps in the literature. Second, we formulated the definition of a Weibo message's degree of influence according to the literature [41,42] and from expert advice, and we used only 3 indexes: retweets, comments, and likes. The number of users who viewed a message is an excellent index to evaluate the degree of influence, but restrictions of the Weibo platform meant that we could not ascertain the number of views for every message. If the Weibo platform cancels this restriction in the future, then views should be included in the evaluation index. Third, we did not evaluate whether the pictures and videos added in the Weibo health-promoting messages accurately matched the points of the message. Such an omission may have affected our results. We still suggest that the publisher add appropriate pictures and videos as much as possible to health-promoting messages. In the future, researchers can further examine the impact of pictures and videos on health-promoting messages on social media.

Conclusions

In this study, we identified the factors that could affect the degree of influence of health-promoting messages on the Sina Weibo platform. A total of 389 health-promoting messages were included in this work. The use of gain-framed messages and statistical expressions could improve the influence of messages for all 3 themes (ie, childhood obesity, smoking, and cancer). Although adding pictures and videos to messages did not significantly improve the influence of messages about childhood obesity and cancer, we still contend that adding appropriate pictures and videos as much as possible when producing health-promoting messages is a good strategy. We encourage users from organizations in health fields to release more health-promoting messages. When public health institutions and professionals release such messages, the framework, organization, and content of the messages must be considered. In this way, health-promoting messages may become more influential.

Acknowledgments

This research was funded by the Student Research and Innovation Experiment Project of the College of Medical Informatics, Chongqing Medical University, China (grant number 2019C002); the Special Research Project of Philosophy and Social Science of Chongqing Medical University, China (grant number 201712); the Intelligent Medicine Research Project of Chongqing Medical University in 2019 (ZHYX2019006); the Chinese Scholarship Council of the Ministry of Education (grant number 201808505165); and the Ministry of Education in China's Project of Humanities and Social Sciences (research on the mechanism and intervention countermeasures of framing effects in the formation of adolescent health literacy No. 20YJCZH043). The funders had no role in the design of the study; the collection, analyses, or interpretation of data; the writing of the manuscript; or the decision to publish the results.

Authors' Contributions

XH and QR designed the experiments. QR and ZZ performed the experiments. QR analyzed the data. YL helped analyze the data. XH, YZ, LB, and ZZ helped draft the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Coding principle for the health-promoting messages.
[DOCX File , 18 KB - medinform_v8i10e20558_app1.docx]

References

1. Yang F, Wendorf Muhamad J, Yang Q. Exploring Environmental Health on Weibo: A Textual Analysis of Framing Haze-Related Stories on Chinese Social Media. Int J Environ Res Public Health 2019 Jul 04;16(13) [FREE Full text] [doi: 10.3390/ijerph16132374] [Medline: 31277378]
2. China Internet Information Center. The 44th Statistical Report on the Development of Internet in China 2019. China Internet Information Center. URL: http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201908/P020190830356787490958.pdf [accessed 2020-02-20]
3. 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 Apr 23;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
4. Pagoto S, Waring ME, Xu R. A Call for a Public Health Agenda for Social Media Research. J Med Internet Res 2019 Dec 19;21(12):e16661 [FREE Full text] [doi: 10.2196/16661] [Medline: 31855185]
5. Yaya S, Uthman O, Amouzou A, Bishwajit G. Mass media exposure and its impact on malaria prevention behaviour among adult women in sub-Saharan Africa: results from malaria indicator surveys. Glob Health Res Policy 2018;3:20 [FREE Full text] [doi: 10.1186/s41256-018-0075-x] [Medline: 29998191]
6. The White Book of Chinese Weibo's Market. Sina Weibo. 2019. URL: http://data.weibo.com/report/reportDetail?id=433 [accessed 2020-09-07]
7. Alhabash S, McAlister AR. Redefining virality in less broad strokes: Predicting viral behavioral intentions from motivations and uses of Facebook and Twitter. 2014 Feb 24;17(8):1317-1339. [doi: 10.1177/1461444814523726]
8. Brooke C. Critique of Information. 2003 Mar;16(1):111-114. [doi: 10.1108/itp.2003.16.1.111.1]
9. Chou WS, Hunt YM, Beckjord EB, Moser RP, Hesse BW. Social media use in the United States: implications for health communication. J Med Internet Res 2009 Nov 27;11(4):e48 [FREE Full text] [doi: 10.2196/jmir.1249] [Medline: 19945947]
10. Jiang S, Beaudoin CE. Smoking Prevention in China: A Content Analysis of an Anti-Smoking Social Media Campaign. J Health Commun 2016 Jul;21(7):755-764. [doi: 10.1080/10810730.2016.1157653] [Medline: 27232655]
11. Wang S, Paul MJ, Dredze M. Social media as a sensor of air quality and public response in China. J Med Internet Res 2015 Mar 26;17(3):e22 [FREE Full text] [doi: 10.2196/jmir.3875] [Medline: 25831020]
12. Guo Y, Goh DH. "I Have AIDS": Content analysis of postings in HIV/AIDS support group on a Chinese microblog. Computers in Human Behavior 2014 May;34:219-226. [doi: 10.1016/j.chb.2014.02.003]
13. Adams SA. Revisiting the online health information reliability debate in the wake of "web 2.0": an inter-disciplinary literature and website review. Int J Med Inform 2010 Jun;79(6):391-400. [doi: 10.1016/j.ijmedinf.2010.01.006] [Medline: 20188623]
14. Jackson L. Information Complexity and Medical Communication: The Effects of Technical Language and Amount of Information in a Medical Message. Health Communication 1992 Jul;4(3):197-210. [doi: 10.1207/s15327027hc0403_3]

15. Van 't Riet J, Cox AD, Cox D, Zimet GD, De Bruijn G, Van den Putte B, et al. Does perceived risk influence the effects of message framing? Revisiting the link between prospect theory and message framing. Health Psychol Rev 2016 Dec;10(4):447-459. [doi: 10.1080/17437199.2016.1176865] [Medline: 27062974]
16. McCray AT. Promoting health literacy. J Am Med Inform Assoc 2005;12(2):152-163. [doi: 10.1197/jamia.M1687] [Medline: 15561782]
17. Shi J, Mao A, Liu C, Dong P, Ren J, Wang K, et al. [Health literacy and awareness of cancer control in urban China, 2005-2017: overall design of a national multicenter survey]. Zhonghua Yu Fang Yi Xue Za Zhi 2020 Jan 06;54(1):108-112. [doi: 10.3760/cma.j.issn.0253-9624.2020.01.020] [Medline: 31914578]
18. Cao W, Zhang X, Xu K, Wang Y. Modeling Online Health Information-Seeking Behavior in China: The Roles of Source Characteristics, Reward Assessment, and Internet Self-Efficacy. Health Commun 2016 Sep;31(9):1105-1114. [doi: 10.1080/10410236.2015.1045236] [Medline: 26861963]
19. Dunn HK, Pearlman DN, Beatty A, Florin P. Psychosocial Determinants of Teens' Online Engagement in Drug Prevention Social Media Campaigns: Implications for Public Health Organizations. J Prim Prev 2018 Oct;39(5):469-481. [doi: 10.1007/s10935-018-0522-y] [Medline: 30194518]
20. Myers R. Promoting healthy behaviors: how do we get the message across? Int J Nurs Stud 2010 Apr;47(4):500-512. [doi: 10.1016/j.ijnurstu.2009.11.017] [Medline: 20031126]
21. Meppelink CS, van Weert JCM, Haven CJ, Smit EG. The effectiveness of health animations in audiences with different health literacy levels: an experimental study. J Med Internet Res 2015 Jan 13;17(1):e11 [FREE Full text] [doi: 10.2196/jmir.3979] [Medline: 25586711]
22. Meppelink CS, Smit EG, Buurman BM, van Weert JCM. Should We Be Afraid of Simple Messages? The Effects of Text Difficulty and Illustrations in People With Low or High Health Literacy. Health Commun 2015;30(12):1181-1189. [doi: 10.1080/10410236.2015.1037425] [Medline: 26372031]
23. Allen M, Preiss RW. Comparing the persuasiveness of narrative and statistical evidence using meta‐analysis. Communication Research Reports 1997 Mar;14(2):125-131. [doi: 10.1080/08824099709388654]
24. Sundar SS. The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. In: Metzger MJ, Flanagin AJ, editors. Digital Media, Youth, and Credibility. Cambridge, MA: MIT Press; 2008:73-100.
25. Rao Q, Bai L, Lv Y, Abdullah AS, Brooks I, Xie Y, et al. Goal-Framing and Temporal-Framing: Effects on the Acceptance of Childhood Simple Obesity Prevention Messages among Preschool Children's Caregivers in China. Int J Environ Res Public Health 2020 Jan 26;17(3):1-15 [FREE Full text] [doi: 10.3390/ijerph17030770] [Medline: 31991873]
26. Zhang D, Gu J, Shao R. A Cluster Analysis of College Students' Health Information Acquisition Channels: Active Seeking and Accidental Exposure. Chin J Journalism Commun 2015 May;37(05):81-93. [doi: 10.13495/j.cnki.cjjc.2015.05.006]
27. Kahneman D, Tversky A. Prospect Theory: An Analysis of Decision under Risk. Econometrica 1979 Mar;47(2):263. [doi: 10.2307/1914185]
28. Striegel-Moore RH, Thompson DR, Affenito SG, Franko DL, Barton BA, Schreiber GB, et al. Fruit and vegetable intake: Few adolescent girls meet national guidelines. Prev Med 2006 Mar;42(3):223-228. [doi: 10.1016/j.ypmed.2005.11.018] [Medline: 16406116]
29. van 't Riet J, Ruiter RAC, Werrij MQ, de Vries H. Investigating message-framing effects in the context of a tailored intervention promoting physical activity. Health Educ Res 2010 Apr;25(2):343-354. [doi: 10.1093/her/cyp061] [Medline: 19841041]
30. Iannotti RJ, Finney LJ, Sander AA, De Leon JM. Effect of clinical breast examination training on practitioner's perceived competence. Cancer Detection and Prevention 2002 May;26(2):146-148. [doi: 10.1016/s0361-090x(02)00029-6]
31. Edwards A, Elwyn G, Covey J, Matthews E, Pill R. Presenting risk information--a review of the effects of "framing" and other manipulations on patient outcomes. J Health Commun 2001;6(1):61-82. [doi: 10.1080/10810730150501413] [Medline: 11317424]
32. Fung IC, Cai J, Hao Y, Ying Y, Chan BSB, Tse ZTH, et al. Global Handwashing Day 2012: a qualitative content analysis of Chinese social media reaction to a health promotion event. Western Pac Surveill Response J 2015;6(3):34-42 [FREE Full text] [doi: 10.5365/WPSAR.2015.6.2.003] [Medline: 26668765]
33. Kahneman D, Tversky A. On the psychology of prediction. Psychol Rev 1973;80(4):237-251. [doi: 10.1037/h0034747]
34. Limon MS, Kazoleas DC. A comparison of exemplar and statistical evidence in reducing counter‐arguments and responses to a message. Communication Research Reports 2004 Jun;21(3):291-298. [doi: 10.1080/08824090409359991]
35. Reinard JC. The Empirical Study of the Persuasive Effects of Evidence: The Status After Fifty Years of Research. Human Comm Res 1988 Sep;15(1):3-59. [doi: 10.1111/j.1468-2958.1988.tb00170.x]
36. Li J, Xu Q, Cuomo R, Purushothaman V, Mackey T. Data Mining and Content Analysis of the Chinese Social Media Platform Weibo During the Early COVID-19 Outbreak: Retrospective Observational Infoveillance Study. JMIR Public Health Surveill 2020 Apr 21;6(2):e18700 [FREE Full text] [doi: 10.2196/18700] [Medline: 32293582]
37. Bai L, Cai Z, Lv Y, Wu T, Sharma M, Shi Z, et al. Personal Involvement Moderates Message Framing Effects on Food Safety Education among Medical University Students in Chongqing, China. Int J Environ Res Public Health 2018 Sep 19;15(9) [FREE Full text] [doi: 10.3390/ijerph15092059] [Medline: 30235903]

38. Suka M, Yamauchi T, Yanagisawa H. Comparing responses to differently framed and formatted persuasive messages to encourage help-seeking for depression in Japanese adults: a cross-sectional study with 2-month follow-up. BMJ Open 2018 Nov 12;8(11):e020823. [doi: 10.1136/bmjopen-2017-020823] [Medline: 30420341]
39. Wong K, Kwong J. Comparing two tiny giants or two huge dwarfs? Preference reversals owing to number size framing. Organizational Behavior and Human Decision Processes 2005 Sep;98(1):54-65 [FREE Full text] [doi: 10.1016/j.obhdp.2005.04.002]
40. Liu X, Lu J, Wang H. When Health Information Meets Social Media: Exploring Virality on Sina Weibo. Health Commun 2017 Dec;32(10):1252-1260. [doi: 10.1080/10410236.2016.1217454] [Medline: 27668831]
41. Shiratuddin N, Hassan S, Hashim N, Sakdan M, Sajat M. Blog influence index: A measure of influential weblog. Int J Virtual Communities Soc Netw 2011;3(3):35-45. [doi: 10.4018/jvcsn.2011070103]
42. Hassan S, Shiratuddin N. Identifying criteria for measuring influence of social media. International Journal of Interactive Communication Systems and Technologies 2013;10(3):91. [doi: 10.5555/2820186.2820189]
43. Starbird K, Palen L. Pass it on?: Retweeting in mass emergency. 2010 Presented at: 7th International ISCRAM Conference; May 2010; Seattle, WA.
44. Xiong X, Zhou G, Huang Y, Ma J. Predicting Popularity of Tweets on Sina Weibo. J Inf Eng Univ 2012;13(4):496-502.
45. Wu H. Research on Active Microblog Prediction Based on LDA and Random Forest Model [In China]. Hefei University of Technology 2017.
46. Rothman AJ, Salovey P. Shaping perceptions to motivate healthy behavior: the role of message framing. Psychol Bull 1997 Jan;121(1):3-19. [doi: 10.1037/0033-2909.121.1.3] [Medline: 9000890]
47. Rothman AJ, Salovey P, Antone C, Keough K, Martin CD. The Influence of Message Framing on Intentions to Perform Health Behaviors. Journal Exp Soc Psychol 1993 Sep;29(5):408-433. [doi: 10.1006/jesp.1993.1019]
48. Levin IP, Schneider SL, Gaeth GJ. All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects. Organ Behav Hum Decis Process 1998 Nov;76(2):149-188. [doi: 10.1006/obhd.1998.2804] [Medline: 9831520]
49. Gallagher KM, Updegraff JA. Health message framing effects on attitudes, intentions, and behavior: a meta-analytic review. Ann Behav Med 2012 Feb;43(1):101-116. [doi: 10.1007/s12160-011-9308-7] [Medline: 21993844]
50. Romanowich P, Lamb RJ. The effect of framing incentives as either losses or gains with contingency management for smoking cessation. Addict Behav 2013 Apr;38(4):2084-2088 [FREE Full text] [doi: 10.1016/j.addbeh.2013.01.007] [Medline: 23403276]
51. Satia JA, Barlow J, Armstrong-Brown J, Watters JL. Qualitative Study to Explore Prospect Theory and Message Framing and Diet and Cancer Prevention-Related Issues Among African American Adolescents. Cancer Nursing 2010;33(2):102-109. [doi: 10.1097/ncc.0b013e3181be5e8a]
52. Sanders Thompson VL. Making decisions in a complex information environment: evidential preference and information we trust. BMC Med Inform Decis Mak 2013;13 Suppl 3:S7 [FREE Full text] [doi: 10.1186/1472-6947-13-S3-S7] [Medline: 24565305]
53. Han B, Fink EL. How Do Statistical and Narrative Evidence Affect Persuasion?: The Role of Evidentiary Features. Argumentation and Advocacy 2017 Feb 02;49(1):39-58. [doi: 10.1080/00028533.2012.11821779]
54. Allen M, Bruflat R, Fucilla R, Kramer M, McKellips S, Ryan D, et al. Testing the persuasiveness of evidence: Combining narrative and statistical forms. Communication Research Reports 2000 Sep;17(4):331-336. [doi: 10.1080/08824090009388781]
55. Zebregs S, van den Putte B, Neijens P, de Graaf A. The differential impact of statistical and narrative evidence on beliefs, attitude, and intention: a meta-analysis. Health Commun 2015;30(3):282-289. [doi: 10.1080/10410236.2013.842528] [Medline: 24836931]
56. Houts P, Doak C, Doak L, Loscalzo M. The role of pictures in improving health communication: a review of research on attention, comprehension, recall, and adherence. Patient Educ Couns 2006 May;61(2):173-190. [doi: 10.1016/j.pec.2005.05.004] [Medline: 16122896]
57. Levie W, Lentz R. Effects of text illustrations: A review of research. Edu Tech Res Dev 1982;30(4):232. [doi: 10.1007/BF02765184]
58. Occa A, Suggs L. Communicating Breast Cancer Screening With Young Women: An Experimental Test of Didactic and Narrative Messages Using Video and Infographics. J Health Commun 2016;21(1):1-11. [doi: 10.1080/10810730.2015.1018611] [Medline: 26147625]
59. Xie X. How Can a Risk Be Increased? An Analysis of Risk Communication Channels. Acta Psychologica Sinica 2008 Sep 19;40(4):456-465. [doi: 10.3724/sp.j.1041.2008.00456]
60. Briggs P, Burford B, De Angeli A, Lynch P. Trust in Online Advice. Social Science Computer Review 2016 Aug 18;20(3):321-332. [doi: 10.1177/089443930202000309]
61. Unger JB, Escobedo P, Allem J, Soto DW, Chu K, Cruz T. Perceptions of Secondhand E-Cigarette Aerosol Among Twitter Users. Tob Regul Sci 2016 Apr;2(2):146-152 [FREE Full text] [doi: 10.18001/TRS.2.2.5] [Medline: 28090560]
62. Freeman B, Chapman S. Is "YouTube" telling or selling you something? Tobacco content on the YouTube video-sharing website. Tob Control 2007 Jun;16(3):207-210. [doi: 10.1136/tc.2007.020024] [Medline: 17565142]

63. O'Connor AM, Pennie RA, Dales RE. Framing effects on expectations, decisions, and side effects experienced: the case of influenza immunization. J Clin Epidemiol 1996 Nov;49(11):1271-1276. [doi: 10.1016/s0895-4356(96)00177-1] [Medline: 8892495]

Edited by C Lovis; submitted 21.05.20; peer-reviewed by J Liu, K Reuter; comments to author 28.07.20; revised version received 21.08.20; accepted 06.09.20; published 09.10.20.

Please cite as:
Rao Q, Zhang Z, Lv Y, Zhao Y, Bai L, Hou X
Factors Associated With Influential Health-Promoting Messages on Social Media: Content Analysis of Sina Weibo
JMIR Med Inform 2020;8(10):e20558
URL: http://medinform.jmir.org/2020/10/e20558/
doi:10.2196/20558
PMID:33034569

Original Paper

Construction of a Digestive System Tumor Knowledge Graph Based on Chinese Electronic Medical Records: Development and Usability Study

Xiaolei Xiu[1*], MSc; Qing Qian[1*], MSc; Sizhu Wu[1], PhD

Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

*these authors contributed equally

Corresponding Author:
Sizhu Wu, PhD
Institute of Medical Information/Medical Library
Chinese Academy of Medical Sciences & Peking Union Medical College
No 3 Yabao Road
Chaoyang District
Beijing, 100020
China
Phone: 86 18510495073
Email: wu.sizhu@imicams.ac.cn

Abstract

Background: With the increasing incidences and mortality of digestive system tumor diseases in China, ways to use clinical experience data in Chinese electronic medical records (CEMRs) to determine potentially effective relationships between diagnosis and treatment have become a priority. As an important part of artificial intelligence, a knowledge graph is a powerful tool for information processing and knowledge organization that provides an ideal means to solve this problem.

Objective: This study aimed to construct a semantic-driven digestive system tumor knowledge graph (DSTKG) to represent the knowledge in CEMRs with fine granularity and semantics.

Methods: This paper focuses on the knowledge graph schema and semantic relationships that were the main challenges for constructing a Chinese tumor knowledge graph. The DSTKG was developed through a multistep procedure. As an initial step, a complete DSTKG construction framework based on CEMRs was proposed. Then, this research built a knowledge graph schema containing 7 classes and 16 kinds of semantic relationships and accomplished the DSTKG by knowledge extraction, named entity linking, and drawing the knowledge graph. Finally, the quality of the DSTKG was evaluated from 3 aspects: data layer, schema layer, and application layer.

Results: Experts agreed that the DSTKG was good overall (mean score 4.20). Especially for the aspects of "rationality of schema structure," "scalability," and "readability of results," the DSTKG performed well, with scores of 4.72, 4.67, and 4.69, respectively, which were much higher than the average. However, the small amount of data in the DSTKG negatively affected its "practicability" score. Compared with other Chinese tumor knowledge graphs, the DSTKG can represent more granular entities, properties, and semantic relationships. In addition, the DSTKG was flexible, allowing personalized customization to meet the designer's focus on specific interests in the digestive system tumor.

Conclusions: We constructed a granular semantic DSTKG. It could provide guidance for the construction of a tumor knowledge graph and provide a preliminary step for the intelligent application of knowledge graphs based on CEMRs. Additional data sources and stronger research on assertion classification are needed to gain insight into the DSTKG's potential.

(JMIR Med Inform 2020;8(10):e18287) doi:10.2196/18287

KEYWORDS

Chinese electronic medical records; knowledge graph; digestive system tumor; graph evaluation

Introduction

Background

Cancer is a leading cause of death worldwide. The International Agency for Research on Cancer estimates that there were 18.1 million new cases of cancer and 9.6 million deaths caused by cancer in 2018 [1]. Nearly 24% (4.3 million) of these cancer cases and 30% (2.9 million) of deaths occurred in China. Digestive tract cancers were responsible for 36.4% of cancer-related deaths in China, compared with <5% in both the United States and United Kingdom [2]. There is a rapid increase of digestive system cancers in China. China is currently challenged by trying to prevent and control digestive cancers.

Electronic medical records (EMRs) are digital versions of paper-based patient charts in clinician offices, clinics, and hospitals that contain detailed clinical information about the occurrence, development, and treatment of the patient's disease [3]. They have important clinical value, such as providing data to support disease screening and prediction [4]. With the full implementation of health information technologies in China, its hospitals have accumulated large amounts of EMRs. However, the utilization rate of EMRs in China is relatively low, and research still focuses on traditional data management and statistical analysis of data from small samples [5-7]. One reason is because medical knowledge in Chinese EMRs (CEMRs) mainly exists in unstructured text, which cannot be understood by computers. In addition, medical knowledge in CEMRs is scattered; for example, the main entity type in the chief complaint field is symptom, and the main entity type in the discharge instructions field is medicine. This scattered knowledge distribution introduces obstacles to the analysis and in-depth mining of CEMRs. Furthermore, CEMRs have a unique sentence structure that is different from ordinary text, and there is no unified clinical medical terminology standard for CEMRs, which results in different clinicians using different expressions for the same medical term. For example, for the Chinese medical term "radical gastrectomy (胃癌根治术)," clinicians may write "胃癌根治" or "根治性胃癌根治术." These characteristics of CEMRs have brought great challenges to the mining and utilization of CEMRs in the big data environment.

The knowledge graph is an emerging knowledge service technology in the era of big data and an important part of artificial intelligence [8]. It has a graph-based data structure composed of nodes (entities) and edges (semantic relations) [9]. Strengths of knowledge graphs are their abilities for information processing and knowledge organization; in addition, they can express various domain concepts and the intricate relationships between them. The knowledge graph provides an ideal technical means for connecting scattered knowledge fragments and integrating information in CEMRs. Using knowledge graph technology to organize and manage medical knowledge scattered in various parts of CEMRs can not only effectively describe and mine the relationship between medical entities and avoid information overload but also reduce the time cost for clinicians to find patient information and improve the knowledge service ability of CEMRs.

This study aimed to construct a semantic-driven digestive system tumor knowledge graph (DSTKG) based on CEMRs. Compared with previous studies, this study focused on the construction of a DSTKG schema and the representation of semantic relationships between medical concepts, with the aims of maximizing the presentation of diagnosis and treatment facts, better assisting knowledge calculation and completing knowledge graph construction, and the subsequent application of intelligent medicine. In addition, this paper introduces the characteristics of digestive system tumor diseases and tumor-related CEMRs for the purpose of improving the performance of knowledge extraction in the process of constructing a DSTKG. In the following sections, the framework of a DSTKG based on CEMRs is presented, and the construction process of a DSTKG is described in detail along with a preliminary assessment of the knowledge graph. The paper also discusses challenges and future steps.

Related Work

The concept of the knowledge graph was formally proposed by Google in 2012 and has been popularized in academia and industry since then. Medicine is an important vertical application field of knowledge graphs. So far, there have been a large number of medical knowledge graphs, such as IBM Watson Health [10], the Partitioned Knowledge Graph built by the University of Maryland [11], and the breast cancer knowledge graph built by Huang [12]. Furthermore, researchers have also begun to study how to construct high-quality health knowledge graphs from EMRs. For instance, Rotmensch et al [13] proposed a method to automatically construct a disease-symptom knowledge graph directly from EMRs using a noisy "OR" model [13]. Kwon et al [14] used interpretable and interactive recurrent neural networks for visual analytics on EMRs. Bean et al [15] applied a knowledge graph to verify adverse drug reactions in EMRs. Medical knowledge graphs have played an important role in medical services such as information retrieval, intelligent question-and-answer, and intelligent diagnosis.

Compared with English medical knowledge graphs, the research of Chinese medical knowledge graphs is still in its infancy, especially based on CEMRs. Part of the reason is that the resources for building Chinese medical knowledge graphs are limited. For instance, there are no public Chinese medical knowledge repositories like Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) and repositories of biomedical ontologies in Chinese like BioPortal. In addition, Chinese is different from English. Chinese has no natural space as a separator, fewer speech components, and random use of punctuation, which makes Chinese natural language processing more difficult. For CEMRs, in addition to the characteristics of Chinese that increase the difficulty of constructing knowledge graphs, their scattered knowledge distribution, unique syntax, a large number of abbreviations, and non-standard expressions of terminology further increase the difficulty of constructing Chinese medical knowledge graphs. Considering these challenges, researchers have made various attempts at creating processes for constructing Chinese medical knowledge graphs. For example, Zhang et al [16] proposed a generative model named the conditional relationship variational autoencoder to reduce the workload of data preprocessing and manual

annotation of the Chinese medical corpus. Various deep learning models were used to improve the performance of named entity recognition (NER) [17-19] and relation extraction (RE) of CEMRs [20,21]. Sheng et al [22] introduced a general framework for a health knowledge graph based on cardiovascular disease EMRs. Zhou et al [23] studied the construction and application of the "knowledge-centric" knowledge graph of traditional Chinese medicine based on ancient Chinese texts. Jie et al [24] built a breast tumor knowledge graph that only contains 3 types of concepts (basic information of the patient, examination, and diagnosis) and 7 kinds of one-way semantic relationships (has_a, instance_of, attribute_of, part_of, owns, diagnosis, and detect) [24]. However, with a one-way semantic relationship, it is difficult to fully express the complex medical process of patients. For example, the semantic relationship between disease and examination is not only the examination to investigate the disease but also the examination revealing the disease. So far, some medical knowledge graphs based on CEMRs have been established, such as those for hypertension and diabetes [25,26].

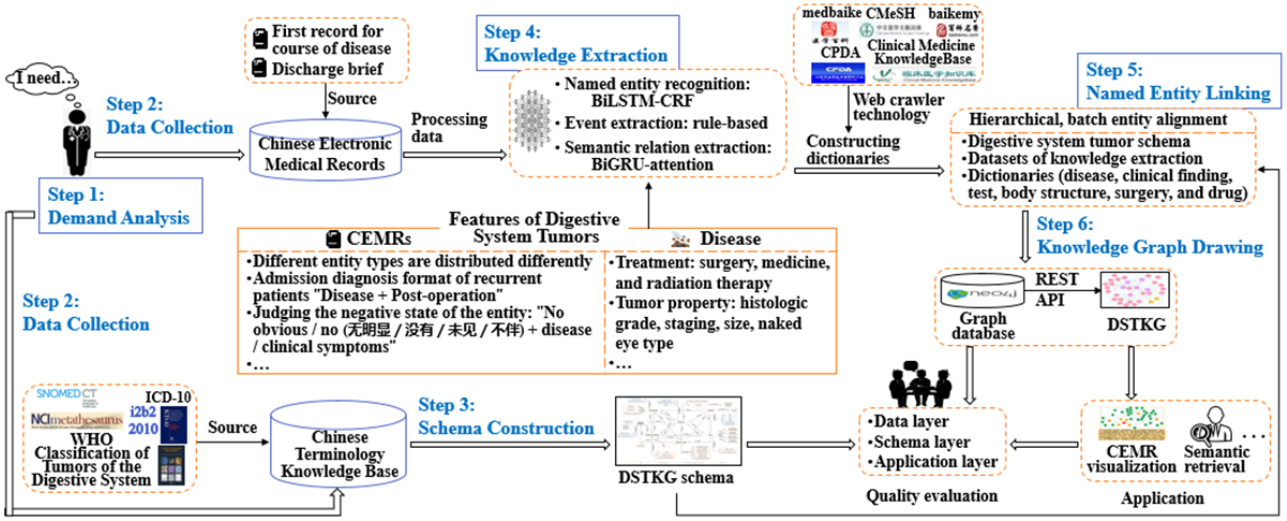
However, compared with the construction of Chinese knowledge graphs in other fields, tumor knowledge graphs are still rare, especially a DSTKG. This is because the purpose of constructing a Chinese medical knowledge graph is not only to serve the grass-roots general practitioners and specialists in large hospitals but also to popularize medical knowledge for patients. Therefore, the establishment of knowledge graphs of common diseases has more extensive value in China. On the other hand, compared with common diseases, the use of knowledge graphs to assist in diagnosis of intractable diseases such as tumors has a very high dimension, so it is more difficult to construct a tumor knowledge graph. Additionally, existing works for constructing Chinese medical knowledge graphs have mainly focused on "data-centered," namely extracting information and establishing straightforward connections [27,28]. However, they pay less attention to the schema of the knowledge graph and semantic relationships, which leads to poor conceptual standardization and scalability of the knowledge graph. With the continuous increase in the incidence and mortality of digestive system tumor diseases in China and the rapid growth of CEMR data, the construction of a DSTKG becomes urgent.

Methods

DSTKG Construction Framework

This study takes the first record for the course of disease and the discharged brief as the data source, focusing on the construction of the DSTKG schema and enriching the types of semantic relationships between concepts. Combining the characteristics of digestive system tumor diseases and the characteristics of tumor EMRs, we designed the DSTKG construction framework based on CEMRs, as shown in Figure 1, which includes 6 steps: demand analysis, data collection, schema construction, knowledge extraction, named entity linking (NEL), and knowledge graph drawing.

Figure 1. Construction framework of a digestive system tumor knowledge graph (DSTKG) based on Chinese electronic medical records (CEMRs). API: application programming interface; BiGRU: bidirectional gated recurrent unit; BiLSTM-CRF: bidirectional long short-term memory with a conditional random field; ICD: International Classification of Diseases; NCI: National Cancer Institute; REST: representational state transfer; SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms; WHO: World Health Organization.



Step 1 involves analyzing the construction purpose and application demands of the knowledge graph. The knowledge graph construction methods (eg, the selection of data sources and schema of the knowledge graph) would change according to the different construction purposes of knowledge graphs. First, we need to make it clear what the knowledge graph is built to do. The purpose of building a DSTKG in this study is to summarize the main contents of CEMRs as systematically and comprehensively as possible, to lay the foundation for the upper-level intelligent service.

In step 2, the data sources used to construct the knowledge graph are collected, according to the purpose of knowledge graph construction. The data sources consist of two parts: knowledge to construct the DSTKG schema and data to populate the DSTKG. CEMRs consist of 53 parts, such as medical record summary, admission record, and therapy record [29]. The patient's medical condition, diagnosis, and treatment are mainly recorded in the inpatient progress notes and discharge brief. The first record for the course of disease is the first diagnosis of the disease after admission, which is the quintessence of the

inpatient progress notes. Hence, the first record for the course of disease and the discharge brief are used as the data sources for the DSTKG in this study.

Step 3 involves constructing a patient-centered DSTKG schema to organize and manage the knowledge in CEMRs. Due to the rigor of medicine, the quality of knowledge used to construct a knowledge graph schema needs to be high. We can collect domain knowledge from various existing and high-quality ontologies or terminologies, such as SNOMED CT and National Cancer Institute (NCI) Metathesaurus.

In step 4, knowledge is extracted from the data sources. This step can be divided into 3 steps: data preprocessing, NER, and semantic RE. With considerations of the diagnosis characteristics of digestive system tumor diseases and the language structure characteristics of tumor CEMRs, we employed deep learning and rule-based methods to extract knowledge from CEMRs.

After knowledge extraction, this study exploited a hierarchical and batch entity alignment strategy in step 5 to realize NEL. In the process, we constructed 6 dictionaries (eg, disease dictionary, drug dictionary) to improve the effect of NEL.

In step 6, knowledge is stored in a Neo4j graph database for preliminary feedback. After that, this study uses DSTKG for semantic retrieval and CEMR visualization. Simultaneously, this research evaluates the quality of DSTKG in order to provide ideas for further improving the knowledge graph.

Data Sources

A knowledge graph can be divided into schema and data layers in a logical structure [30]. The schema of a knowledge graph stores refined knowledge. Accordingly, the knowledge used to construct a DSTKG schema learns from several open-access authoritative terminologies and ontologies, which are Chinese 3.4 version of SNOMED CT, the NCI Metathesaurus, World Health Organization (WHO) classification of digestive system tumors, and the second edition of the10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10). In addition, this study also refers to the US Informatics for Integrating Biology and the Bedside in 2010 (i2b2 2010) to determine the types of semantic relationships in DSTKG.

As for the dataset (731 CEMRs of digestive system tumors) used to build the DSTKG, it is derived from the Clinical Named Entity Recognition (CNER) challenge task of the China Conference on Knowledge Graph and Semantic Computing (CCKS) in 2018 [31]. In fact, a total of 1000 CEMRs were released by the CCKS 2018 CNER challenge task, of which 731 were CEMRs of digestive system tumors. The 731 CEMRs consist of 436 annotated CEMRs and 295 original medical records. It is worth noting that the annotated corpus only has 5 types of entities (“body structure,” “symptom,” “sign,” “surgery,” and “medicine”), which cannot meet the demands for constructing the DSTKG. Hence, under the guidance of a digestive oncology surgeon and following the principles of nonoverlapping and nonnesting, this study had 2 clinical postgraduates manually supplement 3 types of clinical entities (“disease,” “disease type,” and “test”) and 4 properties (“histological grade,” “pathological stage”, “naked eye type,” and “tumor size”) for 436 annotated CEMRs.

Schema Construction

Lying at the core of a knowledge graph, the schema is essentially a semantic network framework that can describe knowledge normatively and objectively. This paper refers to the method of constructing an ontology by Stanford University [32]; with the help of clinical experts and ontology experts, a top-down approach was used to construct a patient-centered DSTKG schema. The Material Management System of Chinese Clinical Medical terms was utilized for schema construction. The construction process can be divided into 4 parts: clarify the purpose of schema construction, determine the domain and scope, consider reusing existing resources, and assess quality (Figure 2).

To improve the standardization and scalability of the DSTKG, we decided to build a patient-centered DSTKG schema. Based on this purpose, we determined that the field and scope of the DSTKG schema are digestive system tumors, such as stomach cancer and colorectal cancer. Then, we reused existing resources (ICD-10, NCI Metathesaurus, WHO classification of digestive system tumors, SNOMED CT, i2b2 2010) and combined the characteristics of cancer diseases to define 7 classes and their data properties, as well as 16 semantic relationships. Finally, we invited 2 ontology experts to evaluate the quality of the DSTKG schema and revise it.

The 7 classes in the DSTKG schema are “patient,” “disease,” “disease type,” “test,” “body structure,” “clinical finding,” and “treatment.” In addition, “disease” is divided into 2 subconcepts: “noncancerous disease” and “cancer.” Its knowledge source is the Chinese 3.4 version of SNOMED CT and the second edition of ICD-10. The knowledge source of “test” is the same as “disease,” which has 4 subconcepts. The classes for “clinical finding” and “body structure” have the same knowledge source, namely the Chinese 3.4 version of SNOMED CT. “Clinical finding” is divided into the 2 subconcepts of “symptom” and “sign.” “Treatment” is divided into 3 subconcepts, namely “surgery,” “medicine,” and “radiotherapy,” which are inspired by NCI Metathesaurus. Most of the concepts in “disease type” are derived from the WHO classification of digestive system tumors. In addition, each concept has its own data properties such as ID, English name, and state. More details can be found in Table 1.

Figure 2. Process of constructing the digestive system tumor knowledge graph (DSTKG) schema. i2b2: Systematized Nomenclature of Medicine-Clinical Terms; ICD: International Classification of Diseases; NCI: National Cancer Institute; SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms; WHO: World Health Organization.

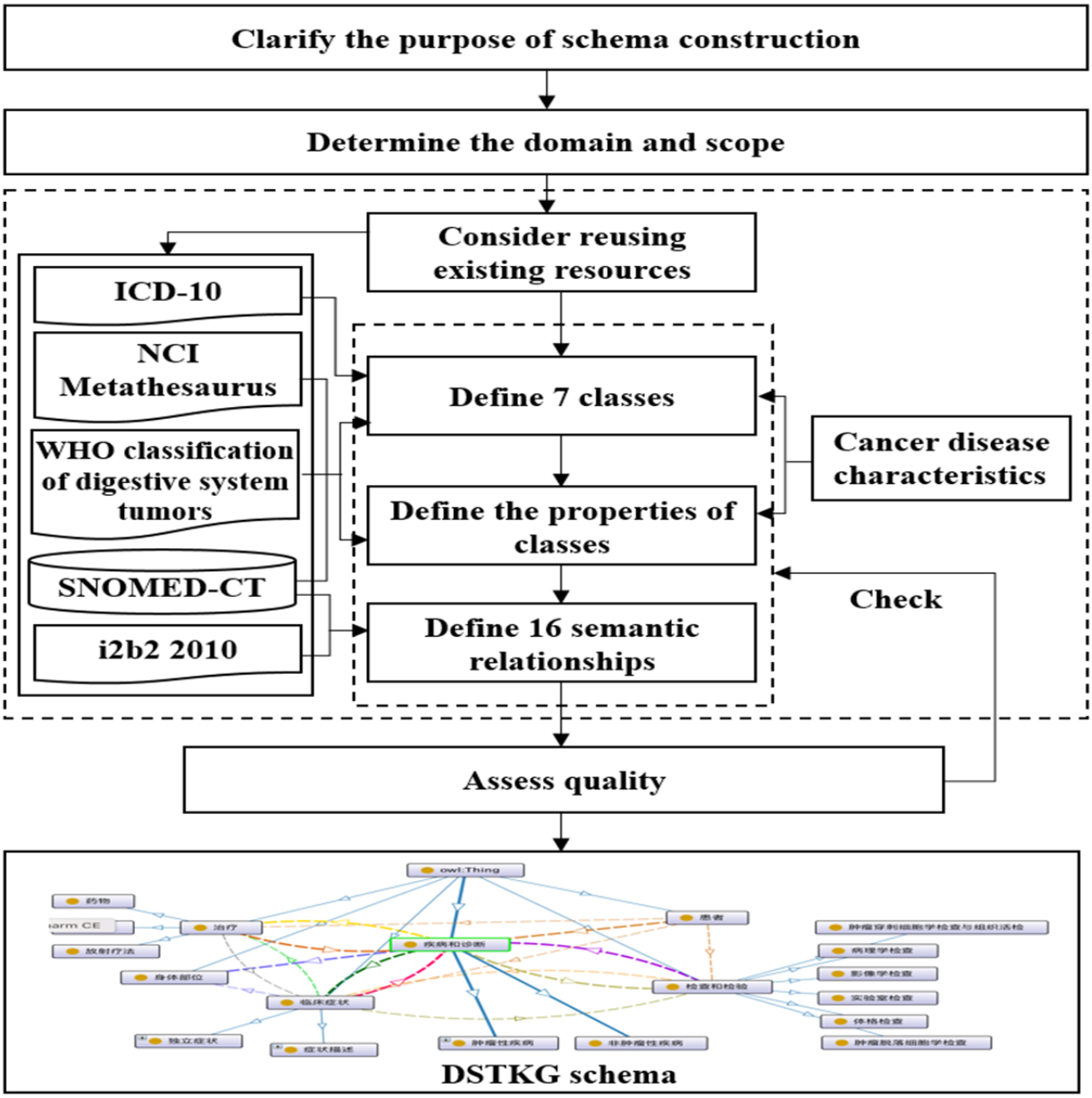


Table 1. Data properties of the digestive system tumor knowledge graph schema.

Class	Data properties
patient	ID, sex, age, occupation, native place
disease	English name, nonpreferred term, state
disease type	English name, nonpreferred term, histological grade, pathological stage, naked eye type, tumor size
clinical finding	English name, nonpreferred term, state
test	English name, nonpreferred term
treatment	English name, nonpreferred term
body structure	English name, nonpreferred term

The semantic relationship is the representation of the semantic correlation between domain concepts that connects the concepts. We defined 16 types of semantic relationships in the DSTKG. For example, the DSTKG connects the concepts “disease” and

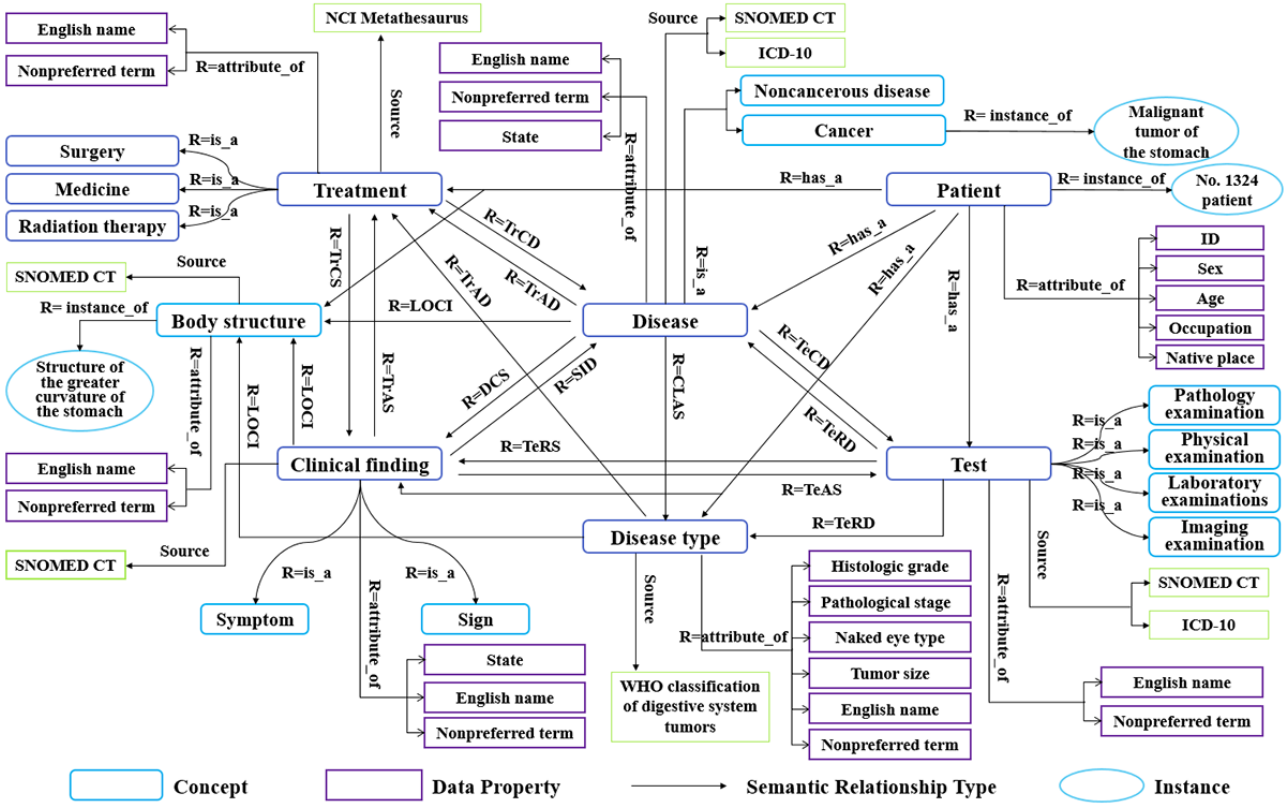
"test" with the semantic relationship "TeCD," which means "test is conducted to investigate the disease." Another semantic relationship between "disease" and "test" is "TeRD," which means "test reveals the disease." A collection of 13 types of semantic relationships is presented in Table 2. The 3 other semantic relationships are "attribute_of," "instance_of," and "is_a." Specifically, the DSTKG connects the concept with its data properties as "attribute_of" and establishes an "instance_of" relation between the concept and its entities. Further, the "is_a" relation is employed to connect the concept and its hyponym. For instance, the DSTKG connects the concepts "Primary malignant neoplasm of the stomach (胃原发性恶性肿瘤)" and "Primary malignant neoplasm of the body of the stomach (胃体原发性恶性肿瘤)" with an edge labeled "is_a" and establishes an "attribute_of" relation between the concept "Primary malignant neoplasm of the stomach (胃原发性恶性肿瘤)" and its data properties (pathological stage) "pT1N0M0."

The schema of the DSTKG that we constructed is shown in Figure 3. This DSTKG schema basically covers the domain concept system and provides a relatively complete framework for constructing the DSTKG.

Table 2. Semantic relationships of the digestive system tumor knowledge graph schema.

Coarse-grained category	Semantic relationship	Definition
Test-disease relation	TeCD	Test is conducted to investigate the disease.
Test-disease relation	TeRD	Test reveals the disease.
Test-clinical finding relation	TeRS	Test reveals the symptoms and signs.
Test-clinical finding relation	TeAS	Test is administered for the symptoms and signs.
Treatment-disease relation	TrAD	Treatment is administered for the disease.
Treatment-disease relation	TrCD	Treatment causes the disease.
Treatment-clinical finding relation	TrAS	Treatment is administered for the symptoms and signs.
Treatment-clinical finding relation	TrCS	Treatment causes the symptoms and signs.
Disease-clinical finding relation	DCS	Disease causes symptoms and signs.
Disease-clinical finding relation	SID	Symptoms and signs indicate the disease.
Disease-disease type relation	CLAS	Cancer disease type
Clinical finding–body structure relation	LOCI	Symptoms and signs are located in the body structure.
Patient-disease, clinical finding, test, and treatment relation	has_a	The patient has a certain disease, clinical finding, test, or treatment.

Figure 3. Schema of the digestive system tumor knowledge graph. CLAS: cancer disease type; DCS: disease causes symptoms and signs; ICD: International Classification of Diseases; LOCI: symptoms and signs are located in the body structure; NCI: National Cancer Institute; REST: representational state transfer; SID: symptoms and signs indicate the disease; SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms; TeCD: test is conducted to investigate the disease; TeRD: test reveals the disease; TeRS: test reveals the symptoms and signs; TrAD: treatment is administered for the disease; TrAS: treatment is administered for the symptoms and signs; TrCS: treatment causes the symptoms and signs; WHO: World Health Organization.



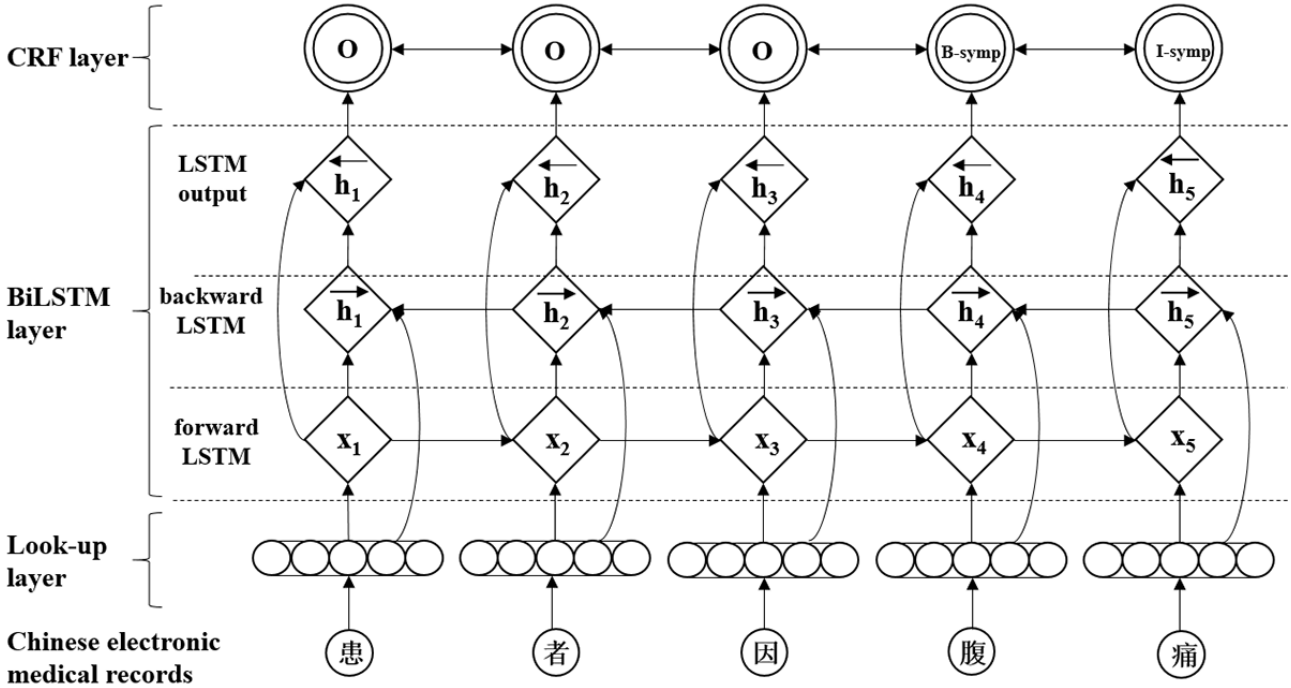
Knowledge Extraction

Knowledge extraction is the process of identifying valid, potentially useful, and ultimately understandable patterns from large data collections. It includes NER and RE [33].

Named Entity Recognition of CEMRs

Chinese NER is usually modeled as a character-level sequence-labeling problem, because a Chinese sentence is a string of characters without an explicit delimiter [34]. For most sequence-labeling tasks, it is beneficial to access both past (left) and future (right) contexts [35]. Bidirectional long short-term memory (BiLSTM) networks with a conditional random field (CRF) can do this. A BiLSTM-CRF model takes advantage of BiLSTM that can simultaneously utilize the past and future input features through the forward and backward LSTM layer. It can use the sentence-level tag information via the CRF layer as well. In terms of identifying various types of clinical entities (such as diagnoses, tests, symptoms, body part, and treatment) from CEMRs, BiLSTM-CRF achieved better performance than the baseline CRF model. Generally speaking, the F1 score of the BiLSTM-CRF model is more than 1% higher than that of the CRF model [34,36]. Therefore, this research employed a char-based BiLSTM-CRF model for NER from CEMRs. The architecture of the BiLSTM-CRF is shown in Figure 4.

Figure 4. Architecture of the bidirectional long short-term memory (BiLTSM) conditional random fields (CRF). symp: symptom.



The BiLSTM-CRF model consists of the look-up, BiLSTM, and CRF layers. In the look-up layer, each character in the sentence is mapped from the one-hot vector to the low-dimensional dense character vector (character embedding) as the initial input feature vector of the model. Then, the BiLSTM layer is used to extract sentence features automatically. Specifically, the forward LSTM and backward LSTM layers take the sequence of character representations $X=(x_1, x_2,\ldots, x_n)$ as input and generate the representation of the left (Equation 1) and right (Equation 2) context for each character.

After that, the LSTM's output layer represents the overall context sequence as  , where h_t is the concatenation of  and 

Finally, the sequence of overall context representations was taken as input for the CRF layer to predict the output label sequence.

To train the BiLSTM-CRF model, we selected 4 types of features to optimize the recognition effect: bag of characters, part-of-speech (POS), position of the character in the sentence, and dictionary features. The Natural Language Processing and Information Retrieval Chinese lexical analysis system (NLPIR-ICTCLAS) was utilized for word segmentation, while POS tags were generated simultaneously [37]. Because we used character-level information, the POS tag of the Chinese character is just the POS tag of the corresponding word that contains that character. In addition, character embeddings were learned through Gensim's word2vec on the 1000 original clinical medical records, while the segmentation information was generated by the Jieba segmentation system.

In this study, the corpus was divided into training, test, and verification sets at a ratio of 8:1:1. For the deep learning model, we set the character embedding dimension to 100, batch size of the model to 50, and learning rate to 0.0004. To alleviate the possible overfitting problem of the model, the dropout was fixed at 0.5. In the CRF model, the content window size was set to 5 to extract character features. As a result, our BiLSTM-CRF model achieved excellent overall performance with an F1 score of 0.9678, precision score of 0.9720, and recall score of 0.9636. Then, we applied the trained BiLSTM-CRF model to tag the unlabeled corpus and manually proofread the results.

For the "status" property of "clinical finding" or "disease," we explored a rule-based approach to tag it. This is because tumor CEMRs have some fixed grammatical usage and syntactic structure. Specifically, a "clinical finding" or "disease" that indicates a negative state usually comes after "无明显 / 没有 / 未见 / 不伴 (not obvious/not seen/no)." For example, "患者2月余前出现上腹闷痛不适, 为饥饿时明显, 无阵发性加剧, 不伴恶心, 呕吐, 返酸, 嗳气等 (The patient started to have epigastric pain and discomfort more than 2 months ago, which was more severe when he was hungry. He has no paroxysmal aggravation, no nausea, no vomiting, no acid regurgitation, no belching, and so on.)" When the entities of "clinical finding" and "disease" with a negative property are determined, the "state" property of the remaining entities of "clinical finding" and "disease" is positive.

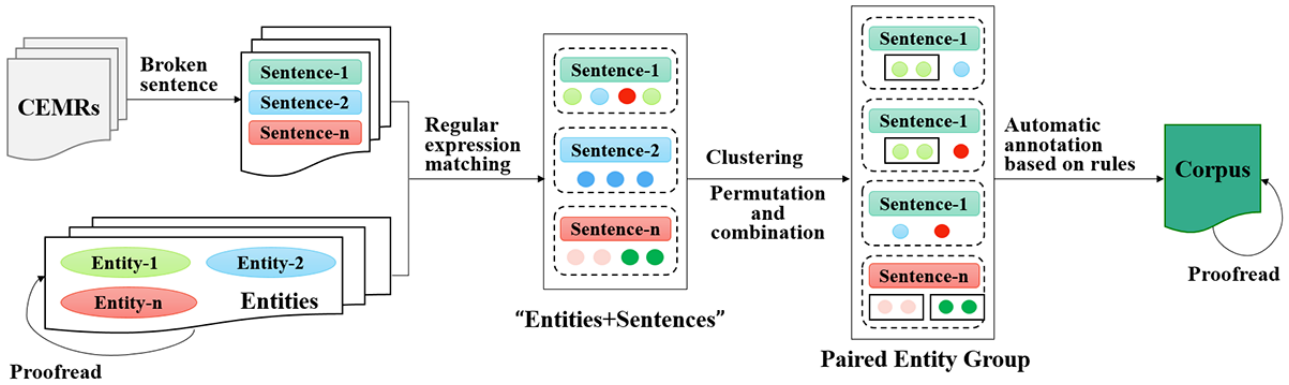
Relationship Extraction of CEMRs

RE can be simply understood as a classification problem: Given 2 entities and the sentence that appear together, distinguish the relationship between the 2 entities. Liu et al [38] found that the attention mechanism used on top of the gated regression unit (GRU) model outperforms the existing state-of-the-art neural network models on the THYME corpus in intrasentence RE from clinical narratives. Furthermore, the GRU model was also

very effective on the Chinese corpus ACE2005 dataset for the entity extraction task. It only embeds Chinese character vectors, Chinese word vectors, and the regional list and can obtain high-level global features (F1=85.3) without any additional features [39]. The basic idea of the GRU model is similar to that of the LSTM model and can be regarded as a variant of LSTM. In some cases, the GRU model can produce the same excellent performance. Additionally, the GRU model can save the information in the long-term sequence and will not clear it or remove it over time because it is not related to prediction. Therefore, it can effectively solve the problem of gradient disappearance of the standard recurrent neural network [40]. Character embedding naturally adapts to Chinese characteristics. Zhou et al [41] proposed that attention-based BiLSTM memory networks outperform most of the existing methods, with only word vectors, on the SemEval-2010 relation classification task. Sentence-level attention dynamically reduces the weights of the wrong labelling problem [42]. Therefore, this paper utilizes a bidirectional GRU (BiGRU) neural network and dual-attention mechanism at the word and sentence levels to extract the relationship.

First, we performed character embedding on each Chinese character in the sentence and then fed the results to the attention GRU model at the word level. Finally, the characteristics of each sentence output are re-entered into the BiGRU, and sentence-level attention is added to solve the problem that global information cannot be expressed [43]. However, due to a lack of an annotated RE corpus at the beginning, we needed to label the corpus first, as shown in Figure 5. The materials are 731 Chinese digestive system tumor EMRs and their entity sets containing position and type information. As can be seen from Table 2, when the type of 2 entities is determined, the type of semantic relationship between them can roughly be determined too. For instance, the semantic relationship between the "disease" and "body structure" classes can only be "LOCI," and the semantic relationship between the "disease" and "test" classes may be "TeCD" or "TeRD." Hence, this paper first breaks the sentences in CEMRs and then uses regular expressions to match the proofread entities and sentences. In tumor CEMRs, the same types of medical entities often appear together and usually have the same semantic relationship with other types of medical entities. Based on this characteristic of tumor CEMRs, we cluster, permute, and combine the clinical entities in a sentence to make sure there are at most 2 types of entities.

Figure 5. Construction process of the relational extraction corpus. CEMR: Chinese electronic medical record.



In the end, according to the aforementioned rules, the RE corpus is labeled automatically and proofread manually. In this research, we annotated 45,607 sentences, of which 80% were used as training data, and the rest were used as a testing set. Moreover, BiGRU-attention achieved good performance with an F1 score of 0.5167.

Named Entity Linking

There is usually a synonymy problem in the process of extracting knowledge from a single data source; that is, an entity has many different entity mentions, such as names, aliases, abbreviations, and even misrepresentations. NEL is an important method to solve the problem of entity ambiguity by mapping entity mentions to standard concepts in the knowledge base [44].

Before carrying out NEL, we constructed 6 dictionaries based on sources such as the State Food and Drug Administration and clinical medical knowledge base. Then, we proposed a hierarchical and batch NEL method based on the DSTKG schema, as illustrated in Figure 6. The process can be divided into 3 steps. The first step is to remove the duplicate entities of each EMR text and match them hierarchically between entities and schema. Specifically, the semantic type of the entity first matches the class; then, the entity exactly matches concepts of this class. If there is no match, an exact match will be made later on the data properties of these concepts. The second step is to use the constructed dictionaries to expand those unmatched entity mentions and repeat the step of hierarchical matching. Last, we used the rank-based approach of cosine similarity to sort those unmatched entities and select the top-ranking entities as the NEL result after disambiguation.

Eventually, the research obtained 9868 entities: 1002 "disease," 452 "disease type," 1874 "body structure," 1606 "treatment," 2786 "clinical finding," 924 "test," and 1224 "properties of disease type." We also obtained 11,005 semantic relationships (Table 3).

Figure 6. Process of hierarchical named entity linking. DSTKG: digestive system tumor knowledge base.

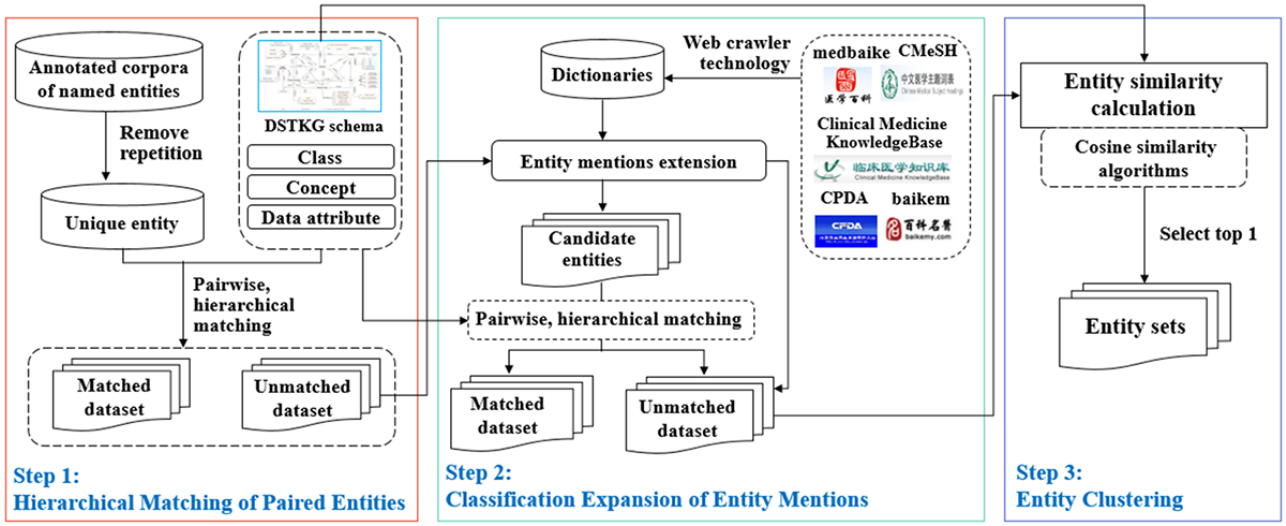


Table 3. Number of each semantic relationship (N=11,005).

Semantic relationship	Definition	Number
TeCD	Test is conducted to investigate the disease.	96
TeRD	Test reveals the disease.	990
TeRS	Test reveals the symptoms and signs.	2035
TeAS	Test is administered for the symptoms and signs.	486
TrAD	Treatment is administered for the disease.	980
TrCD	Treatment causes the disease.	2
TrAS	Treatment is administered for the symptoms and signs.	613
TrCS	Treatment causes the symptoms and signs.	326
DCS	Disease causes the symptoms and signs.	29
SID	Symptoms and signs indicate the disease.	261
LOCI	Symptoms and signs are located in the body structure.	3330
attribute_of	Properties of the entity	1857

Knowledge Graph Drawing

In this study, we chose the Neo4j graph database to draw the DSTKG. A graph database is different from a traditional relational database in that it can store the ontologically structured knowledge and visualize the relationship between entities [45]. As one of the most popular graph databases, Neo4j is an open-source database implemented in Java, which organizes data as nodes, relationships, and properties in the property graph model. Additionally, Neo4j supports an ACID-compliant transactional backend and applies Cypher for retrieving data. The syntax of Cypher is relatively simple, and its performance is not affected by the amount of data. Therefore, we used Neo4j to manage the data and draw the knowledge graph.

Considering the DSTKG's readability, the knowledge graph first displays the top 3-tier structure of the DSTKG by default. The users can take advantage of Neo4j's node expansion to browse the DSTKG. Furthermore, the nodes at different levels are designed to have different sizes and colors for easy distinction. For example, the color of the "patient" class is red, and its nodes are larger than those of other classes, while the nodes of the "entity" are green and the smallest. Similarly, different types of semantic relationships in the DSTKG have different colors (ie, "instance_of" is orange, "TeRS" is pink, and "TrAD" is green). Finally, the DSTKG constructed in this study contains 11,372 entities and 19,276 semantic relationships [46].

Results

After completion of the DSTKG, we used it to browse clinical knowledge and visualize CEMRs. A preliminary evaluation of the DSTKG was also carried out.

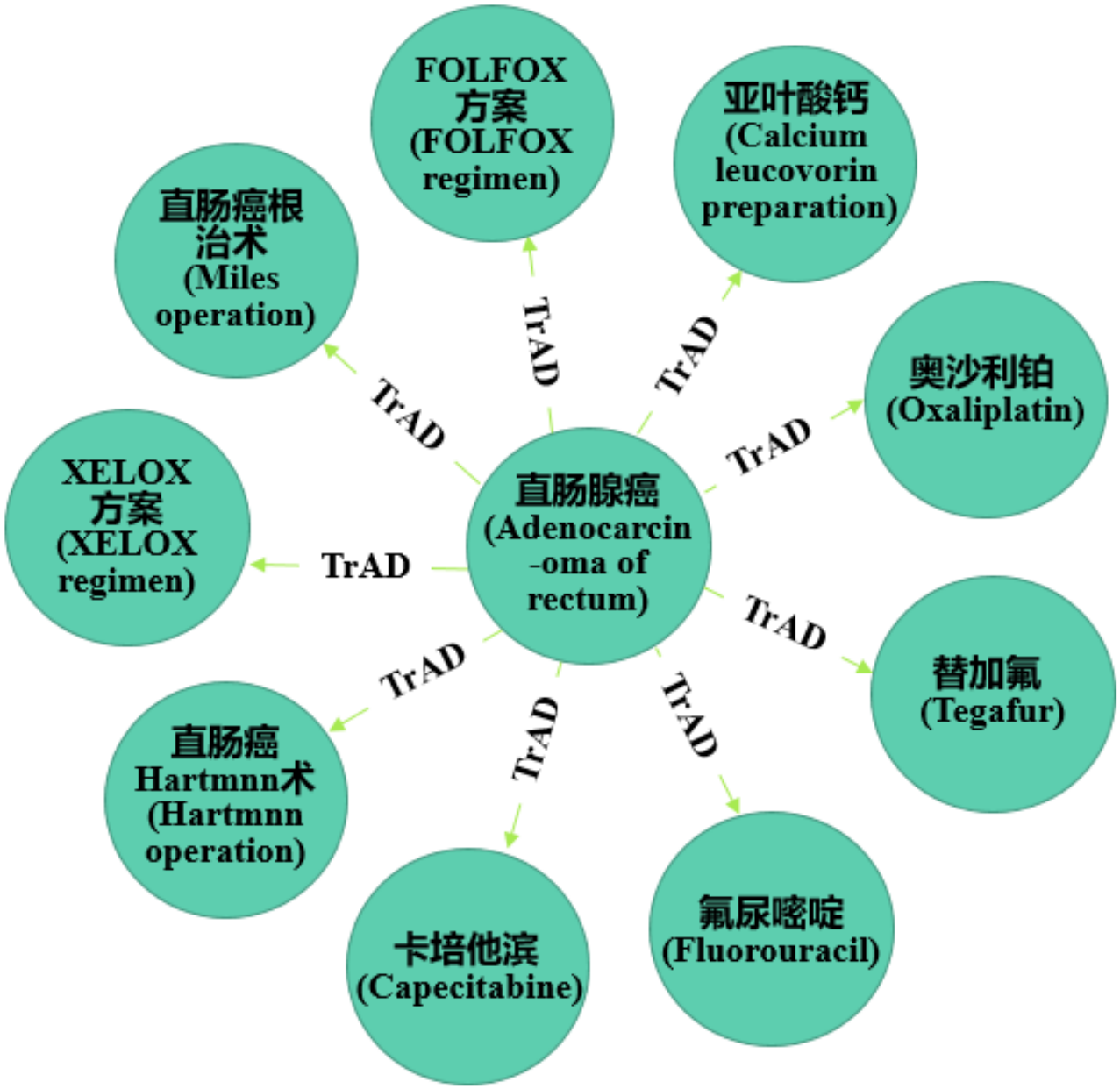
Semantic Retrieval and CEMR Visualization

The knowledge graph implements a transition from a "text-centered" retrieval model to a "things-centered" retrieval model. The DSTKG integrated 731 patients' CEMRs so we can retrieve global or individual patient data. For instance, if you want to study the treatment of rectal adenocarcinoma (直肠腺癌), you could use the following MATCH statement: MATCH

(n:entities)-[r:TrAD]->(c) WHERE n.name='直肠腺癌' RETURN n,c LIMIT 2500. The retrieval results are shown in Figure 7.

As is shown in the retrieval results of the 731 CEMRs (Figure 7), rectal adenocarcinoma had 9 treatments. Moreover, if you want to find the most frequently used treatment, you only need to perform a simple statistical analysis of the query results: 50% (365/731) of patients with rectal adenocarcinoma underwent a Miles operation, 15% (110/731) were administered oxaliplatin, 11% (80/731) of patients received the XELOX regimen, 8% (58/731) were administered capecitabine, 6% (44/731) were administered tegafur, 4% (29/731) were administered the calcium leucovorin preparation, and only 2% (15/731) each were treated with fluorouracil, the FOLFOX regimen, or a Hartmann operation.

Figure 7. Treatment of rectal adenocarcinoma based on a search of Chinese electronic medical records using the search "MATCH (n:entities)-[r:TrAD]->(c) WHERE n.name='直肠腺癌' RETURN n,c LIMIT 2500" in the digestive system tumor knowledge graph. TrAD: treatment is administered for the disease.

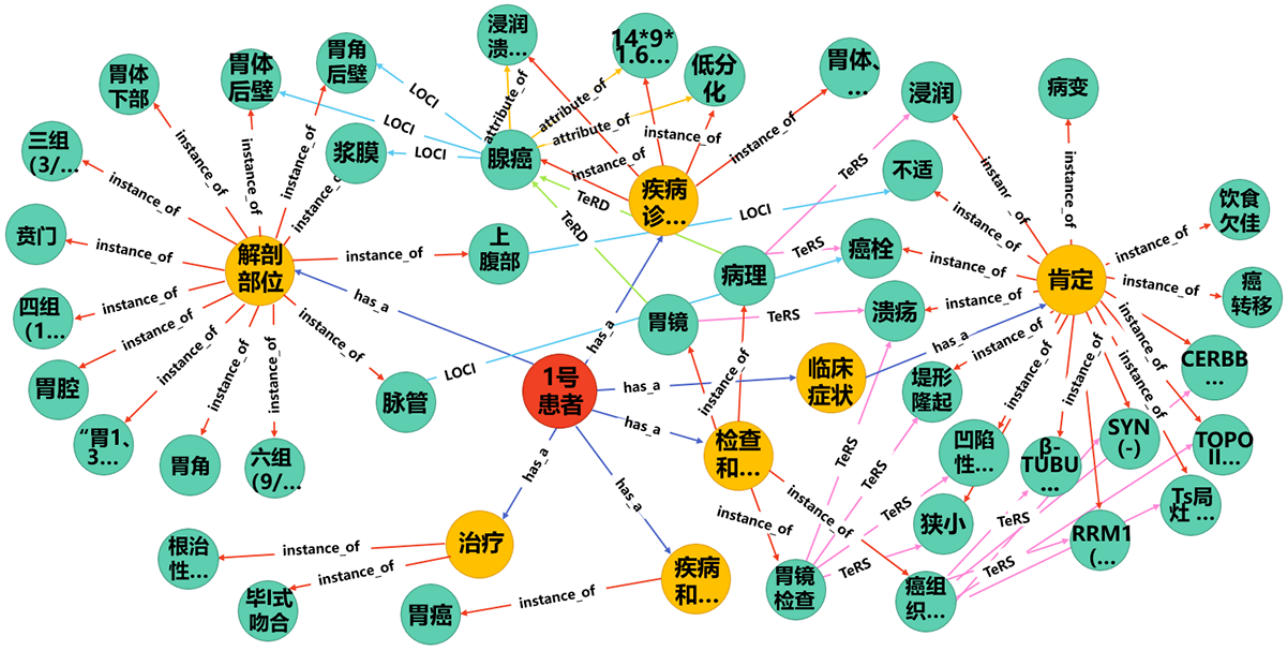


The Miles operation, as a radical operation for lower rectal cancer and anal carcinoma, is widely used in the clinic. However, the Hartmann operation is generally performed in patients because of poor general condition, an inability to tolerate the Miles operation, or the Dixon operation is not suitable due to acute obstruction. The statistical results of this study are consistent with the actual clinical situation. Moreover, the FOLFOX regimen is also a common chemotherapy regimen for gastrointestinal tumors. The drugs included in this regimen are oxaliplatin, calcium leucovorin preparation, and fluorouracil. The reason for the small proportion of FOLFOX regimen may be due to the different writing habits of doctors. Some doctors write the FOLFOX regimen as a specific dosing program, such as oxaliplatin 85 mg/m^2 iv gtt (2 h) D1, calcium leucovorin preparation 400 mg/m^2 iv gtt (2 h) D1, and fluorouracil 2400 to 3000 mg/m^2 iv gtt (continuous 46 h) D1. This will lead to an

increase in the proportion of drugs used to treat rectal adenocarcinoma (such as oxaliplatin and calcium leucovorin preparation) in the statistical results, while the proportion of the FOLFOX regimen decreases. This conjecture proved to be correct by consulting the original CEMRs. Regardless of these findings, these statistical results can provide a certain degree of data to support clinicians in developing treatment plans.

The process of disease diagnosis is generally that the doctor first understands the patient's medical history and then issues the necessary examination documents. After the patient completes the examination, the doctor interprets the examination results and determines the treatment plan or further examination decisions. In this process, the doctor's understanding, interpretation, and judgment of the patient's condition are at the core of clinical work. Based on the DSTKG, we can not only view the treatment plans, clinical examinations, or common clinical findings of digestive system diseases but also query the visualization result of an individual patient's CEMRs from the DSTKG. For example, we queried the knowledge graph for "Patient no. 1" using the following statement: MATCH (n:person)-[r:has_a]->(b)-[re]->(c) WHERE n.name='1号患者' RETURN n,b,c LIMIT 2500. The visualization result is shown in Figure 8. Textbox 1 presents the original CEMR text for this patient. By comparing Figure 8 and Textbox 1, we can see that the patient's knowledge graph basically contains the main clinical entities in the CEMR text. The knowledge graph technology integrated the medical information of patient no. 1 scattered on the various working platforms of the EMR system, such as basic patient information, treatment information, and examination information and displayed them in a visual manner. With the help of the knowledge graph of patient no. 1, clinicians could grasp his main medical information at a conceptual level without having to find his information scattered on various work platforms. This not only is more intuitive but also saves time for clinicians when looking up patient medical information and effectively improves clinicians' work efficiency. On the other hand, the integration of patient medical information can prevent clinicians from misjudgments due to omission of important patient information. Moreover, for patients, most have no medical background, and it is difficult for them to correctly find their own diagnosis and treatment information scattered in different reports. The DSTKG can provide patients with their own EMR knowledge map so they can fully understand their diagnosis and treatment process.

Figure 8. Knowledge graph of "patient no. 1." CLAS: cancer disease type; DCS: disease causes symptoms and signs; LOCI: symptoms and signs are located in the body structure; SID: symptoms and signs indicate the disease; TeCD: test is conducted to investigate the disease; TeRD: test reveals the disease; TeRS: test reveals the symptoms and signs; TrAD: treatment is administered for the disease; TrAS: treatment is administered for the symptoms and signs; TrCS: treatment causes the symptoms and signs.



Textbox 1. The original Chinese electronic medical record text of "patient No. 1."

患者2个月前因上腹部不适于我院就诊, 入院后行胃镜检查示: 于胃体下部, 胃角后壁处可见较深凹陷性病变, 占胃腔1/2周, 中央可见巨大溃疡, 表面覆污苔, 周围粘膜不规则, 有明显浸润并周边呈堤形隆起; 病变部位胃腔狭小, 内镜尚能通过; 贲门部未受侵犯. 胃镜病理 (201600649), 示: (胃角, 胃体后壁) 低分化腺癌. 于2016-01-14在全麻下行根治性胃癌根治术+毕I式吻合, 术后病理示: (201600925) 胃体, 胃窦低分化腺癌, 浸润溃疡型, 体积14*9*1.6CM, 侵穿浆膜, 并于部分脉管内查见癌栓. 累及近端切线, 远端切线及另送"远端切线"未查见癌. 呈三组 (3/3个), 四组 (11/17个), 六组 (9/11个), "胃1,3,7组" (2/5个) 淋巴结癌转移. "胃第六组" (1个) 淋巴结未查见癌., 癌组织免疫组化染色示: RRM1 (-), TS局灶 (+), TOPOII部分 (+), β-TUBULIN-III (-), SYN (-), CERBB-2 (-). 术后给予静脉营养, 抑酸, 补液, 补充白蛋白, 抗感染等对症治疗. 患者恢复良好出院, 现患者为进一步行化疗来我院, 门诊以"胃癌术后"收入我科. 患者自发病以来, 精神可, 饮食欠佳, 睡眠可, 二便正常, 体重体力无明显改变.

DSTKG Preliminary Evaluation

To appraise the reliability and practicability of the DSTKG, we adopted the expert evaluation method to preliminarily evaluate the knowledge graph from 3 aspects: data layer, schema layer, and application layer.

First, we designed a 5-level Likert scale with 9 different quality dimensions, as shown in Table 4. Then, 5 experts were invited to evaluate the quality of the DSTKG: 2 ontology experts, 1 expert in computer science, 1 urology clinical expert, and 1 hepatobiliary surgery clinical expert. At the time of expert assessment, we had not developed an interactive visualized medical knowledge service system based on the DSTKG. Hence, when the experts assessed the quality of DSTKG, we assisted them. The quality evaluation process was divided into 3 steps: (1) introducing the construction process and basic functions of the DSTKG to the experts; (2) assisting the experts to employ the DSTKG stored in Neo4j, such as retrieving specific concepts and semantic relationships or browsing the knowledge graph of specific patients; and (3) rating the DSTKG anonymously using the 5-level Likert scale. Table 4 shows the quality evaluation results.

Table 4. Quality evaluation score and intraclass correlation (ICC) score of the digestive system tumor knowledge graph.

Dimension and metrics	Quality evaluation score (rating)	ICC (95% CI)
Data layer		
Authority	4.73 (excellent)	0.97 (0.73 to 1.00)
Amount of data	2.68 (adequate)	
Schema layer		
Rationality	4.72 (excellent)	0.23 (–0.19 to 1.00)
Scalability	4.67 (excellent)	
Application layer		
Data consistency	4.46 (very good)	0.76 (0.41 to 0.97)
Ease of use	3.69 (adequate)	
Readability of results	4.69 (excellent)	
Accuracy	4.57 (very good)	
Practicability	3.56 (adequate)	

The mean quality evaluation score of the 5 experts was 4.20, indicating that the experts thought the DSTKG was generally good. In addition, the scores of the 2 metrics in the schema layer were higher than 4.60, which was much higher than the average. This indicated that the DSTKG schema had good performance, had reasonable structure, and was easy to expand. At the same time, the scores of "readability of results" and "accuracy" were >4.50. This could be attributed to the good performance of knowledge extraction and the NEL model, as well as the visual design of the DSTKG. However, the metric "amount of data" was scored the lowest, at 2.89, and the scores of "ease of use" and "practicability" were both <3.70. The reason was that the experts thought that the number of CEMRs was insufficient, which had a certain impact on the usability of the knowledge graph and objectivity of evaluation. Furthermore, compared with other experts, the clinical experts had higher requirements for the accuracy of the DSTKG and deemed that there was still a long way to go before the DSTKG was ready for clinical application, so they gave a very low score. About the "ease of use," although the Cypher query language is easy to learn, it is still difficult for users who do not understand programming.

This was a preliminary evaluation. However, it provided direction for improving the DSTKG and indicated the need to better understand and contemplate the application of the DSTKG beyond a solely technological perspective.

Consistency

We performed an interrater reliability analysis using the intraclass correlation coefficient (ICC) and 95% CI. The ICC is commonly used to assess interrater reliability for ordinal, interval, and ratio variables and is suitable when 2 or more coders are used [47]. Reliability is assessed as "perfect" at an ICC value of 1.0, "excellent" at >0.81, "substantial" at 0.80-0.61, "moderate" at 0.60-0.41, "fair" at 0.40-0.21, and "slight" at <0.20 [48,49]. This analysis was conducted using SPSS version 23.0. The results are shown in Table 4.

The ICCs for the DSTKG metrics ranged from 0.23 to 0.97 (Table 4). The ICC for the 5 experts was lowest in the schema layer (0.23), but higher in the other two dimensions (application layer, 0.76; data layer, 0.97), the interrater reliability was "substantial" to "excellent." In general, the intrareviewer item score agreement was acceptable.

Discussion

Principal Findings

This paper proposes a framework for building a DSTKG based on CEMRs and describes the construction of the DSTKG according to the framework. Finally, experts evaluated its quality from different dimensions. Although this DSTKG needs further refinement, it is an attempt to provide the basis for complete and consistent reporting of this rather vague area. The responses

from the experts showed the DSTKG construction framework was scientific and feasible, and the DSTKG had high rationality and some practicality. However, the ICC for the 5 experts was lower in the schema layer (ICC=0.23). On the one hand, because the schema construction is highly specialized, it is difficult for nonprofessionals to understand, so its evaluation varied greatly. On the other hand, the schema of the DSTKG is built by semi-automation. Due to different professional fields, experts have different views on this. For instance, computer science experts believed that semi-automatic construction would affect the scalability of the schema, so they gave a lower score for the scalability of the schema. Our DSTKG also displayed more granular semantic relationships and scalability than previous tumor knowledge graphs. Furthermore, clinicians can grasp patients' medical information more quickly and conveniently than by reading CEMRs, and patients can also better understand and manage their own diseases on the basis of the DSTKG. The DSTKG is expected to contribute to a better representation of CEMRs and form the basis for further semantic research of tumors.

Highly Granular Knowledge Extraction

As the content of CEMRs becomes more complex, highly granular knowledge extraction of CEMR text is becoming increasingly important. The property of the concept is known as "the key feature of the concept," which can describe the concept in a holistic manner. The property description of medical concepts not only helps further narrow the scope of possible diseases but also helps to distinguish patients with similar symptoms and provide personalized treatment. Taking this into account, this research not only defined 7 common types of concepts in the process of clinical diagnosis and treatment but also used a deep learning model or rule-based method to extract conceptual properties, such as the "histological grade" of the disease type and the "state" of disease or clinical finding. Although it is a preliminary attempt, it can provide a reference for subsequent assertion classification and entity property recognition, to further improve the accuracy and practicability of the DSTKG.

Additionally, unlike the breast tumor knowledge graph, which simply defined the semantic relationships between patients and medical concepts, the DSTKG also built rich semantic relationships between medical entities. The DSTKG contained a total of 16 types of semantic relationships. It also went a step further to refine the semantic relationships between concepts. For instance, this study defined 2 types of semantic relationships between the concepts of treatment and disease: "TrAD" for "treatment is administered for the disease" and "TrCD" for "treatment causes the disease." Highly granular semantic relationships can increase the relevance between concepts in the text. This can not only enhance the semantic interpretation of diagnostic results but also facilitate in-depth data analysis and knowledge reasoning.

Easy-to-Extend DSTKG

Many aspects of the DSTKG can be easily adjusted to a designer's focus because of the schema we built. This has several advantages over existing Chinese tumor knowledge graphs. First, the DSTKG schema was based on the international standardized thesauruses, and 6 types of medical dictionaries were used, which made the instantiation of the schema easier. In addition, the concepts in the DSTKG are relatively independent. Therefore, the DSTKG allows designers to easily add or delete a class of concepts and semantic relationships to accommodate their emphasis on a subject matter. For example, if the designer only wants to construct a knowledge graph about the clinical symptoms and examinations of patients with digestive system tumors, he can directly extract the concepts of "patient," "clinical finding," and "examination" and their semantic relationships to construct a schema. Of course, the designer can also easily add other types of concepts such as patient basic information. There is no inherent limit to the number of concepts and semantic relationships that can be included in our schema.

Limitations

A limitation of this study is the lack of extensive evaluation of the DSTKG. Moreover, the lack of additional data sources may pose a challenge to promoting the use of the DSTKG. In addition to adding more data sources, strengthening the research on assertion classification and further enriching the properties of entities are considered to be a necessary next step. Therefore, we plan to develop an interactive knowledge service platform based on the DSTKG in the future, so that the DSTKG can be more widely used and evaluated.

Conclusions

Although CEMRs contain a wealth of medical knowledge, their utilization rate is very low. Knowledge graphs are an emerging knowledge organization technology that provides a novel approach for the deep mining and utilization of CEMRs. In view of this, we proposed a framework for the construction of a DSTKG based on CEMRs and realized the construction of the DSTKG. This research not only contributes to knowledge organization in the field of digestive system tumors but also paves the way for knowledge extraction based on the characteristics of digestive system tumor diseases and tumor CEMRs. More importantly, this research promotes the development of oncology research towards semantics. In addition, the DSTKG can also evolve toward the creation of an interactive knowledge service platform for further evaluation and investigation.

Acknowledgments

The authors would like to thank the CCKS 2018 CNER challenge organizers for providing the data source. The research is supported by the Construction of Online Service System and Standard Penetration of National Population Health Science Data Center (grant #NCMI-KDOIN-201905) and the Precise Medical Database Group of Major Diseases of the Key Special Project of Precision Medical Research in China's National Key R&D Programmes (grant #2016YFC0901602).

Conflicts of Interest

None declared.

Multimedia Appendix 1
The original Chinese electronic medical record text of “patient No. 1”.
[DOCX File , 15 KB - medinform_v8i10e18287_app1.docx]

Multimedia Appendix 2
English translation of the original Chinese electronic medical record text of “patient No. 1”.
[DOCX File , 16 KB - medinform_v8i10e18287_app2.docx]

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 12;68(6):394-424 [FREE Full text] [doi: 10.3322/caac.21492] [Medline: 30207593]
2. Feng R, Zong Y, Cao S, Xu R. Current cancer situation in China: good or bad news from the 2018 Global Cancer Statistics? Cancer Commun (Lond) 2019 Apr 29;39(1):22 [FREE Full text] [doi: 10.1186/s40880-019-0368-6] [Medline: 31030667]
3. What are the differences between electronic medical records, electronic health records, and personal health records? The Office of the National Coordinator for Health Information Technology (ONC). 2019 May 02. URL: https://www.healthit.gov/faq/what-are-differences-between-electronic-medical-records-electronic-health-records-and-personal [accessed 2019-07-21]
4. Manca DP. Rebuttal: Do electronic medical records improve quality of care? Yes. Can Fam Physician 2015 Oct;61(10):e435, e437. [Medline: 26472802]
5. Payne TH, Corley S, Cullen TA, Gandhi TK, Harrington L, Kuperman GJ, et al. Report of the AMIA EHR-2020 Task Force on the status and future direction of EHRs. J Am Med Inform Assoc 2015 Sep;22(5):1102-1110 [FREE Full text] [doi: 10.1093/jamia/ocv066] [Medline: 26024883]
6. 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]
7. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 May 02;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
8. Chun S, Jin X, Seo S, Lee KH, Shin Y, Lee I. Knowledge graph modeling for semantic integration of energy services. 2018 Presented at: IEEE International Conference on Big Data and Smart Computing; January 15-17, 2018; Shanghai, China. [doi: 10.1109/bigcomp.2018.00138]
9. Sang S, Yang Z, Wang L, Liu X, Lin H, Wang J. SemaTyP: a knowledge graph based literature mining method for drug discovery. BMC Bioinformatics 2018 May 30;19(1):193 [FREE Full text] [doi: 10.1186/s12859-018-2167-5] [Medline: 29843590]
10. Ferrucci D, Levas A, Bagchi S, Gondek D, Mueller ET. Watson: Beyond Jeopardy!. Artificial Intelligence 2013 Jun;199-200:93-105. [doi: 10.1016/j.artint.2012.06.009]
11. Pujara J, Miao H, Getoor L, Cohen WW. Ontology-aware partitioning for knowledge graph identification. 2013 Presented at: Workshop on Automated Knowledge Base Construction; October 27-28, 2013; San Francisco, CA. [doi: 10.1145/2509558.2509562]
12. Knowledge Graph of Breast Cancer. 2012. URL: http://wasp.cs.vu.nl/BreastCancerKG/ [accessed 2019-07-22]
13. 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 [FREE Full text] [doi: 10.1038/s41598-017-05778-z] [Medline: 28729710]
14. Kwon BC, Choi M, Kim JT, Choi E, Kim YB, Kwon S, et al. RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. IEEE Trans. Visual. Comput. Graphics 2019 Jan;25(1):299-309. [doi: 10.1109/tvcg.2018.2865027]
15. Bean DM, Wu H, Iqbal E, Dzahini O, Ibrahim ZM, Broadbent M, et al. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. Sci Rep 2017 Nov 27;7(1):16416 [FREE Full text] [doi: 10.1038/s41598-017-16674-x] [Medline: 29180758]
16. Zhang C, Li Y, Du N, Fan W, Yu PS. On the generative discovery of structured medical knowledge. 2018 Dec Presented at: 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; August 2018; London, United Kingdom p. 19-23. [doi: 10.1145/3219819.3220010]
17. Zhao Q, Wang D, Li J, Akhtar F. Exploiting the concept level feature for enhanced name entity recognition in Chinese EMRs. J Supercomput 2019 Jun 10;76(8):6399-6420. [doi: 10.1007/s11227-019-02917-3]
18. Chowdhury S, Dong X, Qian L, Li X, Guan Y, Yang J, et al. A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records. BMC Bioinformatics 2018 Dec 28;19(Suppl 17):499 [FREE Full text] [doi: 10.1186/s12859-018-2467-9] [Medline: 30591015]

19. Ji B, Liu R, Li S, Yu J, Wu Q, Tan Y, et al. A hybrid approach for named entity recognition in Chinese electronic medical record. BMC Med Inform Decis Mak 2019 Apr 09;19(Suppl 2):64 [FREE Full text] [doi: 10.1186/s12911-019-0767-2] [Medline: 30961597]
20. Zhang Z, Zhou T, Zhang Y, Pang Y. Attention-based deep residual learning network for entity relation extraction in Chinese EMRs. BMC Med Inform Decis Mak 2019 Apr 09;19(Suppl 2):55 [FREE Full text] [doi: 10.1186/s12911-019-0769-0] [Medline: 30961580]
21. Chen L, Li Y, Chen W, Liu X, Yu Z, Zhang S. Utilizing soft constraints to enhance medical relation extraction from the history of present illness in electronic medical records. J Biomed Inform 2018 Nov;87:108-117 [FREE Full text] [doi: 10.1016/j.jbi.2018.09.013] [Medline: 30292854]
22. Sheng M, Shao Y, Zhang Y, Li C, Xing C, Zhang H, et al. An Extensible Framework for Health Knowledge Graph. 2019 Presented at: International Conference on Smart Health; July 1-2, 2019; Shenzhen, China. [doi: 10.1007/978-3-030-34482-5_3]
23. Zhou Y, Qi X, Huang Y, Ju F. Research on Construction and Application of TCM Knowledge Graph Based on Ancient Chinese Texts. 2019 Presented at: IEEE/WIC/ACM International Conference on Web Intelligence; October 14-17, 2019; Thessaloniki, Greece p. 144-147. [doi: 10.1145/3358695.3360938]
24. Jei C, Dehua C, Jiajin L. Study on the construction of knowledge graph of breast tumor based on EMR. Computer Applications and Software 2017;34(12):122-126.
25. Sheng M, Zhang H, Zhang Y, Li C, Xing C, Wang J, et al. A Cross-lingual Knowledge Graph Framework for Cardiovascular Diseases. 2019 Presented at: International Conference on Web Information Systems and Applications; September 18-20, 2019; Vienna, Austria p. 20-22. [doi: 10.1007/978-3-030-30952-7_51]
26. Yin S, Chen D, Le J. Deep Neural Network Based on Translation Model for Diabetes Knowledge Graph. 2017 Presented at: International Conference on Advanced Cloud and Big Data (CBD); Aug 13-16, 2017; Shanghai, China. [doi: 10.1109/cbd.2017.62]
27. Li X, Liu H, Zhao X, Zhang G, Xing C. Automatic approach for constructing a knowledge graph of knee osteoarthritis in Chinese. Health Inf Sci Syst 2020 Dec;8(1):12. [doi: 10.1007/s13755-020-0102-4] [Medline: 32175080]
28. Li L, Wang P, Yan J, Wang Y, Li S, Jiang J, et al. Real-world data medical knowledge graph: construction and applications. Artif Intell Med 2020 Mar;103:101817. [doi: 10.1016/j.artmed.2020.101817] [Medline: 32143785]
29. National Health and Family Planning Commission of PRC. Specification for sharing document of electronic medical record. 2014. URL: https://tinyurl.com/yxr26r64 [accessed 2019-11-28]
30. Zhang K, Li K, Ma H, Yue D, Zhuang L. Construction of MeSH-Like Obstetric Knowledge Graph. 2018 Presented at: International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC); Oct 18-20, 2018; Zhengzhou, China p. 18-20. [doi: 10.1109/cyberc.2018.00041]
31. CIPS_SIGKG. CCKS 2018 CNER challenge. 2018 Presented at: China Conference on Knowledge Graph and Semantic Computing; August 14-17 2018; Tianjing, China URL: https://github.com/liuhuanyong/CCKS2018Summary
32. Noy NF, McGuinness DL. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05. 2001 Mar. URL: https://protege.stanford.edu/publications/ontology_development/ontology101.pdf [accessed 2019-05-21]
33. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May 8;1(1):18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
34. Li L, Hou L. Named Entity Recognition in Chinese Electronic Medical Records Based on the Model of Bidirectional Long Short-Term Memory with a Conditional Random Field Layer. Stud Health Technol Inform 2019 Aug 21;264:1524-1525. [doi: 10.3233/SHTI190516] [Medline: 31438213]
35. Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. 2016 Presented at: ACL 2016: Annual meeting of the Association for Computational Linguistics; August 7-12, 2016; Berlin, Germany. [doi: 10.18653/v1/p16-1101]
36. Zhang Y, Wang X, Hou Z, Li J. Clinical Named Entity Recognition From Chinese Electronic Health Records via Machine Learning Methods. JMIR Med Inform 2018 Dec 17;6(4):e50 [FREE Full text] [doi: 10.2196/medinform.9965] [Medline: 30559093]
37. ICTCLAS: Institute of Computing Technology, Chinese Lexical Analysis System. SEWM. URL: http://sewm.pku.edu.cn/QA/reference/ICTCLAS/FreeICTCLAS/English.html [accessed 2019-06-15]
38. Liu S, Wang L, Chaudhary V, Liu H. Attention neural model for temporal relation extraction. In: 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics; June 7, 2019; Minneapolis, MN p. 134-139. [doi: 10.18653/v1/w19-1917]
39. Xiao J, Zhou Z, Luo X. A Method for Chinese Entity Relationship Extraction Based on Bi-GRU. 2019 Presented at: International Conference on Modeling, Analysis, Simulation Technologies and Applications (MASTA); May 26-27, 2019; Hangzhou, China p. 26-27. [doi: 10.2991/masta-19.2019.9]
40. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. 2014 Presented at: NIPS 2014 Deep Learning and Representation Learning Workshop; December 8-13, 2014; Montreal, Canada.

41. Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. 2016 Presented at: Annual meeting of the association for computational linguistics; Aug 7-12, 2016; Berlin, Germany p. 207-212. [doi: 10.18653/v1/p16-2034]
42. Lin Y, Shen S, Liu Z, Luan H, Sun M. Neural Relation Extraction with Selective Attention over Instances. 2016 Presented at: Annual meeting of the association for computational linguistics; Aug 7-12, 2016; Berlin, Germany p. 2124-2133. [doi: 10.18653/v1/p16-1200]
43. Li X, Rao Y, Sun L, Lu Y. Relation Extraction Based on Dual Attention Mechanism. 2019 Presented at: International Conference of Pioneering Computer Scientists, Engineers and Educators; September 20-23, 2019; Guilin, China p. 20-23. [doi: 10.1007/978-981-15-0118-0_27]
44. Zhu G, Iglesias CA. Exploiting semantic similarity for named entity disambiguation in knowledge graphs. Expert Systems with Applications 2018 Jul;101:8-24. [doi: 10.1016/j.eswa.2018.02.011]
45. He X, Zhang R, Rizvi R, Vasilakes J, Yang X, Guo Y, et al. ALOHA: developing an interactive graph-based visualization for dietary supplement knowledge graph through user-centered design. BMC Med Inform Decis Mak 2019 Aug 08;19(Suppl 4):150 [FREE Full text] [doi: 10.1186/s12911-019-0857-1] [Medline: 31391091]
46. Connect to Neo4j. 2020. URL: http://phdatashare.com:7474/browser/ [accessed 2020-09-27]
47. Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutor Quant Methods Psychol 2012;8(1):23-34 [FREE Full text] [doi: 10.20982/tqmp.08.1.p023] [Medline: 22833776]
48. Fleiss JL, Cohen J. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. Educational and Psychological Measurement 2016 Jul 02;33(3):613-619. [doi: 10.1177/001316447303300309]
49. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]

Abbreviations

API: application programming interface
BiGRU: bidirectional gated recurrent unit
BiLSTM-CRF: bidirectional long short-term memory with a conditional random field
CCKS: China Conference on Knowledge Graph and Semantic Computing
CEMR: Chinese electronic medical record
CLAS: cancer disease type
CNER: Clinical Named Entity Recognition
DCS: disease causes symptoms and signs
EMR: electronic medical record
DSTKG: digestive system tumor knowledge graph
GRU: gated regression unit
i2b2: Informatics for Integrating Biology & the Bedside
ICC: intraclass correlation coefficient
ICD-10: The Tenth Revision of the International Statistical Classification of Diseases and Related Health Problems
LOCI: symptoms and signs are located in the body structure
NCI: National Cancer Institute
NEL: named entity linking
NER: named entity recognition
POS: part-of-speech
RE: relation extraction
REST: representational state transfer
SID: symptoms and signs indicate the disease
SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms
TeCD: test is conducted to investigate the disease
TeRD: test reveals the disease
TeRS: test reveals the symptoms and signs
TrAS: treatment is administered for the symptoms and signs
TrAD: treatment is administered for the disease
TrCS: treatment causes the symptoms and signs
WHO: World Health Organization

Edited by G Eysenbach; submitted 17.02.20; peer-reviewed by Z Huang, E Frontoni; comments to author 29.06.20; revised version received 23.08.20; accepted 22.09.20; published 07.10.20.

Please cite as:
Xiu X, Qian Q, Wu S
Construction of a Digestive System Tumor Knowledge Graph Based on Chinese Electronic Medical Records: Development and Usability Study
JMIR Med Inform 2020;8(10):e18287
URL: http://medinform.jmir.org/2020/10/e18287/
doi:10.2196/18287
PMID:33026359

Original Paper

Predictive Models of Mortality for Hospitalized Patients With COVID-19: Retrospective Cohort Study

Taiyao Wang[1,2,3], PhD; Aris Paschalidis[4]; Quanying Liu[5], PhD; Yingxia Liu[6], MD; Ye Yuan[7], PhD; Ioannis Ch Paschalidis[1,2,3], PhD

[1]Department of Electrical and Computer Engineering, Boston University, Boston, MA, United States

[2]Department of Biomedical Engineering, Boston University, Boston, MA, United States

[3]Center for Information and Systems Engineering, Boston University, Boston, MA, United States

[4]Brown University, Providence, RI, United States

[5]Department of Biomedical Engineering, University of Science and Technology, Shenzen, China

[6]Third People's Hospital of Shenzhen, Second Hospital Affiliated to Southern University of Science and Technology, Shenzen, China

[7]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China

Corresponding Author:
Ioannis Ch Paschalidis, PhD
Department of Electrical and Computer Engineering
Boston University
8 Saint Mary's St
Boston, MA, 02215
United States
Phone: 1 6173530434
Email: yannisp@bu.edu

Abstract

Background: The novel coronavirus SARS-CoV-2 and its associated disease, COVID-19, have caused worldwide disruption, leading countries to take drastic measures to address the progression of the disease. As SARS-CoV-2 continues to spread, hospitals are struggling to allocate resources to patients who are most at risk. In this context, it has become important to develop models that can accurately predict the severity of infection of hospitalized patients to help guide triage, planning, and resource allocation.

Objective: The aim of this study was to develop accurate models to predict the mortality of hospitalized patients with COVID-19 using basic demographics and easily obtainable laboratory data.

Methods: We performed a retrospective study of 375 hospitalized patients with COVID-19 in Wuhan, China. The patients were randomly split into derivation and validation cohorts. Regularized logistic regression and support vector machine classifiers were trained on the derivation cohort, and accuracy metrics (F1 scores) were computed on the validation cohort. Two types of models were developed: the first type used laboratory findings from the entire length of the patient's hospital stay, and the second type used laboratory findings that were obtained no later than 12 hours after admission. The models were further validated on a multicenter external cohort of 542 patients.

Results: Of the 375 patients with COVID-19, 174 (46.4%) died of the infection. The study cohort was composed of 224/375 men (59.7%) and 151/375 women (40.3%), with a mean age of 58.83 years (SD 16.46). The models developed using data from throughout the patients' length of stay demonstrated accuracies as high as 97%, whereas the models with admission laboratory variables possessed accuracies of up to 93%. The latter models predicted patient outcomes an average of 11.5 days in advance. Key variables such as lactate dehydrogenase, high-sensitivity C-reactive protein, and percentage of lymphocytes in the blood were indicated by the models. In line with previous studies, age was also found to be an important variable in predicting mortality. In particular, the mean age of patients who survived COVID-19 infection (50.23 years, SD 15.02) was significantly lower than the mean age of patients who died of the infection (68.75 years, SD 11.83; $P<.001$).

Conclusions: Machine learning models can be successfully employed to accurately predict outcomes of patients with COVID-19. Our models achieved high accuracies and could predict outcomes more than one week in advance; this promising result suggests that these models can be highly useful for resource allocation in hospitals.

(JMIR Med Inform 2020;8(10):e21788) doi:10.2196/21788

KEYWORDS

coronavirus; COVID-19; mortality; Wuhan; China; machine learning; logistic regression; support vector machine; predictive modeling

Introduction

The ongoing pandemic due to the novel coronavirus SARS-CoV-2 has caused worldwide disruption; national governments have imposed drastic measures to contain the pandemic, and the global economy has been impacted [1]. SARS-CoV-2 causes a disease called COVID-19, which is marked by symptoms such as cough, fever, chills, and a range of respiratory symptoms [2]. As of the end of July 2020, the total number of confirmed cases of COVID-19 had surpassed 15 million, and the total number of deaths was approaching 650,000 [3,4].

As the virus continues to proliferate, governments, institutions, and hospitals have struggled to allocate resources such as tests, hospital beds, intensive care unit beds, and ventilators. A significant amount of work has already been performed to predict and track the spread of the virus [3-8]. Recent and ongoing efforts are being made to understand the biomarkers and comorbidities associated with severe COVID-19 disease [9-12]. This work has been important in helping hospitals to classify patients in terms of risk. However, infrastructure to predict hospitalization, mortality, or other patient outcomes is lacking. Predicting these outcomes is essential, as it enables clinicians to make informed decisions regarding patients at risk. For example, clinicians can ensure that the proper resources are allocated to patients who are more likely to require critical care and the use of ventilators.

Using blood samples from patients from Tongji Hospital in Wuhan, China, we used supervised machine learning methods to predict mortality following hospitalization. These machine learning models have been used frequently in the literature for a variety of applications. Some examples include predicting the death of patients with sepsis [13,14], identifying patients at high risk of emergency hospital admissions [15], predicting hospitalization due to heart disease [16,17], and predicting diabetes complications [18,19].

The aim of this retrospective cohort study was to develop accurate models to predict mortality among hospitalized patients with COVID-19 using basic demographics and easily obtainable laboratory data.

Methods

Data Collection

Data were collected between January 10 and February 18, 2020, from patients admitted to Tongji Hospital in Wuhan, China. Data collection was approved by the Tongji Hospital Ethics Committee. The records collected included epidemiological, demographic, clinical, and laboratory results as well as mortality following infection with COVID-19. Data originating from pregnant and breastfeeding women or patients aged younger than 18 years and records with more than 20% missing data were excluded from the analysis [20].

Preprocessing

Prior to model development, several preprocessing measures were undertaken. Variables were standardized by subtracting the mean and dividing by the standard deviation. Variable elimination was performed to reduce the complexity of the resulting model, improve the out-of-sample performance, and enhance the interpretability. Redundant variables and variables with more than 30% missing data were removed. In addition, we computed pairwise Spearman correlations between variables and removed one of the variables if the absolute correlation coefficient was >0.8. Furthermore, missing data in the remaining variables were imputed using the median values of the respective variables. This measure enabled us to include as many patients as possible in our analysis and is a well-documented and popular method of inferring missing values.

Model Development

Data from a total of 375 patients were used to develop the models. These patients were split into two groups to obtain a training set and validation set. The training set was used to train and develop the models, and the validation set was used to determine the accuracy of each model. Unless otherwise noted, 70% of the data were reserved for the training set, and the other 30% were reserved for the validation set. After the data were split into training and validation sets, feature selection was performed to remove several variables. Models were trained using the training set and tested on the validation set. This process was repeated five times, and the average performance and its SD were calculated.

Feature selection was performed using ℓ_1-norm regularization and recursive feature elimination with cross-validation. Specifically, we performed ℓ_1-regularized logistic regression (LR) and obtained the coefficients of the model. We then eliminated the variable with the smallest absolute coefficient and performed the LR again to obtain a new model. We continued this iteration to select a model that maximizes a metric equal to the mean performance minus its SD in a validation data set.

Model Selection

Two different types of regularized models were used in this analysis: ℓ_1-regularized logistic regression (L1LR) models and ℓ_1-regularized support vector machine (L1SVM) models. The models were initially fit to patient data that were collected at any time during the patients' length of stay at the hospital. However, due to the possibility that some laboratory tests were performed close to the patients' outcomes (death or survival), the models were also fit to patient data obtained ≤12 hours after admission. By doing this, we could ensure that the patients' outcomes were predicted as far in advance as possible.

LR, in addition to prediction, provides the likelihood associated with the predicted outcome, which can be used as a confidence measure in decision making.

Model Performance

The performance of the models was evaluated by calculating the weighted F1 score on the validation set. The weighted F1 score is defined as the weighted mean of the F1 score of the positive and negative classes, where the F1 score is defined as the harmonic mean of the precision and the recall. The precision, or positive predictive value (PPV), can be expressed as the ratio of the true positives to the sum of the true positives and false positives. The recall is the true positive rate (ie, the ratio of the true positives to the sum of the true positives and false negatives). The weighted F1 score, unlike the F1 score, considers all the possible outcomes (in this case, survival or death). This can combat potential class imbalance issues and evaluate whether the model accurately predicts mortality and survival, both of which are important in our context. In particular, while identifying patients with higher mortality risk can help direct more resources and attention to those patients, identifying patients who are not at risk is also helpful and can free up resources and time that would otherwise be spent on these lower-risk patients. In addition to the weighted F1 score, we also determined the PPV and the negative predictive value (NPV); the latter is defined as the ratio of the true negatives to the predicted negatives, or the precision of the negative class.

Furthermore, to gain additional insight into the roles of specific variables, we developed a "binarized" counterpart to our sparse LR model. Specifically, we defined a threshold for each variable (using the normal range of the variable) and devised a model in which each variable was either 0 (normal) or 1 (abnormal). For this model, we computed the odds ratio (OR) for each variable; this quantifies how the odds of mortality are scaled by the variable being normal vs abnormal while controlling for the remaining variables.

Statistical Power and External Validation

To assess whether our study cohort size was sufficiently large for the models we derived, we conducted a multiple logistic regression power analysis [21]. This analysis tests the null hypothesis that a specific variable has an LR coefficient equal to zero vs the coefficient value obtained by the model. We set the Type I error probability to 0.05 and the Type II error probability to 0.2 (statistical power of 0.8), from which we obtained a minimum sample size for the variable.

Further, to demonstrate that our models are generalizable, we validated our models on a multicenter external data set. This data set contained data from 432 patients from Shenzhen, China, and 110 patients from Wuhan, China. The data set contained very limited information, encompassing the results of three laboratory tests, the times of the laboratory tests, the discharge time, and the outcome for each patient. Given this limited information, we were only able to validate our best-performing L1SVM model, which uses these three laboratory test values.

Results

Patient Demographics and Laboratory Tests

Table S1 in Multimedia Appendix 1 details patient demographics in addition to various laboratory values for the full patient population. The average age of the patients was 58.83 years (SD 16.46). The mean age of the patients who survived COVID-19 infection (50.23 years, SD 15.02) was significantly lower than the mean age of the patients who did not survive the infection (68.75 years, SD 11.83; $P<.001$). The proportion of men (224/375, 60%) and women (151/375, 40%) in the study cohort was similar. However, more male patients succumbed to infection (126/174, 72%, $P<.001$).

Several laboratory tests were found to have statistically different values among patients who survived and died of COVID-19 infection. Patients who succumbed to infection had LDH values that were roughly 4 times larger than those of patients who survived (755.58 compared to 215.77, $P<.001$). Patients who died also had significantly smaller percentages of lymphocytes and eosinophils in their blood ($P<.001$). Furthermore, the mean level of hs-CRP in patients who died was significantly higher than that in patients who survived ($P<.001$).

As detailed in the Methods section, two different approaches were used to model the data. The first approach was to use blood test results obtained throughout the patients' length of stay at the hospital. Although this approach ensured that there were few missing data points, some of the blood samples were tested close to the patients' outcomes (death or discharge from the hospital). To predict a patient's outcome in advance, a second approach was developed using laboratory test results that were obtained ≤12 hours after the patients' admission to hospital.

Models Using All Laboratory Tests

We first present the results of our predictive models using all laboratory tests. These models were developed as noted in the Methods section. Of the 375 total patients, 24 (6.4%) had incomplete measurements and were omitted, leaving a total of 351 patients (93.6%) for model development. The accuracies of the models using all patient laboratory tests were determined on the validation and external test sets described in the Methods. Complete lists of all the models and their accuracies are provided in Table S2 and Table S3 in Multimedia Appendix 1.

The best-performing models were the ℓ_1-regularized logistic regression model using 4 variables selected by recursive feature selection (L1LR 4) and the ℓ_1-regularized support vector machine model using 3 variables selected by recursive feature selection (L1SVM 3). The L1LR 4 model had a weighted F1 score of 96.98% (SD 0.93%) on the validation set, while the L1SVM 3 model had a score of 97.36% (SD 1.10%). The L1SVM 3 model had a weighted F1 score of 94.55% on the external test set of Shenzhen and Wuhan patients.

The L1LR 4 model had an average validation PPV of 97.61% and an average validation NPV of 96.31%. The L1SVM 3 model had a similarly high average PPV and NPV of 98.27% and 96.71%, respectively. On the multicenter external test set, the accuracy of the L1SVM 3 model remained high (94.55%). Furthermore, both models used a small number of variables in their predictions.

The variables used in each of the best-performing models and the corresponding weight of each variable are reported in Table 1. The logistic regression model used four variables: lactose dehydrogenase (LDH), an enzyme that is found in most living

cells and is typically released when there is tissue damage; the percentage of lymphocytes, a class of immune molecules that are found in the body; hypersensitive C-reactive protein (hs-CRP), a protein that is often used as an indication of heart disease and shows increased levels with inflammation and infection; and albumin, which is one of the main proteins found in blood and is important in regulating the pressure of red blood cells as well as transporting nutrients, proteins, and other molecules. The L1SVM 3 model used the same variables, with the exception of albumin.

Table 1. Coefficients showing the weights of the variables for the two best models.

Variable	Coefficient	
	L1LR 4[a]	L1SVM 3[b]
LDH[c]	1.35	1.44
Percentage of lymphocytes	–0.86	–0.47
hs-CRP[d]	0.74	0.34
Albumin	-0.64	N/A[e]

[a]L1LR 4: ℓ_1-regularized logistic regression model using 4 variables selected by recursive feature selection.

[b]L1SVM 3: ℓ_1-regularized support vector machine model using 3 variables selected by recursive feature selection.

[c]LDH: lactose dehydrogenase.

[d]hs-CRP: hypersensitive C-reactive protein.

[e]N/A: not applicable.

The coefficients obtained by both methods are comparable because the variables were standardized. Therefore, a larger absolute coefficient indicates that the corresponding variable is a more significant predictor. A positive coefficient implies a positive correlation with the outcome, while a negative coefficient implies a negative correlation. Of the variables selected by our models, LDH was considered to be the most important (binarized L1LR 4 OR 55.62, 95% CI 11.41-270.97). The next most important variables were the percentage of lymphocytes (binarized L1LR 4 OR 32.17, 95% CI 5.99-172.90) and hs-CRP (binarized L1LR 4 OR 13.12, 95% CI 3.65-47.23). Finally, the L1LR model found that albumin was important in predicting mortality (binarized L1LR 4 OR 4.08, 95% CI 1.45-11.48). To calculate these ORs, we used a binarized model with the following thresholds: LDH values ≥250 were set to 1, and values <250 were set to 0; lymphocyte percentage values <20 were set to 1, and values ≥20 were set to 0; hs-CRP values ≥10 were set to 1, and values <10 were set to 0; albumin values <34 were set to 1, and values ≥34 were set to 0.

As outlined in the Methods section, a power analysis was performed for the L1LR 4-variable model, and the results indicated that our sample size of 351 patients was sufficient. Specifically, this power analysis indicated that the sufficient numbers of patients to find the LR coefficient were 21 for LDH, 63 for hs-CRP, 61 for the percentage of lymphocytes, and 162 for albumin.

In addition to the previously mentioned models, we also trained models with several important variables removed. More specifically, we removed LDH, albumin, and D-D dimer, a protein that is produced by the degradation of a blood clot. The accuracies of these models were slightly lower than those of the models that included these factors. Furthermore, as we removed more variables, the accuracy of the models decreased. The validation accuracy of the L1LR model with LDH removed was 94.90% (SD 2.13%), the validation accuracy of the L1LR model with LDH and albumin removed was 94.51% (SD 2.19%), and the validation accuracy of the L1LR model with LDH, albumin, and D-D dimer removed was 94.14% (SD 2.5%) (Multimedia Appendix 1 Table S2). The models highlighted several other important factors that were not previously indicated to be important, such as the activity of prothrombin, a protein used in blood clot formation; the platelet count – the count of one of the main cells that makes up blood clots; and age. After these variables were removed, the two most important factors were hs-CRP and the percentage of lymphocytes. When fitting a model to the data using only these two factors, the validation accuracy of the model was 94.87% (SD 1.76%).

Models Using Test Results Obtained ≤12 Hours After Admission

To predict the outcome of a patient with COVID-19 soon after admission to the hospital, we developed several L1SVM models using laboratory test results obtained no later than 12 hours after admission. More specifically, we first performed an ℓ_1-regularized logistic regression to perform feature selection and then fed the selected features into an ℓ_1-regularized support vector machine model. The average time between admission and the time the laboratory test was conducted was 8.4 hours (SD 2.6 hours). Furthermore, the average time between the time of the laboratory test and the patient outcome was 11.5 days (SD 7.5 days).

Table 2 details the average F1 scores and SDs for a select number of the models developed based on data collected ≤12 hours from admission. Table S4 in Multimedia Appendix 1 reports the variables selected by these models. For all models, the L1SVM was performed five times and optimized using a validation set. Of the 375 total patients, 114 (30.4%) had missing data and were excluded, leaving 261 patients (69.6%) for analysis. For these 261 patients, 90% of the data were used for training and 10% of the data were kept as a validation set. As before, the models were fit using all the variables, a limited

number of variables, and all variables other than LDH, albumin, and D-D dimer.

All the models performed well, with accuracies >89% and SDs <5%. The number of variables used in each model varied greatly. The L1SVM All model used 18 of the variables provided in the data set, the L1SVM 7 model used 7 variables, the L1SVM model without LDH and albumin used 10 variables, and the L1SVM model without LDH, albumin, and D-D dimer used 12 variables. Of these models, the model that used 7 variables (including LDH, albumin, and D-D dimer) performed the best, with an accuracy of 94.08% (SD 1.81%). When LDH and albumin were removed from the model, the accuracy decreased by approximately 4%.

These L1SVM models highlighted several key variables that were not indicated by the models that included all laboratory tests. In the models that used all variables, LDH and hs-CRP were consistently two of the most important markers. However, the percentage of lymphocytes found in the blood did not appear to be consistently important in this set of models. Interestingly, the number of neutrophils, a different class of immune marker, in the blood was deemed to be an important variable.

Table 2. Performance of select models developed based on data collected ≤12 hours after admission.

Model	Validation set weighted F1 score (%), mean (SD)
L1SVM all[a]	90.39 (3.25)
L1SVM 7[b]	94.08 (1.81)
L1SVM no LDH[c], albumin[d]	89.65 (4.30)
L1SVM no LDH, albumin, D-D dimer[e]	89.64 (4.89)

[a]L1SVM all: ℓ_1-regularized support vector machine model developed using all the variables in the data set.

[b]L1SVM 7: ℓ_1-regularized support vector machine model using 7 variables.

[c]LDH: lactate dehydrogenase.

[d]L1SVM no LDH, albumin: ℓ_1-regularized support vector machine model developed using all variables except LDH and albumin.

[e]L1SVM no LDH, albumin, D-D dimer: ℓ_1-regularized support vector machine model developed using all variables except LDH, albumin, and D-D dimer.

Discussion

Principal Findings

Our developed L1LR and L1SVM models were able to accurately predict the outcomes of patients with COVID-19, as validated by their weighted F1 scores as high as 97%. In general, the models that used laboratory test results from the duration of the patients' hospital stays were more accurate than models that were restricted to laboratory test results obtained ≤12 hours after admission. However, even when the data were restricted, our models achieved accuracies as high as 94%. These models are more useful because they make predictions upon admission of the patient and thus provide sufficient lead time for making decisions regarding staffing and resource allocation. Because the length of stay of most patients was >1 week, our models can predict a patient's outcome more than one week in advance, with accuracies exceeding 90%.

In many ways, our patient cohort represented a typical cohort of hospitalized patients with COVID-19. In particular, individuals who die of the infection tend to be older and male [22-25]. However, the rate of mortality in our study cohort was higher; close to 50% of the patients admitted to hospital died (174/375, 46.4%). This is likely due to the fact that Tongji Hospital admitted higher numbers of patients with severe and critical disease in Wuhan, China.

The performance of the L1SVM model using all patient laboratory tests on an external multicenter data set suggests that our models are generalizable. The performance of the model decreased by <3% when tested on the external data set compared to the validation set. This indicates that our model could be used by other hospitals worldwide to better understand the risk associated with each patient with COVID-19.

Of particularly importance was the ability of the models to perform well with a small number of predictors. Moreover, the models still performed well when certain key predictors, such as LDH, albumin, and D-D dimer, were removed due to these variables' tendency to exhibit abnormalities at a very late stage of the disease when the outcome is inevitable. The ability of the models to perform well even with few variables can prove particularly useful, as this facilitates interpretation. Furthermore, this ability ensures that predictions can be made even when the outcome is not apparent to a sufficiently experienced physician.

In a recent study, a predictive model was developed based on a few key variables [20]. Different machine learning methods were used in this study, and a decision tree was created. The authors found that LDH, percentage of lymphocytes, and hs-CRP were important predictors of mortality; we also found these three variables to be important. The study's models were very accurate, with F1 scores of approximately 95%. The key difference in our study is that we used laboratory test results obtained ≤12 hours after admission and tested the robustness of the models to the absence of several key variables. Therefore, we are confident that our models can accurately predict patient outcomes well in advance, in the absence of key variables, and even when the outcome may not be obvious to a trained physician.

Limitations

One of the main limitations of this study was the relatively targeted study cohort used to derive the models. These patients lived in Wuhan, China, which was the original epicenter of the outbreak of the novel coronavirus SARS-CoV-2. However, one of our models was validated on an external multicenter cohort of patients from Wuhan and Shenzhen; this suggests that this model can be generalized to other patient cohorts, especially in China. It is less clear how well the models generalize to cohorts in other countries, where patient characteristics and care practices may differ.

Conclusions

We developed multiple state-of-the-art supervised machine learning models to predict the outcome of infection with the novel coronavirus SARS-CoV-2. We were able to predict mortality with greater than 90% accuracy, and we identified several important predictors of mortality.

Acknowledgments

This research was partially supported by the National Science Foundation under grants IIS-1914792, DMS-1664644, and CNS-1645681, by the Office of Naval Research under MURI grant N00014-19-1-2571, and by the National Institutes of Health under grant 1R01GM135930. The authors thank the physicians at Tongji Hospital in Wuhan, China, and Dr George Velmahos at Massachusetts General Hospital for useful discussions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Select patient demographics and laboratory test results, performance of all logistic regression and SVM models evaluated on all laboratory test results, and variables and coefficients of select models evaluated using laboratory test results obtained within 12 hours of admission.
[DOCX File , 24 KB - medinform_v8i10e21788_app1.docx]

References

1. Rolling updates on coronavirus disease (COVID-19). World Health Organization. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen [accessed 2020-05-29]
2. Symptoms of Coronavirus. US Centers for Disease Control and Prevention. 2020 May 13. URL: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html [accessed 2020-05-29]
3. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020 May;20(5):533-534 [FREE Full text] [doi: 10.1016/S1473-3099(20)30120-1] [Medline: 32087114]
4. COVID-19 Map. Johns Hopkins Coronavirus Resource Center. URL: https://coronavirus.jhu.edu/map.html [accessed 2020-07-25]
5. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Dis 2020 May;20(5):553-558. [doi: 10.1016/s1473-3099(20)30144-4]
6. Shen C, Chen A, Luo C, Zhang J, Feng B, Liao W. Using Reports of Symptoms and Diagnoses on Social Media to Predict COVID-19 Case Counts in Mainland China: Observational Infoveillance Study. J Med Internet Res 2020 May 28;22(5):e19421 [FREE Full text] [doi: 10.2196/19421] [Medline: 32452804]
7. Gong M, Liu L, Sun X, Yang Y, Wang S, Zhu H. Cloud-Based System for Effective Surveillance and Control of COVID-19: Useful Experiences From Hubei, China. J Med Internet Res 2020 Apr 22;22(4):e18948 [FREE Full text] [doi: 10.2196/18948] [Medline: 32287040]
8. Yasaka TM, Lehrich BM, Sahyouni R. Peer-to-Peer Contact Tracing: Development of a Privacy-Preserving Smartphone App. JMIR Mhealth Uhealth 2020 Apr 07;8(4):e18936 [FREE Full text] [doi: 10.2196/18936] [Medline: 32240973]
9. Guo W, Li M, Dong Y, Zhou H, Zhang Z, Tian C, et al. Diabetes is a risk factor for the progression and prognosis of COVID-19. Diabetes Metab Res Rev 2020 Mar 31:e3319 [FREE Full text] [doi: 10.1002/dmrr.3319] [Medline: 32233013]
10. Frater JL, Zini G, d'Onofrio G, Rogers HJ. COVID-19 and the clinical hematology laboratory. Int J Lab Hematol 2020 Jun;42 Suppl 1:11-18 [FREE Full text] [doi: 10.1111/ijlh.13229] [Medline: 32311826]
11. Lippi G, Plebani M, Henry BM. Thrombocytopenia is associated with severe coronavirus disease 2019 (COVID-19) infections: A meta-analysis. Clin Chim Acta 2020 Jul;506:145-148 [FREE Full text] [doi: 10.1016/j.cca.2020.03.022] [Medline: 32178975]
12. Qin C, Zhou L, Hu Z, Zhang S, Yang S, Tao Y, et al. Dysregulation of Immune Response in Patients With Coronavirus 2019 (COVID-19) in Wuhan, China. Clin Infect Dis 2020 Jul 28;71(15):762-768 [FREE Full text] [doi: 10.1093/cid/ciaa248] [Medline: 32161940]

13. Jaimes F, Farbiarz J, Alvarez D, Martínez C. Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. Crit Care 2005 Apr;9(2):R150-R156 [FREE Full text] [doi: 10.1186/cc3054] [Medline: 15774048]
14. Vieira SM, Mendonça LF, Farinha GJ, Sousa JM. Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. Applied Soft Computing 2013 Aug;13(8):3494-3504. [doi: 10.1016/j.asoc.2013.03.021]
15. Bottle A, Aylin P, Majeed A. Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis. J R Soc Med 2006 Aug;99(8):406-414 [FREE Full text] [doi: 10.1258/jrsm.99.8.406] [Medline: 16893941]
16. Dai W, Brisimi TS, Adams WG, Mela T, Saligrama V, Paschalidis IC. Prediction of hospitalization due to heart diseases by supervised learning methods. Int J Med Inform 2015 Mar;84(3):189-197 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.10.002] [Medline: 25497295]
17. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. Eur Heart J 2017 Feb 14;38(7):500-507 [FREE Full text] [doi: 10.1093/eurheartj/ehw188] [Medline: 27252451]
18. Brisimi TS, Xu T, Wang T, Dai W, Paschalidis IC. Predicting diabetes-related hospitalizations based on electronic health records. Stat Methods Med Res 2019 Dec;28(12):3667-3682. [doi: 10.1177/0962280218810911] [Medline: 30474497]
19. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, et al. Machine Learning Methods to Predict Diabetes Complications. J Diabetes Sci Technol 2018 Mar;12(2):295-302 [FREE Full text] [doi: 10.1177/1932296817706375] [Medline: 28494618]
20. 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 14;2(5):283-288. [doi: 10.1038/s42256-020-0180-7]
21. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. Statist. Med 1998 Jul 30;17(14):1623-1634. [doi: 10.1002/(sici)1097-0258(19980730)17:14<1623::aid-sim871>3.0.co;2-s] [Medline: 9699234]
22. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, the Northwell COVID-19 Research Consortium, et al. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA 2020 May 26;323(20):2052-2059 [FREE Full text] [doi: 10.1001/jama.2020.6775] [Medline: 32320003]
23. Onder G, Rezza G, Brusaferro S. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. JAMA 2020 May 12;323(18):1775-1776. [doi: 10.1001/jama.2020.4683] [Medline: 32203977]
24. Jordan R, Adab P, Cheng K. Covid-19: risk factors for severe disease and death. BMJ 2020 Mar 26;368:m1198. [doi: 10.1136/bmj.m1198] [Medline: 32217618]
25. Smith-Ray R, Roberts EE, Littleton DE, Singh T, Sandberg T, Taitel M. Distribution of Patients at Risk for Complications Related to COVID-19 in the United States: Model Development Study. JMIR Public Health Surveill 2020 Jun 18;6(2):e19606 [FREE Full text] [doi: 10.2196/19606] [Medline: 32511100]

Abbreviations

hs-CRP: hypersensitive C-reactive protein
L1LR: ℓ1-regularized logistic regression
L1SVM: ℓ1-regularized support vector machine
LDH: lactate dehydrogenase
LR: logistic regression
NPV: negative predictive value
OR: odds ratio
PPV: positive predictive value

Edited by G Eysenbach; submitted 25.06.20; peer-reviewed by E Mahmoudi; comments to author 18.07.20; revised version received 28.07.20; accepted 15.09.20; published 15.10.20.

Please cite as:
Wang T, Paschalidis A, Liu Q, Liu Y, Yuan Y, Paschalidis IC
Predictive Models of Mortality for Hospitalized Patients With COVID-19: Retrospective Cohort Study
JMIR Med Inform 2020;8(10):e21788
URL: http://medinform.jmir.org/2020/10/e21788/
doi:10.2196/21788
PMID:33055061

Viewpoint

Data Object Exchange (DOEx) as a Method to Facilitate Intraorganizational Collaboration by Managed Data Sharing: Viewpoint

Ronald G Hauser[1], MD; Ankur Bhargava[2], MD; Ronald Talmage[3], PhD; Mihaela Aslan[4], PhD; John Concato[5,6], MD

[1]Department of Laboratory Medicine, Yale University School of Medicine, New Haven, CT, United States
[2]Center for Medical Informatics, Yale University, New Haven, CT, United States
[3]Information Technology, Veterans Affairs Puget Sound Healthcare, Seattle, WA, United States
[4]Clinical Epidemiology Research Center, Veterans Affairs Connecticut Healthcare, West Haven, CT, United States
[5]Department of Medicine, Yale University School of Medicine, New Haven, CT, United States
[6]Medical Service, Veterans Affairs Connecticut Healthcare, West Haven, CT, United States

Corresponding Author:
Ronald G Hauser, MD
Department of Laboratory Medicine
Yale University School of Medicine
333 Cedar Street
New Haven, CT,
United States
Phone: 1 2039325711 ext 3120
Email: ronald.hauser@yale.edu

Abstract

Background: To help reduce expenses, shorten timelines, and improve the quality of final deliverables, the Veterans Health Administration (VA) and other health care systems promote sharing of expertise among informatics user groups. Traditional barriers to time-efficient sharing of expertise include difficulties in finding potential collaborators and availability of a mechanism to share expertise.

Objective: We aim to describe how the VA shares expertise among its informatics groups by describing a custom-built tool, the Data Object Exchange (DOEx), along with statistics on its usage.

Methods: A centrally managed web application was developed in the VA to share informatics expertise using database objects. Visitors to the site can view a catalog of objects published by other informatics user groups. Requests for subscription and publication made through the site are routed to database administrators, who then actualize the resource requests through modifications of database object permissions.

Results: As of April 2019, the DOEx enabled the publication of 707 database objects to 1202 VA subscribers from 758 workgroups. Overall, over 10,000 requests are made each year regarding permissions on these shared database objects, involving diverse information. Common "flavors" of shared data include disease-specific study populations (eg, patients with asthma), common data definitions (eg, hemoglobin laboratory results), and results of complex analyses (eg, models of anticipated resource utilization). Shared database objects also enable construction of community-built data pipelines.

Conclusions: To increase the efficiency of informatics user groups, a method was developed to facilitate intraorganizational collaboration by managed data sharing. The advantages of this system include (1) reduced duplication of work (thereby reducing expenses and shortening timelines) and (2) higher quality of work based on simplifying the adoption of specialized knowledge among groups.

(JMIR Med Inform 2020;8(10):e19267) doi:10.2196/19267

KEYWORDS

information-seeking behavior; information services; communication media; database; database management system

Introduction

To help reduce expenses, shorten timelines, and improve the quality of final deliverables, the Veterans Health Administration (VA) and other health care systems seek to promote the sharing of informatics expertise among user groups. This expertise within informatics user groups often develops through individual or small group experience, based on a unique interest or need. Informatics groups with related interests are likely to benefit from each other's expertise, but only if a mechanism exists to offer, find, and exchange expertise. This ability is called knowledge management, and it is described as a utilized, accessible, and efficient virtual system for knowing who is doing what, how, and with what effects [1]. Some authors believe this ability is rarely available, even in the best health care organizations [1].

Knowledge management is especially difficult in large, as opposed to small, health care organizations. As health care systems grow, they tend to become more "loosely coupled" (impersonal and disaggregated). Loosely coupled health care systems operate with tight functional integration within any unit, but few structures or processes tie the organization's units together [1]. This scenario has been referred to as a "silo mentality" [2]. Efficiently finding a suitable collaborator also becomes more difficult, because the possible number of collaborators increases with the size of an organization. According to the Metcalfe law, the number of potential collaborators within a health care system has a squared (n^2) proportional increase [3]. For example, 100 users can form approximately 5000 total collaborations, but 200 users can form nearly 20,000 collaborations. In addition to organizational structure and the combinatorial scale of potential connections between groups, the lack of physical proximity (eg, operation across multiple time zones) and perceived diversity of purpose (eg, financial and patient satisfaction) also represent barriers to collaboration.

Traditional methods to share knowledge have unique considerations when applied to a health care informatics ecosystem. User groups sharing knowledge through the deployment of applications (eg, Docker) will likely generate security concerns. Similar security concerns will also likely exist in sharing source code (eg, GitHub), which could be made into an application. Datasets, as a knowledge-sharing mechanism, may contain protected health information as a necessity. Public data repositories would therefore be reluctant to host data with protected health information (eg, Machine Learning Repository at the University of California Irvine). Transmission of protected health information between groups within a health care system would likely require administrative oversight (such as an approved media of transmission) to mitigate the risk of a data breach. Didactic lectures represent another option to share knowledge, but consumers of shared knowledge may find it inefficient to implement expertise described in a lecture. The VA health care system, while offering many of these existing knowledge-sharing mechanisms, sought additional options.

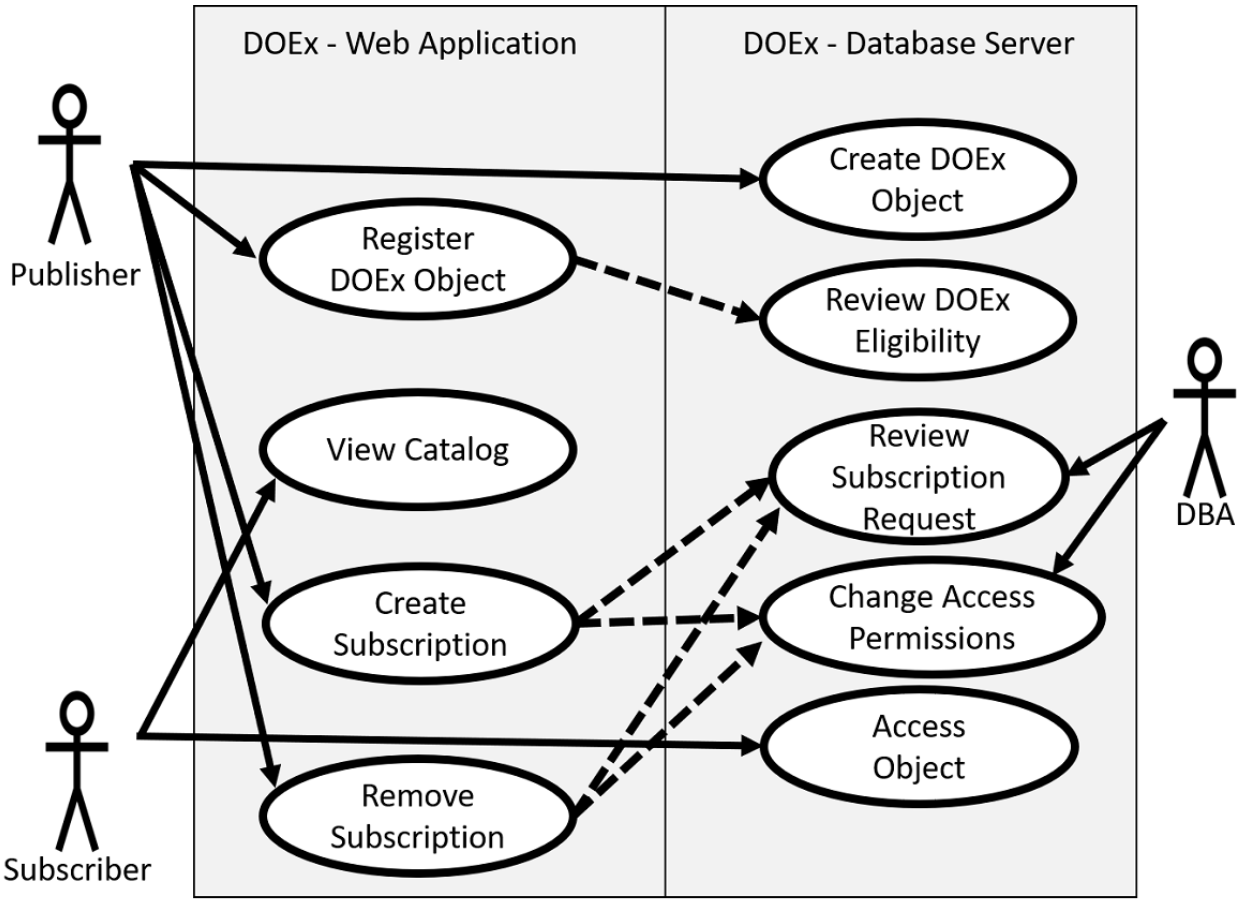
The VA sought to create an environment where informatics user groups could find and exchange data through a secure channel. To find the desired expertise, users browse a catalog of work, where each element in the catalog represents a database object (eg, database table). They can subscribe to catalog elements of interest, which provides access to the object. Experts publish their work to the catalog or share it privately with other user groups. Permissions are centrally administered, and the exchange of data takes place on a secure database server shared between user groups. Inherent advantages of this system include reduced costs and shortened timelines through a decrease in redundant work. Higher-quality deliverables are a likely result, based on the adoption of specialized, rather than generic, knowledge. The design of the VA's solution to promote the sharing of informatics expertise (the Data Object Exchange [DOEx]) and statistics on its usage are described.

Methods

Overview of the DOEx

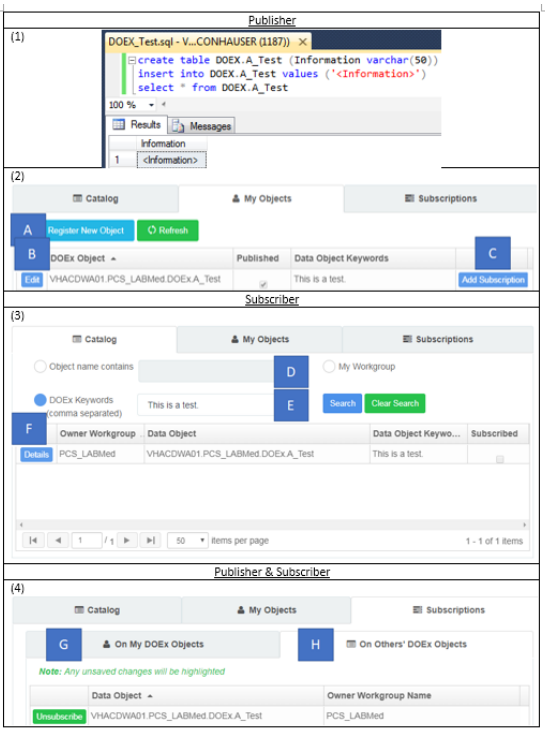
To facilitate collaboration, the VA Business Intelligence Service Line (BISL) designed the DOEx. The DOEx operates with a publish-subscribe design pattern (Figure 1) [4]. Similar to popular subscription services (eg, Wall Street Journal and Netflix), a publisher produces content that subscribers consume. In the DOEx, publishers are collections of individuals, operating as workgroups, with a shared goal or interest. Workgroups may publish their work in a catalog, which all other workgroups can browse. Alternatively, workgroups may publish their work through the DOEx without advertising it in the public catalog. Workgroups can subscribe to published objects they find in the public catalog or learn about through private collaborations. In either case, the subscribing workgroup can request a subscription from the publishing workgroup, which provides them read-only access to the requested object. A walkthrough of the workflow used by the publisher and subscriber is shown in Figure 2.

Figure 1. Use case diagram of the Data Object Exchange (DOEx). The DOEx consists of the following two parts: a web application and a database server (gray boxes). Publishers create and register DOEx database objects. Subscribers become aware of shared objects through the DOEx catalog or private communication with the publisher (not shown). Publishers control subscriptions to their DOEx objects through the creation and removal of subscribers. The database administrator (DBA) provides administrative oversight to ensure adherence to the terms of service.



The DOEx consists of the following two components: a web application and a database server. The web application allows publishers to manage their subscriptions and control the content they choose to publish. The web application also hosts the catalog of published work and provides metadata about each available subscription (eg, object owner and email address, object description, and object location on the database server). The database server contains a collection of databases, typically one per workgroup. Each workgroup database contains a set of data objects (eg, data tables and views), which may be shared through the DOEx. Tables designed for sharing through the DOEx are placed by the publisher in a database schema named "DOEx." Additionally, DOEx objects may be designated as either public or private. Public DOEx objects have a record in the DOEx catalog, which advertises them to potential subscribers (Figures 2 and 3). Private DOEx objects, by comparison, do not exist in the DOEx catalog. Private DOEx objects generally form in the context of existing collaborations between publishers and subscribers.

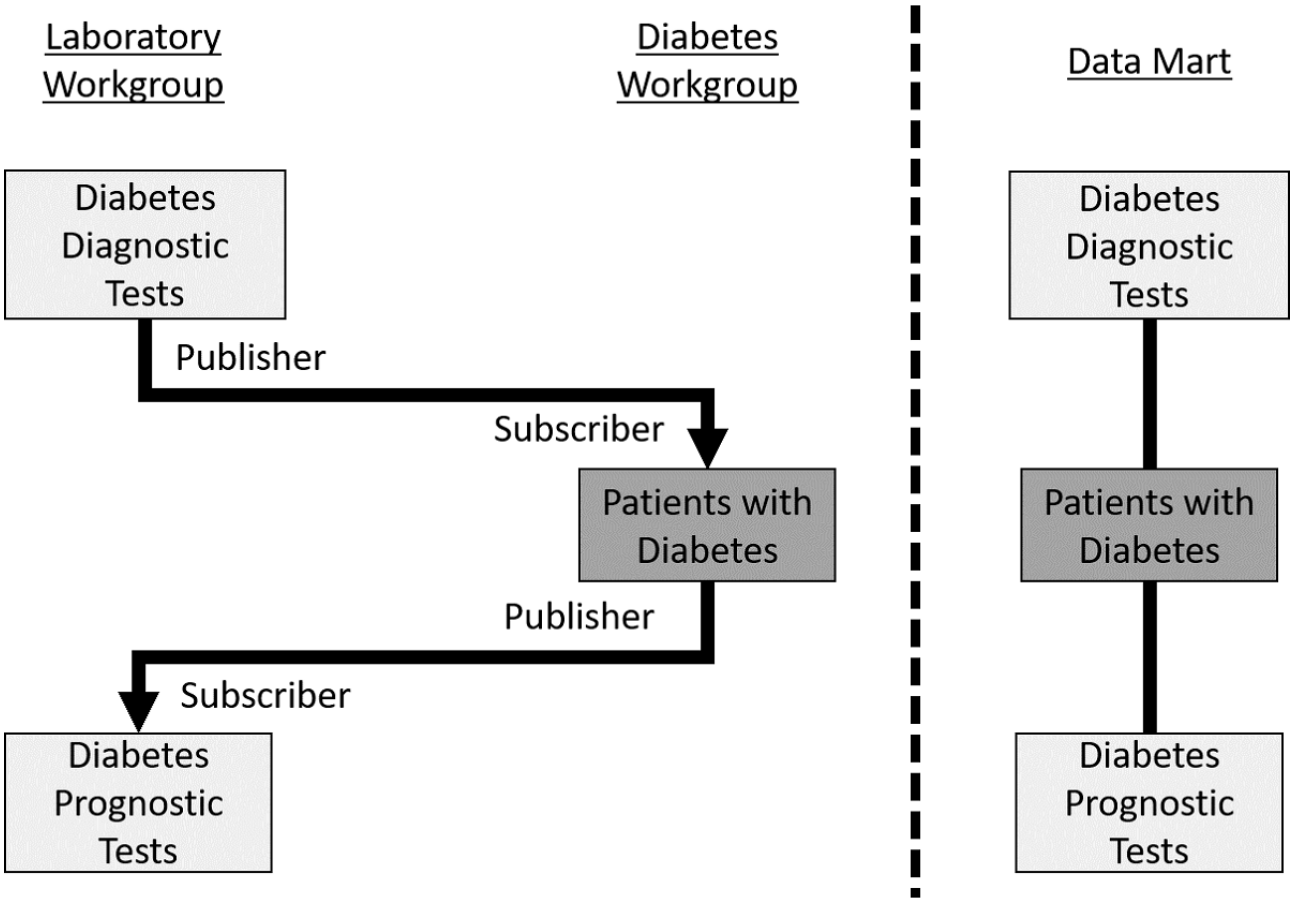
Figure 2. Data Object Exchange (DOEx) web interface. (1) The publisher creates a database object. Example provided in Microsoft SQL Server. (2) The publisher registers an object in the DOEx. (A) "Register New Object" - This button allows the publisher to share the database object via the DOEx. The publisher must provide a description of the object and choose if it will be displayed in the public catalog. (B) "Edit" - Edit the object description and its inclusion in the public catalog. (C) "Add Subscription" - The publisher allows subscribers read-only access to a DOEx object. (3) A subscriber searches the DOEx catalog for objects of interest. (D) Search for public objects by name. (E) Search for public objects by keyword. (F) "Details" - Display the object's description and publisher's email. Subscribers email the publisher to request access. The publisher adds the subscriber with 1C. (4) Publishers and subscribers can manage their subscriptions. (G) Publishers can manage subscriptions. (H) Subscribers can manage subscriptions.



Alongside the publisher and subscriber, a database administrator also participates in the setup and management of object sharing. The web application alerts the database administrator to a publisher-subscriber DOEx request. The database administrator reviews the shared object to ensure compliance with the terms of service of the DOEx. Certain types of data, such as patient names and social security numbers, cannot be shared between workgroups. When these terms are met, the database administrator approves the request.

All modifications to DOEx objects are communicated via email to publishers, subscribers, and database administrators.

Figure 3. Creating a diabetes data mart with the Data Object Exchange (DOEx) and the dynamic data pipeline design pattern. (Left) The publisher/subscriber relationship between the laboratory and diabetes workgroups. Each rectangle represents a dynamic data design pattern. The arrows denote the data processing sequence. (Right) A subset of the final data mart: a database table of patients with diabetes and their prognostic tests.



Data Security Considerations

Permissions granted to workgroups vary based on their need to access protected health information and/or personally identifiable information. When subscribers and publishers have different permissions to protected health information/personally identifiable information, they cannot share objects through the DOEx. This prevents the sharing of sensitive data with workgroups that do not have the necessary permissions.

To limit violations of the terms of service, publishers cannot swap one DOEx object for another with the same name. This type of modification inactivates the sharing of the object with its subscribers. Publishers can only swap one DOEx object for another after removing the subscribers and repeating the process of object registration, which requires database administrator review.

In theory, the web application could perform object registration and subscription services, but this approach would require the web application to have elevated permissions on the databases. Given concerns for security, such services are not currently made available.

Results

Usage Description

The DOEx went live in May 2017, and as of January 2019, 707 database objects were shared among 758 workgroups. The public catalog contains 217 (30.7%) database objects, and the remaining 490 (69.3%) objects are shared privately; these 707 objects have 1202 subscribers. Of the available DOEx objects, 230 have multiple subscribers, and the 10 most popular DOEx objects have between 10 and 40 subscribers.

Design Patterns

Experience with the DOEx has led to the emergence of the follow three common DOEx design patterns: static data, dynamic data, and dynamic data pipelines.

The static data design pattern works well for data that does not change or changes very slowly over time. Examples of static data shared through the DOEx include the laboratory test standard known as Logical Observation Identifiers Names and Codes (LOINC), which is updated twice each year. Accordingly, the static data design pattern, the simplest of the three, requires a single DOEx object.

The dynamic data design pattern works well for data that updates frequently. Typical uses for this design pattern include maintenance of study populations, such as an up-to-date list of patients with diabetes in the health care system. This design

pattern involves the following two DOEx objects: the "dynamic data table" containing up-to-date data and the "update time table" containing a timestamp of the last update. Subscribers to a dynamic design pattern subscribe to both objects. The subscriber uses the timestamp in the "update time table" as a signal to review the updated "dynamic data table."

The third design pattern, the dynamic data pipeline, uses two or more dynamic data design patterns in sequence (Figure 3). Users employ this design pattern to maintain an up-to-date data mart [5]. For example, to construct a data mart of patients with diabetes, the first dynamic data design pattern would aggregate all diagnostic tests for diabetes. A second dynamic data design pattern would assemble the patients with diabetes, derived from those diagnostic tests. A third dynamic data design pattern would aggregate the prognostic tests used to monitor the progression of the cohort's disease. This pipeline executes in a stepwise fashion as follows: the first dynamic data design pattern updates on a schedule, and the remaining dynamic data design patterns update in response. More than one workgroup is likely to contribute to a dynamic data pipeline.

Discussion

Summary of the DOEx

This report introduces the DOEx, an application designed to facilitate the sharing of expertise among informatics user groups within the largest integrated health care system in the United States, the VA. The DOEx currently hosts over 1200 securely managed collaborations. These collaborations are entirely voluntary, rather than centrally planned, and presumably exist because they enhance work performance. No other platform to share expertise among informaticists exists at this scale in any other health care system. Although the impact of these collaborations is not formally quantified, by its very nature, the DOEx will reduce redundant work by allowing users to leverage already existing expertise to achieve their objectives. This reduces costs and shortens the time required to produce deliverables.

The DOEx has also likely resulted in higher quality deliverables, as user groups now take advantage of specialized knowledge, which would be prohibitive for them to recreate. An example of this specialized knowledge includes the care assessment needs (CAN) score, a predictive model for death and readmission [6]. Users throughout our health care system can subscribe to the CAN score DOEx object to incorporate this predictive model into their diverse needs. This work includes, for example, the identification of patients at hospital discharge at risk for readmission, which is an operational focus [7]. Alternatively, researchers have explored the relationship between CAN scores and physical function [8].

Implementation of the DOEx by health care systems besides our own could likely be easily achieved. The premise on which the DOEx operates is simple. It securely manages the authorization and therefore the sharing of database objects among workgroups. All major database vendors offer diverse options for denoting authorization, including Microsoft (SQL Server) [9], Oracle (Oracle Database) [10], and IBM (DB2) [11].

Users of the DOEx have reported positive effects on the VA's informatics ecosystem through allowing DOEx publishers to publicly showcase the content they produce and still maintain control over it. For example, users can register their work in the public catalog, viewable by all users in the health care system, and can unsubscribe users who violate the established collaborative agreement. In this manner, the DOEx promotes a concept deemed psychological ownership ("a bonding such that the organizational member feels a sense of possessiveness toward the target of ownership even though no legal claim exists") [12]. Researchers have shown this encourages further innovation [13,14]. Additional benefits include extrarole performance (defined as behaviors of employees above their stated job requirements) that promote the smooth functioning of an organization [12,15]. For example, the publisher of a widely subscribed DOEx object may become known throughout an organization as a subject matter expert and thus gain additional motivation to share expertise. Additionally, users have conveyed they will review the public DOEx catalog prior to pursuing their assigned task to mitigate the risk of redundant effort.

In addition to user benefits, the DOEx also facilitates the efficient operation of a health care database. Subscribers can access DOEx objects in place, rather than making a duplicate copy on their workgroup database. This reduces the memory requirements of the database. Similarly, when a subscriber recycles the product of another workgroup, they do not need to create it themselves. This reduces the computational burden on the database. These optimizations benefit all users of the health care system by increasing database performance.

Alternatives to the DOEx

Consistent with other authors, we found one example of data sharing within a health care system [1]. The Informatics for Integrating Biology and the Bedside (i2b2) Hive operates as a collection of interoperable services. Services are provided by cells that communicate through a web interface [16]. The design of the i2b2 and DOEx appear to have different use cases. The focus of i2b2 is research since it operates as a National Institutes of Health–funded National Center for Biomedical Computing (NCBC). In contrast, the DOEx was conceived to support operational work, rather than research, within the VA.

Limitations

The DOEx, given all its stated advantages, also has limitations. The publisher-subscriber workgroups must form and maintain an element of trust. This trust can be fostered early in the collaboration, ideally prior to a subscription, by defining the relationship such as a "terms of use" or software license (eg, Apache and GNU General Public License [GPL]).

Subscribers must also trust the content of publishers. The DOEx does not require the exchange of the methods used to create database objects; therefore, some user groups may find it difficult to incorporate work from another group without additional details on the methods.

The DOEx is currently used only by the operations community, including individuals tasked with supporting the day-to-day business operations of the VA. It is not available within the VA research community at this time, given that peer-to-peer sharing of expertise is prohibited on the research database server. Deployment of the DOEx within the research community in the future may facilitate and expand collaboration.

The DOEx also provides read-only (unidirectional) access by subscribers from publishers. Modification of the DOEx to allow read and write (bidirectional) access to DOEx objects would allow subscribers to request individualized output from a publisher. Web services utilize this type of bidirectional communication, such as the model for collaboration found in the i2b2.

At present, the majority of DOEx objects (490/707, 69.3%) are shared privately. The DOEx catalog may not currently contain the breadth of available expertise, and consequently, experts may hesitate to offer their expertise via the DOEx. Improving engagement and awareness of the DOEx (eg, email communications and presentations) will likely improve the number, scope, and quality of offerings in the catalog.

Conclusion

Sharing of expertise within a health care system's informatics community includes the need to develop a workflow allowing workgroups to find, offer, and exchange expertise in a secure manner. To address this need, the DOEx promotes shared informatics expertise across workgroups within a health care system, and it reduces costs and shorten timelines through a decrease in redundant work. The DOEx also produces higher-quality deliverables, based on the adoption of specialized knowledge.

Conflicts of Interest

None declared.

References

1. Dearing JW, Greene SM, Stewart WF, Williams AE. If we only knew what we know: principles for knowledge sharing across people, practices, and platforms. Transl Behav Med 2011 Mar 27;1(1):15-25 [FREE Full text] [doi: 10.1007/s13142-010-0012-0] [Medline: 24073028]
2. McCartney M. Margaret McCartney: Breaking down the silo walls. BMJ 2016 Sep 26;354:i5199. [doi: 10.1136/bmj.i5199] [Medline: 27672074]
3. Metcalfe's law. Wikipedia. URL: https://en.wikipedia.org/wiki/Metcalfe%27s_law [accessed 2018-12-28]
4. Publish–subscribe pattern. Wikipedia. URL: https://en.wikipedia.org/wiki/Publish%E2%80%93subscribe_pattern [accessed 2019-01-27]
5. Star schema. Wikipedia. URL: https://en.wikipedia.org/wiki/Star_schema [accessed 2020-10-13]
6. Box T, Fihn S. Care Assessment Need (CAN) Score and the Patient Care Assessment System (PCAS): Tools for Care Management. U.S. Department of Veterans Affairs. URL: https://www.hsrd.research.va.gov/for_researchers/cyber_seminars/archives/video_archive.cfm?SessionID=713 [accessed 2020-10-13]
7. Hobson A, Curtis A. Improving the care of veterans: The role of nurse practitioners in team-based population health management. J Am Assoc Nurse Pract 2017 Nov;29(11):644-650. [doi: 10.1002/2327-6924.12506] [Medline: 28857487]
8. Serra MC, Addison O, Giffuni J, Paden L, Morey MC, Katzel L. Physical Function Does Not Predict Care Assessment Need Score in Older Veterans. J Appl Gerontol 2019 Mar 29;38(3):412-423 [FREE Full text] [doi: 10.1177/0733464817690677] [Medline: 28380717]
9. Authorization and Permissions in SQL Server. Microsoft. URL: https://docs.microsoft.com/en-us/dotnet/framework/data/adonet/sql/authorization-and-permissions-in-sql-server [accessed 2020-10-13]
10. Authorization: Privileges, Roles, Profiles, and Resource Limitations. Oracle. URL: https://docs.oracle.com/cd/B19306_01/network.102/b14266/authoriz.htm#DBSEG5000 [accessed 2020-10-13]
11. DB2 Version 9.7 for Linux, UNIX, and Windows. IBM. URL: https://www.ibm.com/support/knowledgecenter/en/SSEPGG_9.7.0/com.ibm.db2.luw.admin.sec.doc/doc/c0006307.html [accessed 2020-10-13]
12. Vandewalle D, Van Dyne L, Kostova T. Psychological Ownership: An Empirical Examination of its Consequences. 2016 Jul 26;20(2):210-226. [doi: 10.1177/1059601195202008]
13. Saha CN, Bhattacharya S. Intellectual property rights: An overview and implications in pharmaceutical industry. J Adv Pharm Technol Res 2011 Apr;2(2):88-93 [FREE Full text] [doi: 10.4103/2231-4040.82952] [Medline: 22171299]
14. Reiner BI. Intellectual property in medical imaging and informatics: the independent inventor's perspective. J Digit Imaging 2008 Mar 3;21(1):3-8 [FREE Full text] [doi: 10.1007/s10278-007-9096-6] [Medline: 18175182]
15. Bateman TS, Organ DW. Job Satisfaction and the Good Soldier: The Relationship Between Affect and Employee "Citizenship". Academy of Management Journal 1983 Dec 01;26(4):587-595. [doi: 10.2307/255908]
16. The i2b2 Hive and the Clinical Research Chart. i2b2. URL: https://www.i2b2.org/software/files/PDF/current/HiveIntroduction.pdf [accessed 2020-10-13]

Abbreviations

CAN: care assessment needs
DOEx: Data Object Exchange
VA: Veterans Health Administration

Edited by G Eysenbach; submitted 10.04.20; peer-reviewed by M deBaca, D He; comments to author 01.09.20; revised version received 15.09.20; accepted 15.09.20; published 27.10.20.

Please cite as:
Hauser RG, Bhargava A, Talmage R, Aslan M, Concato J
Data Object Exchange (DOEx) as a Method to Facilitate Intraorganizational Collaboration by Managed Data Sharing: Viewpoint
JMIR Med Inform 2020;8(10):e19267
URL: http://medinform.jmir.org/2020/10/e19267/
doi:10.2196/19267
PMID:33107829

Original Paper

Personalized Web-Based Cognitive Rehabilitation Treatments for Patients with Traumatic Brain Injury: Cluster Analysis

Alejandro Garcia-Rudolph[1,2,3], PhD; Alberto Garcia-Molina[1,2,3], PhD; Eloy Opisso[1,2,3], PhD; Jose Tormos Muñoz[1,2,3], PhD

[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

Corresponding Author:
Alejandro Garcia-Rudolph, PhD
Institut Guttmann Hospital de Neurorehabilitacio
Cami de Can Ruti s/n
Badalona
Spain
Phone: 34 93 497 77 00
Email: alejandropablogarcia@gmail.com

Abstract

Background: Traumatic brain injury (TBI) is a leading cause of disability worldwide. TBI is a highly heterogeneous disease, which makes it complex for effective therapeutic interventions. Cluster analysis has been extensively applied in previous research studies to identify homogeneous subgroups based on performance in neuropsychological baseline tests. Nevertheless, most analyzed samples are rarely larger than a size of 100, and different cluster analysis approaches and cluster validity indices have been scarcely compared or applied in web-based rehabilitation treatments.

Objective: The aims of our study were as follows: (1) to apply state-of-the-art cluster validity indices to different cluster strategies: hierarchical, partitional, and model-based, (2) to apply combined strategies of dimensionality reduction by using principal component analysis and random forests and perform stability assessment of the final profiles, (3) to characterize the identified profiles by using demographic and clinically relevant variables, and (4) to study the external validity of the obtained clusters by considering 3 relevant aspects of TBI rehabilitation: Glasgow Coma Scale, functional independence measure, and execution of web-based cognitive tasks.

Methods: This study was performed from August 2008 to July 2019. Different cluster strategies were executed with Mclust, factoextra, and cluster R packages. For combined strategies, we used the FactoMineR and random forest R packages. Stability analysis was performed with the fpc R package. Between-group comparisons for external validation were performed using 2-tailed t test, chi-square test, or Mann-Whitney U test, as appropriate.

Results: We analyzed 574 adult patients with TBI (mostly severe) who were undergoing web-based rehabilitation. We identified and characterized 3 clusters with strong internal validation: (1) moderate attentional impairment and moderate dysexecutive syndrome with mild memory impairment and normal spatiotemporal perception, with almost 66% (111/170) of the patients being highly educated (P<.05); (2) severe dysexecutive syndrome with severe attentional and memory impairments and normal spatiotemporal perception, with 49.2% (153/311) of the patients being highly educated (P<.05); (3) very severe cognitive impairment, with 45.2% (42/93) of the patients being highly educated (P<.05). We externally validated them with severity of injury (P=.006) and functional independence assessments: cognitive (P<.001), motor (P<.001), and total (P<.001). We mapped 151,763 web-based cognitive rehabilitation tasks during the whole period to the 3 obtained clusters (P<.001) and confirmed the identified patterns. Stability analysis indicated that clusters 1 and 2 were respectively rated as 0.60 and 0.75; therefore, they were measuring a pattern and cluster 3 was rated as highly stable.

Conclusions: Cluster analysis in web-based cognitive rehabilitation treatments enables the identification and characterization of strong response patterns to neuropsychological tests, external validation of the obtained clusters, tailoring of cognitive web-based tasks executed in the web platform to the identified profiles, thereby providing clinicians a tool for treatment personalization, and the extension of a similar approach to other medical conditions.

(JMIR Med Inform 2020;8(10):e16077) doi:10.2196/16077

KEYWORDS

cluster analysis; traumatic brain injury; web-based rehabilitation

Introduction

Background

Every year, more than 50 million people worldwide experience a traumatic brain injury (TBI). It is estimated that about half the world's population will have one or more TBIs in their lifetime. TBI is the leading cause of mortality in young adults and a major cause of death and disability across all ages worldwide, as recently reported in *The Lancet Neurology* [1]. Cognitive impairments due to TBI are the significant sources of morbidity in the affected individuals, their family members, and in the society. Disturbances in attention, memory, and executive functioning are the most common cognitive consequences of TBI at all levels of severity [2,3]. The clinical picture of TBI is characterized by a wide heterogeneity because of the nature and location of the injury [4]. Patients with TBI can show various combinations of motor, cognitive, behavioral, psychosocial, and environmental issues that have a huge impact on everyday activities [5], and these issues can greatly interfere with the effectiveness of rehabilitation interventions. It has been proposed that the efficacy of the rehabilitation would increase if programs moved from disease-centered to person-centered issues such that the rehabilitation is tailored to individual needs [6,7]. A number of studies have suggested that brain injury does not have any prototypical pattern of cognitive performance and outcome but may be best characterized by heterogeneity, both in regard to cognitive deficit and ultimate level of functioning [8]. TBI is an extremely heterogeneous disorder ranging from mild reversible conditions, often characterized as concussion, to severe massively destructive trauma, sometimes resulting in death. Saatman et al [9] highlighted the problem as follows: "The heterogeneity of TBI is considered as one of the most significant barriers to finding effective therapeutic interventions."

Clustering in TBI

TBI is a heterogeneous disease, and the mechanism/location of injury, premorbid functioning, secondary complications, and numerous other factors can influence cognitive performance [10]. As cognitive performance is a robust indicator of the current functioning and the prognostic outcome [11], it is critical to identify subgroups of patients who have distinct cognitive profiles that, in turn, can assist in treatment planning and patient care [12]. This can be empirically accomplished using cluster analysis, which is a multivariate classification technique that allows for statistical grouping of like cases into homogeneous subsets (or clusters) based on their similarity across one or more characteristics. Cluster analysis allows for the identification of homogeneous subgroups wherein cognitive heterogeneity is present based on the similarities in performance on neuropsychological tests.

Cluster analysis has been extensively applied in the study of TBI in the last 30 years [13-31]. Nevertheless, we have identified several common limitations such as the number of TBI patients that were clustered (<100 in many studies), the clustering approaches (only hierarchical clustering and k-means and not discussing other possible techniques), the specific implementation of such techniques (most of them restricted to only commercial products), as well as the lack of relation between the obtained clusters and rehabilitation tasks. The details are presented in Supplementary Material Table A1 (see Multimedia Appendix 1).

Web-Based Cognitive Rehabilitation and Cluster Analysis

Cognitive rehabilitation has been playing an ever-increasing role in the treatment of patients with TBI who have cognitive deficits. The data gathered support the idea that improvements attributed to rehabilitation may generalize beyond task-specific skills [32]. Since the number of patients that could be eligible for this type of treatment is ever increasing, it is essential to develop new strategies that may improve access without elevating the costs to deliver such care [33]. The incorporation of computers and information technology-based systems in current clinical practice contributes to optimizing cognitive interventions, that is, their intensity, personalization, patient adherence, and quality of professional monitoring [34,35]. The types of cognitive rehabilitation programs that are the most effective in improving cognitive skills are still unclear [36]. Approaches that are designed to accommodate each individual's cognitive strengths and weaknesses, offer instant item-specific feedback, and dynamically adapt the rehabilitation program accordingly appear to be the most effective, especially in populations with particular cognitive needs [37]. The objective of this study was to contribute to the personalization of web-based cognitive rehabilitation and to identify and characterize subgroups of patients who have distinctive profiles obtained from standard neuropsychological tests administered to patients before starting the rehabilitation.

Main Characteristics of This Study

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

Guttmann, NeuroPersonalTrainer

Guttmann, NeuroPersonalTrainer (GNPT)) is the web-based cognitive rehabilitation platform used in this study. GNPT addresses the desired features outlined in the previous section in the following manner.

1. It uses a baseline cognitive evaluation based on standardized neuropsychological tests to individualize the training regimen.
2. It continually adapts the difficulty level according to the subject's performance by using an interactive-adaptive system.
3. It provides detailed graphic and verbal feedback after each rehabilitation task execution.

This study focuses on the baseline cognitive evaluation to individualize rehabilitation. Personalization of cognitive rehabilitation is accomplished by using a baseline cognitive

evaluation, the results of which determine the individual content and the level of subsequent training for each participant. During rehabilitation, personalization is maintained by an adaptive feature that continually measures the subject's performance, adapts the difficulty level of the training tasks, and provides detailed graphic and verbal performance feedback during and after each task. Because the rehabilitation regimen is designed based on the results of the cognitive evaluation and because the program continually adapts to each person's strengths and weaknesses, it is unlikely that 2 participants can receive the same regimen with regard to the choice of tasks, amount, and intensity of rehabilitation in each cognitive domain.

Baseline Assessment: International Classification of Functioning Disability and Health

Baseline cognitive evaluation is performed in GNPT using the conceptual framework of the International Classification of Functioning, Disability and Health (ICF) [4]. The ICF belongs to a family of international classifications developed by the World Health Organization. ICF aims to provide a unified and standard language and framework for the description of health and health-related status. Direct punctuations obtained by patients in neuropsychological tests are mapped to the ICF 0-4 scale, representing the level of impairment, and they are expressed using ICF as complete disability (4), severe disability (3), moderate disability (2), mild disability (1), and no problem (0). The baseline assessment consists of the following 12 functions: categorization, divided attention, flexibility, inhibition, planning, selective attention, sequencing, spatial and temporal perception, sustained attention, verbal memory, visual gnosis, and working memory.

Individual Clustering Approaches

While numerous clustering algorithms have been published and new ones continue to appear, there is no single algorithm that has been shown to dominate other algorithms across all application domains [38]. Therefore, as an initial step, we proposed to study different clustering approaches in our application domain (the assessment instruments described in the previous section), and we tried different number of clusters (k). Clustering algorithms can be broadly divided into 2 groups: hierarchical and partitional (hierarchical has been applied in most publications presented in Table A1, Multimedia Appendix 1). In this study, we applied the following hierarchical and partitional algorithms: a hierarchical agglomerative algorithm AGNES (AGglomerative NESting), a hierarchical divisive DIANA (DIvisive ANAlysis), the classic k-means implementation, 2 partitional alternatives, that is, PAM (Partitioning Around Medoids) and CLARA (Clustering LARge Applications) [39], and a model-based clustering using the MClust software [40,41] (details are presented in Table A1, Multimedia Appendix 1).

Combined Approaches: Principal Component Analysis and Random Forest

As alternatives to individual clustering approaches, in this work, we present 2 combined approaches: principal component analysis (PCA) and random forest.

PCA can be viewed as a denoising method, which separates signal and noise: the first dimensions extract the essential parts of the information while the last ones are restricted to noise. Without the noise in the data, the clustering is more stable than the one obtained from the original distances. Consequently, if a hierarchical tree is built from another subsample of individuals, the shape of the top of the hierarchical tree remains approximately the same. PCA is thus considered as a preprocessing step before performing clustering methods [42]. PCA has been scarcely applied in previous research, as shown in Table A1 (Multimedia Appendix 1). In this study, we propose an integrated approach of PCA and hierarchical clustering.

Another recently proposed dimensionality reduction strategy is random forest. It consists of a collection or ensemble of classification trees, wherein each tree is grown with a different bootstrap sample of the original data. Each tree votes for a class and the majority rule is used for the final prediction. Random forests can be used in both supervised and unsupervised learning. In unsupervised random forests, the data is classified without a priori classification specifications. Synthetic classes are generated randomly and the trees are grown. Despite the synthetic classes, similar samples will end up in the same leaves of the trees owing to each tree's branching process. The proximity of the samples can be measured and a proximity matrix is constructed. In this study, we propose the application of an unsupervised random forest integrated with the PAM clustering method [43].

Study Objectives

We proposed to identify and characterize cognitive profiles in a web-based cognitive rehabilitation platform by using cluster analysis with the following specific aims:

1. Apply state-of-the-art cluster validity indices (CVIs) to different cluster strategies (hierarchical, partitional, and model-based) to identify meaningful classes.
2. Apply combined strategies of dimensionality reduction and clustering by using PCA and random forests to improve the obtained CVIs.
3. Characterize the identified profiles by using demographic and clinically relevant variables.
4. Study the external validity of the obtained clusters by considering 2 relevant aspects of TBI rehabilitation: functional independence measure (FIM) assessment (as well as Glasgow Coma Scale [GCS] for severity) at admission and rehabilitation and cognitive training tasks executed all along the rehabilitation process.

Methods

Participants

Our study consisted of patients with TBI who were admitted in the Rehabilitation Unit of the Acquired Brain Injury Department of a tertiary institution (Institut Guttmann, Spain). The period of the study was from August 2008 to July 2019.

This study was performed in accordance with the Declaration of Helsinki of the World Medical Association and approved by the ethics committee of the Clinical Research of this institution. Signed informed consent was obtained from every patient or

their relatives after full explanation of the procedures. The inclusion criteria for the study were as follows: adult patients with the diagnosis of TBI and without any previous comorbidities leading to disability. Participants were excluded for illiteracy and inability to undergo formal cognitive evaluation for clinical reasons (eg, excessive sleepiness, bedridden patients, or uncontrolled sharp pain).

Cognitive Evaluation: ICF Mapping

Initial cognition assessments used as input to cluster analysis were obtained through standardized administration of neuropsychological tests on admission; most of them were also applied to the state-of-the-art cluster analysis, as shown in Table A1 (Multimedia Appendix 1): Wisconsin Card Sorting Test, Barcelona Test, Rey Auditory Verbal Learning test, Wechsler Adult Scale III (digit span forward and backward), and Trial Making Test (Part A and Part B). All direct punctuations obtained by patients in each test were then mapped to the 0.4 ICF values. Details on the mapping of assessment instruments to ICF are presented in a previous study [44].

Individual Cluster Analysis Approaches: Proposed Implementations

In this study, we took the 12 cognitive functions assessments (each one ranging from 0 to 4) as input to clustering techniques. For agglomerative hierarchical clustering, we applied the hclust function of the stats R package [45] and the AGNES function of the cluster [46] R package. For divisive hierarchical clustering, we applied the DIANA function of the cluster R package. The eclust function of the factoextra [47] R package was applied for the classic k-means implementation. The PAM function of the cluster R package was applied for PAM clustering, and similarly, the CLARA function of the same package was applied. For model-based clustering, the MClust [48] R package was applied.

Combined Cluster Analysis Approaches: Unsupervised Random Forest Method

We proceeded using the following steps [43]:

1. The unsupervised random forest algorithm was used to generate a proximity matrix using the randomForest [49] R package.
2. PAM clustering of this first proximity matrix generated the initial classes.
3. A supervised random forest analysis of the initial classes allowed the calculation of out-of-bag error rates and the determination of the importance of the variables in relation to their contribution to accuracy in the classification.
4. Repeated the unsupervised random forest analysis with the most important variables to generate a second proximity matrix.
5. Repeated PAM clustering using the second proximity matrix to generate the new classes.
6. We then calculated the CVIs with the cluster.stats function of the fpc R package.

Combined Approaches: PCA Method

We then considered an alternative approach, which combined dimensionality reduction and clustering: the hierarchical clustering on principal components (HCPC) function of the FactoMineR [50] R package. It involves the following steps:

1. Compute the principal components: PCA function for quantitative variables
2. Compute hierarchical clustering: It is performed using the Ward's criterion on the selected principal components. Ward criterion is used because it is based on the multidimensional variance like PCA.
3. Choose the number of clusters based on the hierarchical tree: An optimal partitioning is proposed by HCPC to cut the hierarchical tree obtained using the AGNES technique.
4. Perform k-means clustering to improve the initial partition obtained from hierarchical clustering. The final partitioning solution, obtained after consolidation with k-means, can be (slightly) different from the one obtained with the hierarchical clustering.

Performance Measures: Internal Validation and Stability

We then proposed to compare the internal validity (based only on the clustered data) of the resulting clusters based on the CVIs. These include average silhouette width [51], average Pearson gamma [52], entropy [53], Dunn index [52], and within-between cluster ratio (a higher metric of the former 3 statistics and a smaller within-between cluster ratio indicating a better fitting; eg, Clinical Cancer Research [54]). We focused especially on average silhouette width based on the conclusions in a recent review [55]. We applied the cluster.stats function of the fpc R package [56] to each of the proposed techniques for different number k of clusters, in order to obtain the CVIs. We focused on the average silhouette width by considering the following criteria [51]: 0.71-1.0, a strong structure has been found; 0.51-0.70, a reasonable structure has been found; 0.26-0.50, a weak structure has been found and could be artificial; and <0.25, no substantial structure has been found. In order to assess if the cluster holds up under plausible variations in the dataset (stability), our approach was to perform bootstrap resampling to evaluate the stability of a given cluster [57]. The cluster stability of each cluster in the original clustering is the mean value of its Jaccard coefficient over all the bootstrap iterations.

Performance Measures: External Validation

As in previous publications presented in Table A1 (Multimedia Appendix 1), in order to validate any cluster solution, it is important to compare the resulting clusters on variables that were not included in the original clustering process [25]. Various demographic variables were examined for this purpose. Regarding statistical analysis, first, analysis of the homogeneity of variance by Levene's test and normality of distribution by the Kolmogorov-Smirnov test were conducted. Chi-square tests were conducted for most of these variables because of their ordinal nature (eg, gender), whereas analyses of variance were performed with interval variables such as age. $P<.05$ was considered statistically significant. We included external variables that were described in previous studies such as gender, age, age ranges, education level, FIM [58], and severity at admission measured using the GCS. In Table A2 (Multimedia Appendix 1), we have included a detailed description of FIM and GCS.

A standard cognitive rehabilitation treatment in GNPT takes 2-5 months, which is distributed in 2-5 sessions a week, and each session is composed of 4-10 cognitive training tasks. GNPT integrates a set of about 100 web-based cognitive tasks, each of which mainly addresses one of the 12 functions described above. Typically, each patient executes a different number of tasks along with treatment and in a different order. For each execution, the patient obtains an immediate result (ranging from 0 to 100, as the percentage of compliance) [59].

Results

Sample Description

A final sample of 574 adult patients with TBI who performed web-based cognitive rehabilitation training in the GNPT platform were included in this study. The study was performed from August 1, 2008 to July 1, 2019. Of the 574 patients, 105 (18.3%) were women and 469 (81.7%) were men. Their distribution in the age ranges was as follows: 241 (42.0%) in the 17-30 years range, 259 (45.1%) in the 31-55 years range, and 74 (12.9%) in the >56 years range. With respect to the education level, of the 574 patients, 9 (1.6%) patients had completed primary education, 259 (45.1%) had completed secondary education, 205 (35.7%) completed tertiary education, and 101 (17.6%) completed post-tertiary education. The data of the severity of TBI at admission was available for 455 of the 574 patients (79.3%) by using the GCS, and the data were as follows: 44 (9.6%) had mild head injury, 57 (12.5%) had moderate head injury, and 354 (77.8%) had severe head injury.

Baseline Clustering

In order to run the implementations of the different algorithms presented in the Methods section, input parameters were selected as mentioned in previous state-of-the-art publications presented in Table A1 (Euclidean distance and Ward criteria). As the initial preprocessing phase, we performed Spearman correlation analysis by using the corrplot [60] R package in order to identify highly correlated variables. Figure 1 shows the correlation matrix among the 12 initial variables, which is colored according to the correlation coefficient. We observed the following 3 variables with $r>0.80$ and $P<.001$: flexibility, sequencing, and working memory. Therefore, we removed them for clustering.

Table 1 shows the internal validation results for different k values and for the 6 proposed clustering techniques.

Figure 1. Correlogram of the initial set of cognitive variables. CAT: categorization; DIV, divided attention; FLEX: flexibility; INH: inhibition; PLAN: planning; SEL: selective attention; SEQ: sequencing; SPTEMP: spatiotemporal perception; SUS: sustained attention; VERB: verbal memory; VISGN: visual gnosis; WORK: working memory.

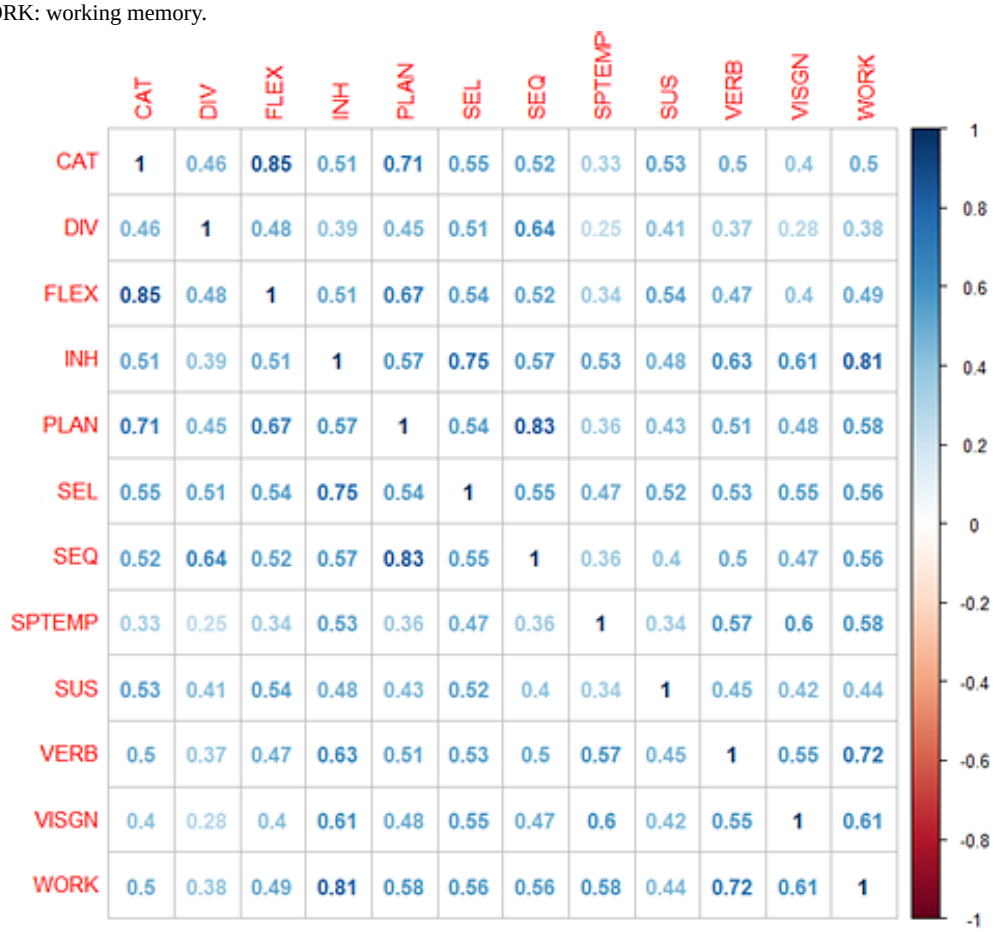


Table 1. Internal validation of the proposed techniques for different number of clusters.

k for the different clusters	Average silhouette width	Pearson gamma	Entropy	Dunn index	Within-between cluster ratio
AGNES (AGglomerative NESting)					
2	0.2947696	0.4446782	0.4193705	0.1466471	0.6827765
3	0.3120659	0.6255316	0.9566682	0.1212678	0.5789247
4	0.2664549	0.619668	1.055581	0.1280369	0.5791948
5	0.2177597	0.6209335	1.173125	0.1324532	0.5722961
6	0.2196451	0.5775517	1.445714	0.1336306	0.5558157
DIANA (DIvisive ANAlysis)					
2	0.384397	0.5435043	0.5454738	0.09667365	0.6445731
3	0.3427734	0.6395711	1.020228	0.1125088	0.5581218
4	0.2918808	0.6340931	1.171331	0.1178511	0.5513584
5	0.2603633	0.615897	1.311116	0.1178511	0.5523393
6	0.2569563	0.547962	1.66604	0.1178511	0.5064622
K-means					
2	0.3683991	0.534673	0.5815533	0.09712859	0.6497419
3	0.3580276	0.6373444	1.0217	0.1125088	0.5584067
4	0.3010858	0.5778943	1.342837	0.1212678	0.5373173
5	0.2906528	0.5374244	1.602541	0.1212678	0.5092503
6	0.2925957	0.5358799	1.759259	0.1360828	0.4798148
PAM (Partitioning Around Medoids)					
2	0.3950954	0.541951	0.5103802	0.09407209	0.6434552
3	0.3558112	0.6379774	1.022812	0.1125088	0.557843
4	0.2912489	0.5634748	1.351603	0.1195229	0.5445628
5	0.2801431	0.5516723	1.543992	0.1360828	0.5210802
6	0.2889038	0.5414131	1.736641	0.1360828	0.4846027
CLARA (Clustering LARge Applications)					
2	0.3917212	0.544271	0.524071	0.09667365	0.6431833
3	0.3496284	0.6216692	1.038644	0.1125088	0.5626294
4	0.292958	0.5662381	1.362308	0.1125088	0.5388617
5	0.2809109	0.5496645	1.567395	0.125	0.5131381
6	0.2890209	0.5298343	1.76809	0.125	0.4816711
MClust					
2	0.2518519	0.4154374	0.6893485	0.1005038	0.7113782
3	0.2897848	0.5629296	1.070214	0.1091089	0.5900842
4	0.2389266	0.5382842	1.208638	0.1097643	0.592125
5	0.2462358	0.5178902	1.506097	0.1118034	0.553266
6	0.1889053	0.4791652	1.569551	0.1118034	0.5790436

Random Forest: Classification Errors

We then calculated random forest classification with 2000 trees as input parameters, and we obtained the following overall out-of-bag errors for the different k values: 1.05% (k=3), 3.83% (k=4), and 5.23% (k=5). In Supplementary Material Table A3 (Multimedia Appendix 1), we present the confusion matrix for the different k values. When calculating variable importance, there was a loss of 20% in accuracy when removing the less important variable (visual gnosis) and 25% loss when removing inhibition, as shown in Supplementary Material Figure A2 (Multimedia Appendix 1). Therefore, no variable was removed, and we did not proceed to steps 4 and 5 of the methodology.

PCA

Since FactoMineR uses a singular value decomposition algorithm, the PCA is calculated over the standardized correlation matrix, wherein a matrix of 40 uncorrelated components is obtained. Table S1 in Supplementary Material (Multimedia Appendix 1) shows the percentage of variance and the eigenvalues for the first 9 components of this matrix. The remaining components (31) correspond to a residual amount of variance. By selecting only the first 3 principal components, we reduced the dimensionality of the multivariate description so that the graphical representation and its subsequent interpretation were simplified. The first 3 principal components described 75.53% of the total variance. The first component described 55.04% of the variance, the second one described 13.42%, and the third component described 7.06%. In the case of the goodness of fit, we relied on the following metrics to verify the choice of the first 3 components: the root mean square of the residuals is 0.05 and the fit based upon off-diagonal values is 0.99.

We then ran the HCPC function with the following parameters: min=2, max=10, distance=Euclidean, criteria=Ward, and agglomerative hierarchical clustering.

When specifying min=2 and max=10 as parameters, HCPC identified the optimal k value maximizing the inertia gain. As shown in Supplementary Material Figure A3 (Multimedia Appendix 1), inertia gain dramatically decreased after the third class; therefore k=3 is the optimal partition proposed by HCPC.

Internal Validation: Summary of the Results

When testing HCPC internal validation with the same indicators as presented in Table 1, we obtained the following CVIs: within-between ratio, 0.3706104; entropy, 0.9873104; Dunn index, 1.849996; Pearson gamma, 0.6511913; and average silhouette width, 0.515794. These CVIs clearly outperformed the CVIs presented in Table 1. For the individual approaches, the best average silhouette width was obtained by PAM for k=2 (0.395) and by k-means for k=3 (0.358). When the average silhouette width ranges from 0.26 to 0.50, the identified structure is weak and can be artificial. We focused especially on the average silhouette width, based on the conclusions in a recent CVI review [55], where 30 different indices with 720 synthetic and 20 real datasets were compared. A group of 10 indices was found to be the most recommended, with silhouette at the top in both synthetic and real datasets. Nevertheless, when considering the other CVIs in Table 1, the within-between ratio (the lower the better) HCPC was also the lowest, and Pearson gamma (the higher the better) was also higher for HCPC than any other in Table 1.

In relation to the random forest approach, when calculating variable importance, there was a loss of 20% in accuracy when removing the less important variable (visual gnosis) and 25% loss when removing inhibition. A previous study [43] removed variables leading to less than 5% loss in accuracy. In our case, no variable was removed, and therefore, we did not proceed to steps 4 and 5 of the methodology.

Characterization of the Final Clusters

As presented in Table 2, the following clusters were found: cluster 1 (n=170), cluster 2 (n=311), and cluster 3 (n=93).

Table 2 shows statistically significant results for the education level of the participants as well as for all the involved cognitive functions. Analysis of cluster rationale indicated that cluster 1 is characterized by the highest level of education with almost 66% (66/170, 38.8% + 45/170, 26.5%) of its participants having tertiary or post-tertiary education. Meanwhile less than half of the participants in the other two clusters reach such educational levels: 49.2% (42/311, 13.5% + 111/311, 35.7%) of cluster 2 participants and 45.2% (14/93, 15.1% + 28/93, 30.1%) of cluster 3 participants. Furthermore, cluster 3 was characterized as complete impairment in all cognitive functions. Therefore, this cluster was characterized as very severe cognitive impairment. Meanwhile, cluster 1 presented mild impairment in working memory, visual gnosis, spatiotemporal perception, and inhibition and moderate impairment in categorization, divided attention, flexibility, planning, and sequencing. We characterized this cluster as highly educated, moderate attentional impairment, and moderate dysexecutive syndrome with mild memory impairment, and good spatiotemporal perception. Cluster 2 presented severe impairment in executive functioning (flexibility, categorization, and planning) and presented the highest degree of impairment in divided attention, as well as severe impairment in selective attention. Therefore, this cluster was characterized by severe dysexecutive syndrome with severe attentional and memory impairment and good spatiotemporal perception.

Table 2. Univariant analysis of the obtained clusters (N=574).

	Cluster 1, n=170	Cluster 2, n=311	Cluster 3, n=93	*P* value
Age (years), mean (SD)	43.3 (14.4)	43.1 (15.2)	43.1 (14.5)	
Gender, n (%)				.84
Women	30 (17.6)	56 (18.0)	19 (20.4)	
Men	140 (82.4)	255 (82.0)	74 (79.6)	
Education level, n (%)				<.05
Post-tertiary	45 (26.5)	42 (13.5)	14 (15.1)	
Primary	6 (3.53)	3 (0.96)	0 (0.0)	
Secondary	53 (31.2)	155 (49.8)	51 (54.8)	
Tertiary	66 (38.8)	111 (35.7)	28 (30.1)	
Age range (years), n (%)				.12
17-30 years	61 (35.9)	131 (42.1)	49 (52.7)	
31-55 years	86 (50.6)	138 (44.4)	35 (37.6)	
56+ years	23 (13.5)	42 (13.5)	9 (9.68)	
Baseline assessments, mean (SD)				
Categorization	2.14 (1.20)	3.72 (0.64)	4.00 (0.00)	<.001
Divided attention	2.34 (1.53)	3.94 (0.23)	4.00 (0.00)	<.001
Flexibility	2.12 (1.17)	3.58 (0.74)	4.00 (0.00)	<.001
Inhibition	0.64 (0.89)	2.34 (1.25)	4.00 (0.00)	<.001
Planning	2.09 (1.10)	3.56 (0.69)	4.00 (0.00)	<.001
Selective attention	1.58 (0.86)	3.29 (0.85)	4.00 (0.00)	<.001
Sequencing	2.06 (1.14)	3.57 (0.69)	4.00 (0.00)	<.001
Spatial and temporal perception	0.17 (0.44)	0.37 (0.64)	4.00 (0.00)	<.001
Sustained attention	1.35 (1.22)	3.03 (1.28)	3.71 (0.73)	<.001
Verbal memory	1.75 (1.01)	2.65 (0.95)	4.00 (0.00)	<.001
Visual gnosis	0.23 (0.59)	0.95 (1.30)	4.00 (0.00)	<.001
Working memory	0.73 (0.89)	1.95 (1.16)	4.00 (0.00)	<.001

External Validation

We performed twofold external validation: (1) by using demographic and clinical variables (age, gender, education level, age ranges) and then by using FIM and GCS evaluations at admission and (2) considering all cognitive tasks executed by the patients in GNPT during the period under study. We found no statistically significant differences when considering age, gender, or age ranges. The total number of available FIM assessments at admission was 439 of the original 574 participants (76.5%). Table 3 shows the number of participants, the mean, median, and IQRs for total FIM as well as the motor and cognitive subtotals for each cluster.

Table 3. Total functional independence measure, cognitive, and motor subtotals by cluster (N=439).

Measures	Cluster 1, n=138	Cluster 2, n=238	Cluster 3, n=63	*P* value
Total functional independence measure				<.001
Mean (SD)	87.88 (33.55)	71.303 (38.07)	68.698 (39.26)	
Median (Q1, Q3)	96.50 (65.25, 117.00)	73.000 (35.00, 108.00)	73.000 (28.00, 105.00)	
IQR	18.00-126.00	18.00-126.00	18.00-126.00	
Cognitive functional independence measure				<.001
Mean (SD)	26.96 (7.99)	22.58 (9.77)	21.452 (10.29)	
Median (Q1, Q3)	29.00 (23.00, 33.00)	25.00 (15.00, 31.00)	22.00 (13.00, 30.00)	
IQR	5.00-35.00	5.00-35.00	5.00-35.00	
Motor functional independence measure				<.001
Mean (SD)	60.91 (27.175)	48.72 (30.02)	47.58 (30.47)	
Median (Q1, Q3)	68.50 (40.00, 85.75)	48.00 (18.00, 79.00)	42.000 (14.00, 76.00)	
IQR	13.00-91.00	13.00-91.00	13.00-91.00	

Regarding total FIM, patients in the 3 clusters required assistance for up to 25% of the tasks but cluster 3 was quite close to requiring assistance for 50% of the tasks. When considering the motor subtotal score with a maximum possible score of 91, patients in cluster 1 obtained 60.91, while cluster 2 obtained less than 50 and cluster 3 obtained 47.58. Regarding the cognition subtotal score (maximum score 35), cluster 1 was almost 30 while clusters 2 and 3 were close to 20.

In relation to GCS, the total number of available GCS assessments at admission was 455 (79.3%) of the original 574 participants. Table 4 shows the number of participants, mean, median, and IQRs for each cluster, and it shows the highest values for cluster 1, followed by cluster 2, and the lowest for cluster 3. Further, the IQR for cluster 3 ranged from 3 to 7, which was lower than that in clusters 1 and 2.

Regarding the second external validation, in GNPT, each task addresses a specific cognitive function. Table 5 shows the number of tasks for each function executed by cluster, with a total of 151,763 executions during the whole period under study.

Table 4. Total Glasgow Coma Scale measures by cluster (N=455).

Glasgow coma scale measures, *P*<.006	Cluster 1, n=136	Cluster 2, n=241	Cluster 3, n=78
Mean (SD)	7.19 (3.76)	6.40 (3.39)	5.50 (2.80)
Median (Q1, Q3)	7.00 (4.00, 10.00)	6.00 (4.00, 8.00)	4.50 (3.00, 7.00)
IQR	3.00-15.00	3.00-15.00	3.00-14.00

Table 5. Total task executions by cluster for all participating patients.

Task execution	Cluster 1, n=41,374	Cluster 2, n=89,577	Cluster 3, n=20,812	Total, N=151,763
Functions (*P*<.001), n (%)				
Categorization	2137 (5.2)	4257 (4.8)	591 (2.8)	6985 (4.6)
Divided attention	3673 (8.9)	7239 (8.1)	1038 (5.0)	11,950 (7.9)
Flexibility	2470 (6.0)	5149 (5.7)	1642 (7.9)	9261 (6.1)
Inhibition	2565 (6.2)	5605 (6.3)	1358 (6.5)	9528 (6.3)
Planning	4636 (11.2)	9907 (11.1)	2114 (10.2)	16,657 (11.0)
Selective attention	4776 (11.5)	12,460 (13.9)	4879 (23.4)	22,115 (14.6)
Sequencing	3239 (7.8)	6067 (6.8)	1140 (5.5)	10,446 (6.9)
Sustained attention	2907 (7.0)	9324 (10.4)	3206 (15.4)	15,437 (10.2)
Verbal memory	9230 (22.3)	16,756 (18.7)	3162 (15.2)	29,148 (19.2)
Visual gnosis	657 (1.6)	2830 (3.2)	75 (0.4)	3562 (2.3)
Working memory	5084 (12.3)	9983 (11.1)	1607 (7.7)	16,674 (11.0)

Figure 2 shows the tasks result boxplots for 5 representative functions. Cluster 1 (at the left of each subplot) shows higher performance (punctuations closer to 100) than cluster 2, with cluster 3 showing lower punctuations. As shown in Table 2, for example, for the categorization function, the respective mean values for clusters 1, 2, and 3 were as follows: 2.14 (1.20), 3.72 (0.64), and 4.00 (0.00). The Figure 2 boxplots for the categorization function somehow reflect such different levels. Figure 3 represents the obtained results in every task execution for 2 functions: verbal memory and working memory. Verbal memory was the function with the largest number of executions, as shown in Table 5: 19.2% (29,148 of the total 151,763 task executions). In Figure 3, we present only cluster 1 (blue) and cluster 2 (red) in order to visually show their results, summarized weekly and plotted yearly during the whole period under study. Figure 3 shows that the working memory tasks have been integrated to the system in 2010, whereas verbal memory task executions started in 2008. For verbal tasks, cluster 1 patients outperformed cluster 2 during almost the whole period under study. Working memory tasks behave similarly, with a higher performance of cluster 2 patients.

Figure 2. Tasks results boxplots for 5 cognitive functions: cluster 1 (red), cluster 2 (green), and cluster 3 (blue). CAT: categorization; DIV: divided attention; SEL: selective attention; SUS: sustained attention; VISGN: visual gnosis.

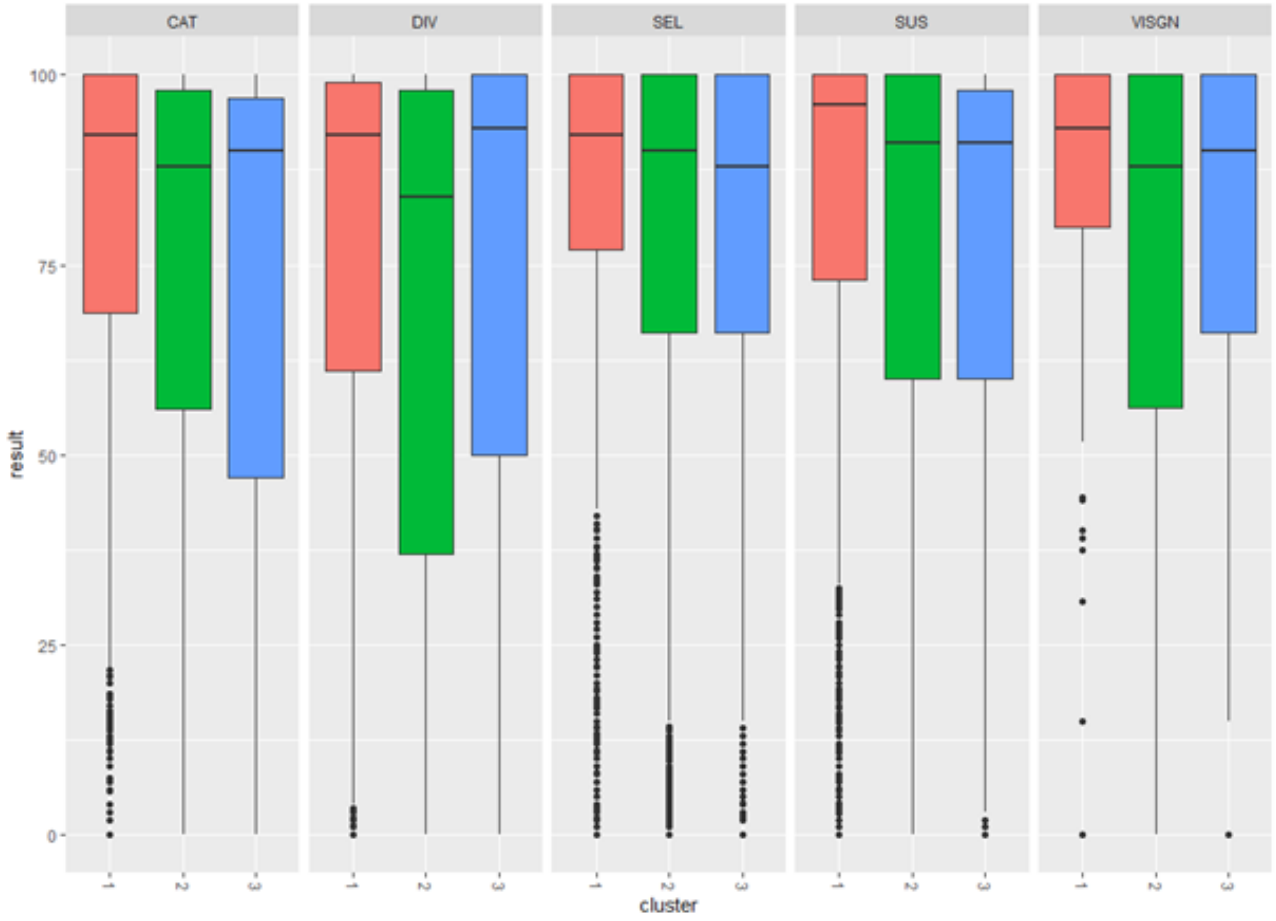
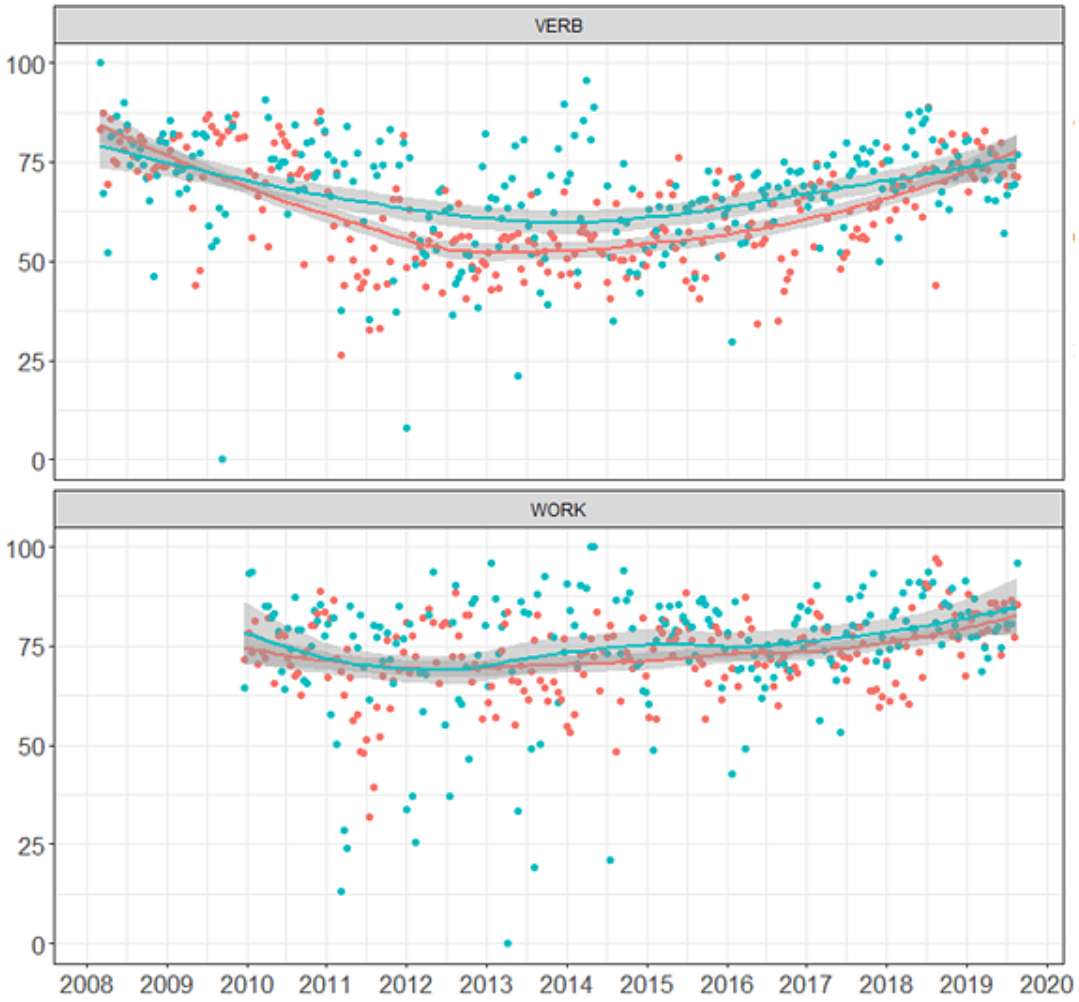


Figure 3. Mean values of the results in task executions summarized weekly, cluster 1 (blue) and cluster 2 (red). VERB: verbal memory; WORK: working memory.



Stability

Values between 0.60 and 0.75 indicate that the cluster is measuring a pattern in the data, but there is no high certainty about which points should be clustered together. Clusters with stability values above 0.85 can be considered highly stable (they are likely to be real clusters). The obtained values by cluster were 0.7524206, 0.6647378, and 0.9910572. Therefore, there were 2 clusters with stability >0.75. As a rule of thumb, clusters with a stability value less than 0.60 should be considered unstable, which is not our case. Therefore, meaningful valid clusters as the ones identified in our study should not disappear if the data set is changed in a nonessential way. Nevertheless, it could also be of interest whether clusters remain stable under the addition of outliers; such cases should be individually considered by clinicians (eg, in case of the lowest GCS assessment values).

Discussion

Principal Findings

In this study, we proposed the application of cluster analysis to a chronic health condition in a GNU framework by using a set of publicly available R libraries (R-3.5.1) in the context of a web-based cognitive platform. We proposed 6 specific clustering techniques (ie, PAM, CLARA, AGNES, DIANA, k-means, and MClust) and 2 combined approaches (HCPC=PCA+AGNES and random forest+PAM) and evaluated them by using state-of-the-art CVIs. It is straightforward to apply both the individual techniques and the combined approaches to other acquired brain injury populations in the same web-based platform (GNPT) or in others. For example, in the Multimedia Appendix 1, we present an initial correlation analysis for patients who had an ischemic stroke that we will address in future work. We obtained the best CVIs with the combined HCPC=PCA+AGNES hierarchical clustering, with average silhouette over 52%; therefore, a *reasonable structure has been found.* We performed stability analysis, and clusters 1 and 2 were rated as 0.60 and 0.75, indicating that the clusters are measuring a pattern, and cluster 3 was rated as highly stable. We identified 3 clearly different profiles. Cluster 1 was characterized as highly educated, moderately distracted, with dysexecutive syndrome and good working memory. Cluster 2 was characterized as severe dysexecutive syndrome and severely distracted. Cluster 3 identified a group of patients with severe symptoms in all the involved functions. External validity in functional independence confirms this characterization by means of severity using GCS and functionality in the activities of daily living, especially when considering the motor FIM subtotal. When considering the performance in the cognitive tasks

executed during the whole period, task results confirmed the identified profiles, with cluster 1 visual representation showing higher values during the whole period than cluster 2. Similar results were obtained when visualizing cluster 3.

Clinical Implications

The actual GNPT implementation integrates an automatic therapy planning functionality, the intelligent therapy assistant (ITA) [61]. The ITA provides therapists with a recommended schedule of cognitive tasks to be executed by each patient during a given period of time. The recommendations provided by the ITA can always be manually modified by therapists according to their own clinical criteria. The ITA takes a predefined set of patient's cognitive profiles as the starting point, which have been obtained using the baseline cognitive evaluation (mapped to ICF as described in the Methods section) as input to CA. When a new patient starts cognitive training in GNPT, the ITA dynamically assigns the patient to the appropriate cluster. The ITA then schedules different cognitive tasks during a user-defined rehabilitation period to the new patient, according to several criteria (eg, usage score, improvement score, clinical score) as described in previous studies. Therefore, the first clinical implication involves the ITA starting point to configure patients' treatments. During therapy, when the patient executes a task (and obtains the result ranging from 0 to 100), GNPT automatically generates another version of the task with a higher or lower difficulty level—increasing the difficulty if the result was "too high" or decreasing the difficulty if the result was "too low" [62]. A second clinical implication involves linking cognitive profiles with performance in task execution. As shown in Figure 3, this allows therapists to identify patterns in performance, for example, results seem to be too close to 50 for cluster 2 in verbal memory tasks during the 2013-2016 period. The current clinical working hypothesis in relation to patient's performance in GNPT tasks is that the optimal range of results is 65-85 [63]. Therefore, Figure 3 (top, verbal memory) suggests that difficulty levels in such tasks might have been too high for patients in cluster 2 during the 2013-2016 period. A more appropriate approach regarding the optimal range of results could be to consider such ranges to vary in relation to clusters. Therefore, a patient in cluster 1 would have a different optimal range than a patient in cluster 2. The next step is to consider the optimal range of the results depending on the cognitive profiles identified by cluster analysis (instead of considering a fixed optimal range as it is now). Future work should also include comparing ITA current cluster analysis results [61] with clusters 1, 2, and 3 obtained in this work for patients with TBI. The integration of cluster analysis as the initial phase of an ITA process also allows for a straightforward extension of a similar approach to other medical conditions, for example, patients who had a stroke, as we present in the Supplementary Material (Multimedia Appendix 1).

Limitations of This Study

First, we conducted a single-center study; an advantage of this is that data were obtained and included by clinicians trained in neurological rehabilitation, and all patients were managed under the same TBI rehabilitation protocols. The GNPT platform is already integrated into the clinical practice of several acquired brain injury centers; nevertheless, their patients were not included in this analysis. A multicenter TBI study may include an initial preprocessing phase, wherein patients are grouped according to their initial GCS severity in order to avoid additional heterogeneity. Thereafter, cluster analysis techniques, as those proposed in this study, may be applied within such groups. External validation assessments, common to all participating centers, is also an important aspect to be addressed in this future multicenter study. Second, the health area studied belongs mainly to the urban population, with a small rural population or populations from other regions.

Third, our analysis lacked computerized tomography or magnetic resonance imaging examinations that describe the presence of contusion, hematoma, hemorrhage, ischemia, or other signs of parenchymal lesion on frontal, temporal, parietal, occipital, and cerebellar lobes or diffuse axonal injury. Fourth, our sample did not include any patient with missing data. All data used as input to cluster analysis are complete. Although there are several R packages addressing the subject (MICE, MissForest, HMISC), we decided to address the problem of missing data in a separate future analysis in order to consider not only the possible imputation strategies but also the reasons for missing data and include such reasons when characterizing the clusters. Fifth, our analysis did not include indicators of mental health or other comorbidities. Persons who experience TBI may have 1 or more preexisting medical comorbidities at the time of injury (eg, alcohol use and depression). Other medical conditions may occur simultaneously with TBI, such as orthopedic trauma, or these conditions may develop afterward as a direct consequence of the TBI such as epilepsy. Still, other medical comorbidities may begin months or years following injury in comparison to uninjured control groups. Studies have suggested that individuals with TBI have more than twice the rates of pain, growth hormone deficiency, insomnia, fatigue, new-onset stroke, urinary incontinence, and epilepsy [64]. Therefore, we aim to include comorbidity analysis in future research studies.

Comparison with Prior Work

We have worked with public GNU libraries, as opposed to the state-of-the-art publications presented in Table A1, wherein most techniques were implemented using commercial packages [15-18,20-23,25-27,29-31]. Previous research presented in Table A1 applied clustering techniques in a batch mode as desktop applications. In our case, the work was integrated in the context of a web-based cognitive training platform. Our baseline assessment consisted of 12 cognitive functions, thereby allowing for a comprehensive description of the patient's profiles, involving cognitive aspects addressed by such different functions, ranging from visual attention to gnosis. Meanwhile, previous clustering research presented in Table A1 addresses specific functions—only one of them in most cases: memory [14,16,18-21,24-26,30], executive functions [17,21,31], or attention [22]. We have proposed different clustering techniques and applied state-of-the-art CVIs to all of them. We have taken advantage of the web-based platform by increasing the number of participants, whereas in only 3 of the 20 studies in Table A1, n is larger than 300 [20,25,30]. We have included the whole set of cognitive tasks performed by all participants as part of the

external validation during the whole period under study (more than 150,000 task executions). We have visually mapped such executions to the obtained clusters along time. To the best of our knowledge, the linking of specific rehabilitation tasks to the obtained clusters has not been yet performed in the state-of-the-art publications presented in Table A1.

Conclusions

Cluster analysis in web-based cognitive rehabilitation treatments allows for identifying and characterizing strong patterns of response to neuropsychological tests, externally validating the obtained clusters by using important aspects of TBI rehabilitation such as severity or functional independence in activities of daily life, tailoring cognitive web-based tasks available in the web platform to the identified profiles by providing clinicians a tool for treatment personalization, which were not addressed in previous traditional cluster analyses, and straightforward extension of a similar approach to patients with other medical conditions, for example, for patients who have had a stroke.

Acknowledgments

This study was partially funded by the INNOBRAIN project: New Technologies for Innovation in Cognitive Stimulation and Rehabilitation (COMRDI15-1-0017). ACCIÓ-Comunitat RIS3CAT d'innovació en salut NEXTHEALTH (COM15-1-0004) cofinanced this project under the FEDER Catalonia 2014-2020 Operational Program.

Authors' Contributions

JTM and AGM conceived the study, AGR and AGM collected, selected, and cleaned the data; they also analyzed the data. AGR drafted the initial manuscript. AGM, EO, and JTM revised the manuscript critically for important intellectual content and approved the final manuscript. AGR, AGM, EO, and JTM received funding for the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Previous studies of cluster analysis of traumatic brain injury based on neuropsychological tests; R code and plots of applied techniques, random forest approach details, and principal component analysis approach details.
[DOCX File , 1603 KB - medinform_v8i10e16077_app1.docx]

References

1. Maas A, Menon D, Adelson P. Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research. Lancet Neurol 2017. [doi: 10.1016/S1474-4422(17)30371-X]
2. Sohlberg MM, Mateer CA. Cognitive Rehabilitation:An interactive Neuropsychological Approach. New York, USA: Guilford Publications; 2018:8147.
3. Stuss, T. D, Winocur, G, Robertson I. Cognitive Neurorehabilitation: Evidence And Application. Cambridge, United Kingdom: Cambridge University Press; 2008.
4. World HO. International Classification of Functioning, Disability, and Health. 2020. URL: https://apps.who.int/iris/bitstream/handle/10665/42407/9241545429.pdf?sequence=1 [accessed 2020-05-20]
5. Turner-Stokes L, Disler PB, Nair A, Wade DT. Multi-disciplinary rehabilitation for acquired brain injury in adults of working age. Cochrane Database Syst Rev 2005 Jul 20(3):CD004170. [doi: 10.1002/14651858.CD004170.pub2] [Medline: 16034923]
6. Sansonetti D, Nicks RJ, Unsworth C. Barriers and enablers to aligning rehabilitation goals to patient life roles following acquired brain injury. Aust Occup Ther J 2018 Dec;65(6):512-522. [doi: 10.1111/1440-1630.12492] [Medline: 29920690]
7. Plant SE, Tyson SF, Kirk S, Parsons J. What are the barriers and facilitators to goal-setting during rehabilitation for stroke and other acquired brain injuries? A systematic review and meta-synthesis. Clin Rehabil 2016 Sep;30(9):921-930 [FREE Full text] [doi: 10.1177/0269215516655856] [Medline: 27496701]
8. Allen DN, Goldstein G. In: Allen DN, Goldstein G, editors. Cluster analysis in neuropsychological research: Recent applications. New York: Springer; 2013.
9. Saatman KE, Duhaime A, Bullock R, Maas AI, Valadka A, Manley GT, Workshop Scientific TeamAdvisory Panel Members. Classification of traumatic brain injury for targeted therapies. J Neurotrauma 2008 Jul;25(7):719-738 [FREE Full text] [doi: 10.1089/neu.2008.0586] [Medline: 18627252]
10. Reitan R, Wolfson D. The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation (2nd ed.). Germany: Tucson: Neuropsychology Press; 1993.
11. Hanks RA, Millis SR, Ricker JH, Giacino JT, Nakese-Richardson R, Frol AB, et al. The predictive validity of a brief inpatient neuropsychologic battery for persons with traumatic brain injury. Arch Phys Med Rehabil 2008 May;89(5):950-957. [doi: 10.1016/j.apmr.2008.01.011] [Medline: 18452745]

12. Spitz G, Ponsford JL, Rudzki D, Maller JJ. Association between cognitive performance and functional outcome following traumatic brain injury: a longitudinal multilevel examination. Neuropsychology 2012 Sep;26(5):604-612. [doi: 10.1037/a0029239] [Medline: 22799747]
13. Crosson B, Greene R, Roth D, Farr S, Adams R. WAIS-R pattern clusters after blunt-head injury. Clinical Neuropsychologist 1990 Aug;4(3):253-262. [doi: 10.1080/13854049008401908]
14. Haut MW, Shutty MS. Patterns of verbal learning after closed head injury. Neuropsychology 1992;6(1):51-58. [doi: 10.1037/0894-4105.6.1.51]
15. Malec JF, Machulda MM, Smigielski JS. Cluster analysis of neuropsychological test results among patients with traumatic brain injury (TBI): Implications for a model of TBI-related disability. Clinical Neuropsychologist 1993 Jan;7(1):48-58. [doi: 10.1080/13854049308401887]
16. Millis SR, Ricker JH. Verbal learning patterns in moderate and severe traumatic brain injury. J Clin Exp Neuropsychol 1994 Aug;16(4):498-507. [doi: 10.1080/01688639408402661] [Medline: 7962354]
17. Donders J, strom D. Factor and cluster analysis of the Intermediate Halstead Category Test. Child Neuropsychology 1995 Apr;1(1):19-25. [doi: 10.1080/09297049508401339]
18. Deshpande SA, Millis SR, Reeder KP, Fuerst D, Ricker JH. Verbal learning subtypes in traumatic brain injury: a replication. J Clin Exp Neuropsychol 1996 Dec 04;18(6):836-842. [doi: 10.1080/01688639608408306] [Medline: 9157108]
19. Wiegner S, Donders J. Performance on the California Verbal Learning Test After Traumatic Brain Injury. J Clin Exp Neuropsychol 1999 Apr;21(2):159-170. [doi: 10.1076/jcen.21.2.159.925] [Medline: 10425514]
20. Curtiss G, Vanderploeg R, Spencer J, Salazar A. Patterns of verbal learning and memory in traumatic brain injury. J Int Neuropsychol Soc 2001 Jul 27;7(5):574-585. [doi: 10.1017/S1355617701755051]
21. Demery JA, Pedraza O, Hanlon RE. Differential profiles of verbal learning in traumatic brain injury. J Clin Exp Neuropsychol 2002 Sep;24(6):818-827. [doi: 10.1076/jcen.24.6.818.8400] [Medline: 12424655]
22. Chan RCK, Hoosain R, Lee TMC, Fan YW, Fong D. Are there sub-types of attentional deficits in patients with persisting post-concussive symptoms? A cluster analytical study. Brain Inj 2003 Feb;17(2):131-148. [doi: 10.1080/0269905021000010168] [Medline: 12519640]
23. van der Heijden P, Donders J. WAIS-III factor index score patterns after traumatic brain injury. Assessment 2003 Jun;10(2):115-122. [doi: 10.1177/1073191103010002001] [Medline: 12801182]
24. Mottram L, Donders J. Cluster subtypes on the California verbal learning test-children's version after pediatric traumatic brain injury. Dev Neuropsychol 2006;30(3):865-883. [doi: 10.1207/s15326942dn3003_6] [Medline: 17083297]
25. Donders J. A confirmatory factor analysis of the California Verbal Learning Test--Second Edition (CVLT-II) in the standardization sample. Assessment 2008 Jun;15(2):123-131. [doi: 10.1177/1073191107310926] [Medline: 18187398]
26. DeJong J, Donders J. Cluster subtypes on the California Verbal Learning Test-Second Edition (CVLT-II) in a traumatic brain injury sample. J Clin Exp Neuropsychol 2010 Nov;32(9):953-960. [doi: 10.1080/13803391003645640] [Medline: 20408004]
27. Thaler NS, Linck JF, Heyanka DJ, Pastorek NJ, Miller B, Romesser J, et al. Heterogeneity in Trail Making Test performance in OEF/OIF/OND veterans with mild traumatic brain injury. Arch Clin Neuropsychol 2013 Dec;28(8):798-807. [doi: 10.1093/arclin/act080] [Medline: 24145667]
28. Harman-Smith YE, Mathias JL, Bowden SC, Rosenfeld JV, Bigler ED. Wechsler Adult Intelligence Scale-Third Edition profiles and their relationship to self-reported outcome following traumatic brain injury. J Clin Exp Neuropsychol 2013;35(8):785-798. [doi: 10.1080/13803395.2013.824554] [Medline: 23947758]
29. Zimmermann N, Pereira N, Hermes-Pereira A, Holz M, Joanette Y, Fonseca RP. Executive functions profiles in traumatic brain injury adults: Implications for rehabilitation studies. Brain Inj 2015 May 07;29(9):1071-1081. [doi: 10.3109/02699052.2015.1015613] [Medline: 25950264]
30. Sherer M, Davis LC, Sander AM, Nick TG, Luo C, Pastorek N, et al. Factors Associated with Word Memory Test Performance in Persons with Medically Documented Traumatic Brain Injury. Clin Neuropsychol 2015;29(4):522-541. [doi: 10.1080/13854046.2015.1052763] [Medline: 26063081]
31. Ringdahl EN, Becker ML, Hussey JE, Thaler NS, Vogel SJ, Cross C, et al. Executive Function Profiles in Pediatric Traumatic Brain Injury. Dev Neuropsychol 2019;44(2):172-188. [doi: 10.1080/87565641.2018.1557190] [Medline: 30590952]
32. Jaeggi SM, Studer-Luethi B, Buschkuehl M, Su Y, Jonides J, Perrig WJ. The relationship between n-back performance and matrix reasoning—implications for training and transfer. Intelligence 2010 Nov;38(6):625-635. [doi: 10.1016/j.intell.2010.09.001]
33. Gates NJ, Sachdev PS, Fiatarone Singh MA, Valenzuela M. Cognitive and memory training in adults at risk of dementia: A Systematic Review. BMC Geriatr 2011 Sep 25;11(1):55. [doi: 10.1186/1471-2318-11-55]
34. Cha Y, Kim H. Effect of computer-based cognitive rehabilitation (CBCR) for people with stroke: a systematic review and meta-analysis. NeuroRehabilitation 2013;32(2):359-368. [doi: 10.3233/NRE-130856] [Medline: 23535800]
35. Kueider AM, Parisi JM, Gross AL, Rebok GW. Computerized cognitive training with older adults: a systematic review. PLoS One 2012;7(7):e40588 [FREE Full text] [doi: 10.1371/journal.pone.0040588] [Medline: 22792378]
36. Thompson G, Foth D. Cognitive-Training Programs for Older Adults: What Are they and Can they Enhance Mental Fitness? Educational Gerontology 2005 Sep;31(8):603-626. [doi: 10.1080/03601270591003364]

37. Whitmer AJ, Gotlib IH. Switching and backward inhibition in major depressive disorder: the role of rumination. J Abnorm Psychol 2012 Aug;121(3):570-578. [doi: 10.1037/a0027474] [Medline: 22468767]
38. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognition Letters 2010 Jun;31(8):651-666. [doi: 10.1016/j.patrec.2009.09.011]
39. Han J, Kamber M, Pei J. Data Mining: Concepts And Techniques, Third Edition (the Morgan Kaufmann Series In Data Management Systems). United States of America: Morgan Kaufmann; 2019. URL: http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf [accessed 2020-04-03]
40. Flynt A, Daepp MIG. Diet-related chronic disease in the northeastern United States: a model-based clustering approach. Int J Health Geogr 2015 Sep 04;14:25 [FREE Full text] [doi: 10.1186/s12942-015-0017-5] [Medline: 26338084]
41. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. R J 2016 Aug;8(1):289-317 [FREE Full text] [Medline: 27818791]
42. Husson F, Le S, Pagès J. Exploratory Multivariate Analysis By Example Using R (chapman & Hall/crc Computer Science & Data Analysis). Boca Raton, Florida, United States of America: CRC Press; 2019.
43. Conrad DJ, Bailey BA. Multidimensional clinical phenotyping of an adult cystic fibrosis patient population. PLoS One 2015;10(3):e0122705 [FREE Full text] [doi: 10.1371/journal.pone.0122705] [Medline: 25822311]
44. Subirats L, Lopez-Blazquez R, Ceccaroni L, Gifre M, Miralles F, García-Rudolph A, et al. Monitoring and Prognosis System Based on the ICF for People with Traumatic Brain Injury. Int J Environ Res Public Health 2015 Aug 18;12(8):9832-9847 [FREE Full text] [doi: 10.3390/ijerph120809832] [Medline: 26295252]
45. R-core. stats v3.6.1. R-core R-core@R-project. URL: https://www.rdocumentation.org/packages/stats/versions/3.6.1 [accessed 2019-08-30]
46. Maechle M. cluster v2.1.0. Finding Groups in Data: Cluster Analysis. URL: https://www.rdocumentation.org/packages/cluster/versions/2.1.0 [accessed 2020-03-01]
47. Kassambara A. factoextra v1.0.5. Extract and Visualize the Results of Multivariate Data Analyses. URL: https://www.rdocumentation.org/packages/factoextra/versions/1.0.5 [accessed 2019-08-30]
48. Scrucca L. mclust v5.4.5. Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation. URL: https://www.rdocumentation.org/packages/mclust/versions/5.4.5 [accessed 2019-08-30]
49. Liaw A. randomForest v4.6-14. Breiman and Cutler's Random Forests for Classification and Regression. URL: https://www.rdocumentation.org/packages/randomForest/versions/4.6-14 [accessed 2019-08-30]
50. Husson F. FactoMineR v1.42. Multivariate Exploratory Data Analysis and Data Mining. URL: https://www.rdocumentation.org/packages/FactoMineR/versions/1.42 [accessed 2019-08-30]
51. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 1987 Nov;20:53-65. [doi: 10.1016/0377-0427(87)90125-7]
52. Halkidi M, Batistakis Y, Vazirgiannis M. On Clustering Validation Techniques. Journal of Intelligent Information Systems 2001;17:107-145. [doi: 10.1023/a:1012801612483]
53. Meilă M. Comparing clusterings—an information based distance. Journal of Multivariate Analysis 2007 May;98(5):873-895. [doi: 10.1016/j.jmva.2006.11.013]
54. Chimge N, Baniwal SK, Luo J, Coetzee S, Khalid O, Berman BP, et al. Opposing effects of Runx2 and estradiol on breast cancer cell proliferation: in vitro identification of reciprocally regulated gene signature related to clinical letrozole responsiveness. Clin Cancer Res 2012 Feb 01;18(3):901-911 [FREE Full text] [doi: 10.1158/1078-0432.CCR-11-1530] [Medline: 22147940]
55. Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I. An extensive comparative study of cluster validity indices. Pattern Recognition 2013 Jan;46(1):243-256. [doi: 10.1016/j.patcog.2012.07.021]
56. Hennig C. fpc v2.2-3. Flexible Procedures for Clustering. URL: https://www.rdocumentation.org/packages/fpc/versions/2.2-3 [accessed 2019-08-30]
57. Zumel N, Mount J. Practical Data Science With R. New York, United States of America: Manning Publications; 2014.
58. Assis CSD, Batista LDC, Wolosker N, Zerati AE, Silva RDCGE. Functional independence measure in patients with intermittent claudication. Rev Esc Enferm USP 2015 Oct;49(5):756-761 [FREE Full text] [doi: 10.1590/S0080-623420150000500007] [Medline: 26516744]
59. García-Rudolph A, Gibert K. Understanding effects of cognitive rehabilitation under a knowledge discovery approach. Engineering Applications of Artificial Intelligence 2016 Oct;55:165-185. [doi: 10.1016/j.engappai.2016.06.007]
60. Wei T. corrplot v0.84. Visualization of a Correlation Matrix. URL: https://www.rdocumentation.org/packages/corrplot/versions/0.84 [accessed 2019-08-30]
61. Solana J, Cáceres C, García-Molina A, Chausa P, Opisso E, Roig-Rovira T, et al. Intelligent Therapy Assistant (ITA) for cognitive rehabilitation in patients with acquired brain injury. BMC Med Inform Decis Mak 2014 Jul 19;14:58 [FREE Full text] [doi: 10.1186/1472-6947-14-58] [Medline: 25038823]
62. Messaris P, Humphreys L. Digital media: Transformations in human communication. Berlin: Peter Lang, 2006; 2018:8147.

63. García-Rudolph A, Gibert K. A data mining approach to identify cognitive NeuroRehabilitation Range in Traumatic Brain Injury patients. Expert Systems with Applications 2014 Sep;41(11):5238-5251. [doi: 10.1016/j.eswa.2014.03.001]
64. Hammond F, Corrigan J, Ketchum JM, Malec JF, Dams-O Connor K, Hart T, et al. Prevalence of Medical and Psychiatric Comorbidities Following Traumatic Brain Injury. J Head Trauma Rehabil 2019;34(4):E1-E10. [doi: 10.1097/HTR.0000000000000465] [Medline: 30608311]

Abbreviations

AGNES: AGglomerative NESting
CLARA: Clustering LARge Applications
CVI: cluster validity index
DIANA: DIvisive ANAlysis
FIM: functional independence measure
GCS: Glasgow Coma Scale
GNPT: Guttman, NeuroPersonalTrainer
HCPC: hierarchical clustering on principal components
ICF: International Classification of Functioning, Disability and Health
ITA: intelligent therapy assistant
PAM: Partitioning Around Medoids
PCA: principal component analysis
TBI: traumatic brain injury

Edited by G Eysenbach; submitted 31.08.19; peer-reviewed by S Ge, J Salisbury, B Smith, JM Cogollor; comments to author 28.09.19; revised version received 26.01.20; accepted 14.05.20; published 06.10.20.

Please cite as:
Garcia-Rudolph A, Garcia-Molina A, Opisso E, Tormos Muñoz J
Personalized Web-Based Cognitive Rehabilitation Treatments for Patients with Traumatic Brain Injury: Cluster Analysis
JMIR Med Inform 2020;8(10):e16077
URL: https://medinform.jmir.org/2020/10/e16077
doi:10.2196/16077
PMID:33021482

Original Paper

How High-Risk Comorbidities Co-Occur in Readmitted Patients With Hip Fracture: Big Data Visual Analytical Approach

Suresh K Bhavnani[1,2], PhD; Bryant Dang[2], BSc; Rebekah Penton[3], DNP, RN, AGPCNP-BC; Shyam Visweswaran[4], MD, PhD; Kevin E Bassler[5], PhD; Tianlong Chen[2], PhD; Mukaila Raji[6], FACP, MS, MD; Rohit Divekar[7], MBBS, PhD; Raed Zuhour[8], MD; Amol Karmarkar[9], PhD; Yong-Fang Kuo[1], PhD; Kenneth J Ottenbacher[9], PhD

[1]Preventive Medicine and Population Health, University of Texas Medical Branch, Galveston, TX, United States

[2]Institute for Translational Sciences, University of Texas Medical Branch, Galveston, TX, United States

[3]School of Nursing, University of Texas Medical Branch, Galveston, TX, United States

[4]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, United States

[5]Department of Physics, University of Houston, Houston, TX, United States

[6]Division of Geriatric Medicine, Department of Internal Medicine, University of Texas Medical Branch, Galveston, TX, United States

[7]Division of Allergic Diseases, Mayo Clinic, Rochester, MN, United States

[8]Radiation Oncology, University of Texas Medical Branch, Galveston, TX, United States

[9]Department of Rehabilitation Sciences, University of Texas Medical Branch, Galveston, TX, United States

Corresponding Author:
Suresh K Bhavnani, PhD
Preventive Medicine and Population Health
University of Texas Medical Branch
301 University Blvd
Galveston, TX, 77555-0129
United States
Phone: 1 409 772 1928
Email: subhavna@utmb.edu

Abstract

Background: When older adult patients with hip fracture (HFx) have unplanned hospital readmissions within 30 days of discharge, it doubles their 1-year mortality, resulting in substantial personal and financial burdens. Although such unplanned readmissions are predominantly caused by reasons not related to HFx surgery, few studies have focused on how pre-existing high-risk comorbidities co-occur within and across subgroups of patients with HFx.

Objective: This study aims to use a combination of supervised and unsupervised visual analytical methods to (1) obtain an integrated understanding of comorbidity risk, comorbidity co-occurrence, and patient subgroups, and (2) enable a team of clinical and methodological stakeholders to infer the processes that precipitate unplanned hospital readmission, with the goal of designing targeted interventions.

Methods: We extracted a training data set consisting of 16,886 patients (8443 readmitted patients with HFx and 8443 matched controls) and a replication data set consisting of 16,222 patients (8111 readmitted patients with HFx and 8111 matched controls) from the 2010 and 2009 Medicare database, respectively. The analyses consisted of a supervised combinatorial analysis to identify and replicate combinations of comorbidities that conferred significant risk for readmission, an unsupervised bipartite network analysis to identify and replicate how high-risk comorbidity combinations co-occur across readmitted patients with HFx, and an integrated visualization and analysis of comorbidity risk, comorbidity co-occurrence, and patient subgroups to enable clinician stakeholders to infer the processes that precipitate readmission in patient subgroups and to propose targeted interventions.

Results: The analyses helped to identify (1) 11 comorbidity combinations that conferred significantly higher risk (ranging from $P<.001$ to $P=.01$) for a 30-day readmission, (2) 7 biclusters of patients and comorbidities with a significant bicluster modularity ($P<.001$; Medicare=0.440; random mean 0.383 [0.002]), indicating strong heterogeneity in the comorbidity profiles of readmitted patients, and (3) inter- and intracluster risk associations, which enabled clinician stakeholders to infer the processes involved in the exacerbation of specific combinations of comorbidities leading to readmission in patient subgroups.

Conclusions: The integrated analysis of risk, co-occurrence, and patient subgroups enabled the inference of processes that precipitate readmission, leading to a comorbidity exacerbation risk model for readmission after HFx. These results have direct

implications for (1) the management of comorbidities targeted at high-risk subgroups of patients with the goal of pre-emptively reducing their risk of readmission and (2) the development of more accurate risk prediction models that incorporate information about patient subgroups.

(JMIR Med Inform 2020;8(10):e13567) doi:10.2196/13567

KEYWORDS

unplanned hospital readmission; visual analytics; bipartite networks; precision medicine

Introduction

Background

Although it is well known that hip fractures (HFx) in older adults are a leading cause of morbidity, long-term functional impairment, and mortality [1], these outcomes are exacerbated when such patients are readmitted to the hospital within 30 days of hospital discharge after surgery, in addition to doubling their risk of 1-year mortality [2].

While many readmissions are unavoidable, unplanned hospital readmissions can easily negate the functional gains painstakingly achieved through weeks of post–acute rehabilitation and can increase the risk of infections acquired during hospital stays [3]. This loss is over and above the costs to caregivers and relatives who have to relive the stress of the original HFx episode, reorganize their work schedules to care for the patient, resulting in loss of productivity, and restart rehabilitation after discharge [3]. Across all conditions, unplanned readmissions cost almost US $17 billion annually in the United States [4], making them an ineffective use of costly resources and therefore closely scrutinized as a marker for poor quality by the Centers for Medicare & Medicaid Services (CMS) [5]. Consequently, the CMS instituted the Hospital Readmissions Reduction Program (HRRP) [6], which has imparted penalties on hospitals if their 30-day readmission rates exceeded the national average.

Although such incentives initially appeared to improve the readmission rates in US hospitals [7], recent reports argue that the start of the HRRP coincided with an increase in mortality among older adults [6,8]. This could have been because, as hospitals tightened their policies for readmission, many older adult patients were denied care, resulting in increased mortality. Furthermore, the decrease in readmission rates might merely reflect changes in the administrative and billing practices rather than an improvement in care [9]. These results suggest a need for more targeted research to comprehend the processes that precipitate readmission and clinical interventions that address the underlying causes of hospital readmission.

Methods Used to Analyze the Risk of Pre-Existing Comorbidities in Hospital Readmission

As hospital readmissions in the older adult HFx population are predominantly for reasons not related to the HFx surgery [10], several studies have focused on using supervised machine learning methods to determine how pre-existing comorbidities (defined as one or more conditions or diseases co-occurring with a primary condition such as HFx) increased the risk of readmission [2,10-14]. Most of these studies have focused on using logistic regression to analyze the risk of readmission of single comorbidities. For example, a recent study using Medicare data conducted for the CMS, analyzed patients with total hip or total knee arthroplasty to construct a logistic regression model with variables including 29 comorbidities to predict readmission [14]. Although the above descriptive and predictive approaches have provided important insights into the role of comorbidities in the readmission of patients with HFx, such studies do not focus on understanding how multiple comorbidities *co-occur* within and across *subgroups* of patients, a critical step in the design of targeted interventions to reduce readmissions.

Although the *co-occurrence* of pre-existing comorbidities has not yet been analyzed in readmitted patients with HFx, it has been analyzed in other index conditions [15-18], such as chronic obstructive pulmonary disease (COPD), and in patient populations, such as in older adults [19-21]. Such studies have focused on using unsupervised machine learning methods such as clustering (eg, hierarchical and partitioning clustering), dimensional reduction methods (eg, principal component analysis), and visual analytics (eg, network visualization and analysis). These include a recent questionnaire-based study of senior Australians that compared several unsupervised clustering methods to analyze patterns of multimorbidities (2 or more co-occurring conditions or diseases irrespective of an index condition) in the population [20]. The results found frequent co-occurrences, such as high blood pressure and diabetes, across the study population. Another study used unipartite networks (where nodes represented comorbidities, and edges between pairs of comorbidities represented the frequency of co-occurrence in patients) to identify clusters of frequently co-occurring comorbidities [21].

Although these studies have revealed the feasibility and appropriateness of using unsupervised methods to analyze the co-occurrences of comorbidities, they have typically focused on a unipartite analysis (clustering of only comorbidities) of the data and therefore cannot reveal complex patterns of patient heterogeneity hidden within those co-occurrences. Furthermore, such analyses cannot reveal the nature and degree of overlap among such subgroups. Understanding the complexities in such overlapping patient subgroups and their risk for readmission has direct relevance to clinician stakeholders in inferring the underlying processes involved in precipitating readmission and for the design of targeted interventions to reduce the risk of readmission.

Therefore, we explored an approach that integrates a supervised combinatorial method with an unsupervised bipartite network to address 3 questions: (1) Which combinations of comorbidities confer high risk for readmission in patients with HFx? (2) How do high-risk comorbidities co-occur within and across subgroups of readmitted patients with HFx? (3) What is the association

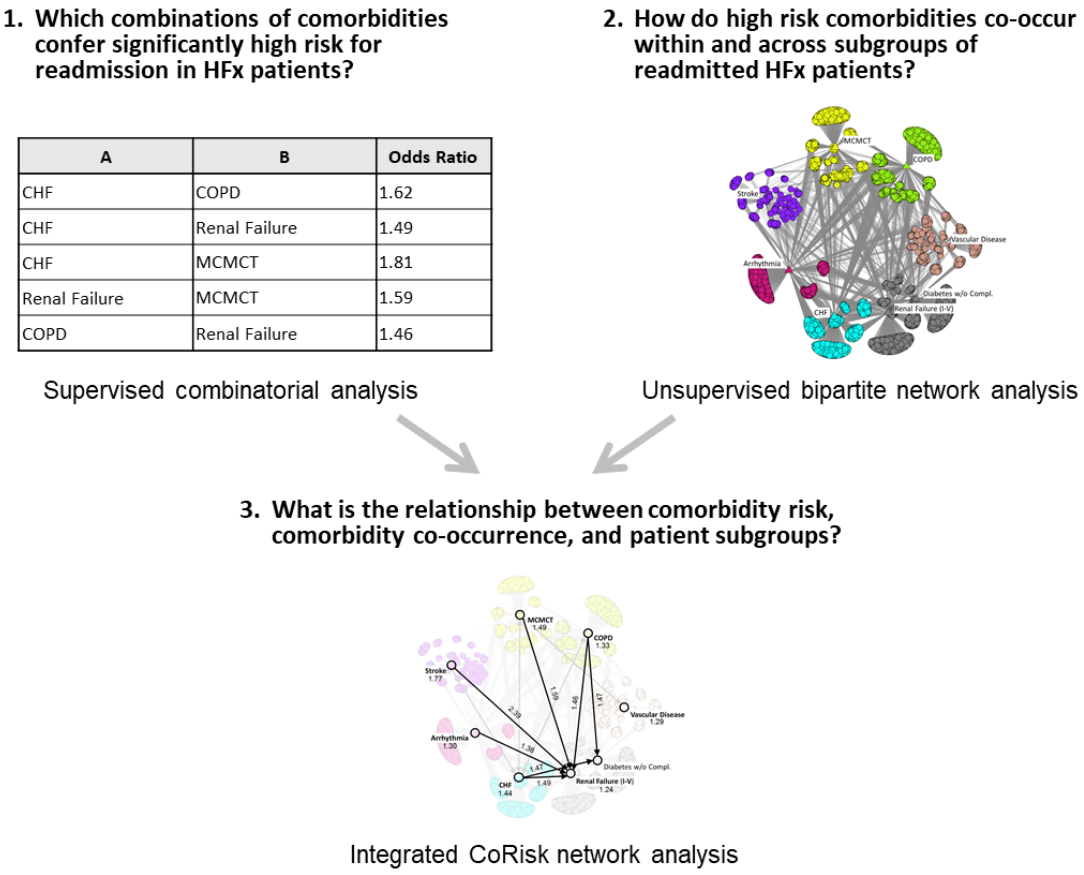
between comorbidity risk, comorbidity co-occurrence, and patient subgroups?

Methods

Overview

As shown in Figure 1, we addressed our 3 research questions by using a supervised machine learning method to address the first question, an unsupervised visual analytical method to address the second question, and an integrated visualization of both results to address the third question. Our goal was to analyze which combinations of comorbidities confer high risk for readmission and how those high-risk comorbidities co-occur within and across patient subgroups. This integrated visual analytical approach was designed to explicitly enable clinician stakeholders using a team-centered informatics [22] approach to comprehend the complex association of comorbidity risk, comorbidity co-occurrence, and patient subgroups, with the goal of designing targeted interventions, a cornerstone of precision medicine.

Figure 1. Overview of the analytical method based on 3 research questions. The steps and data shown are schematic to illustrate the overall approach and are elaborated on in the analytical method section.



Data Selection

Our data consisted of a *training data set* extracted from the 2010 inpatient Medicare claims data (the most current Medicare data set to which we had access) and a *replication data set* extracted from the 2009 inpatient Medicare claims data (the next most current dataset). In 2010, Medicare provided health insurance to approximately 48 million Americans, of which 40 million were older adults (≥65 years), representing 93% of all older adult Americans. Furthermore, the eligible claims were from 6204 medical institutions from across the United States, thereby confirming this to be one of the few data sets that are highly representative of the US older adult population and its care.

As is commonly done in analytical studies of claims data, we used Medicare Severity-Diagnosis Related Group (MS-DRG) codes to define our population. The MS-DRG codes are used by physicians to categorize Medicare beneficiaries into payment groups for the purposes of billing. We operationally defined patients with HFx as those who were discharged from an acute care hospital with the MS-DRG codes 480, 481, or 482. To isolate the association of pre-existing comorbidities with the risk of readmission and to maintain homogeneity of our study population, we included only patients without hospital complications. Furthermore, we included only patients who were enrolled in Medicare part A but not in a health maintenance organization (a type of health insurance that limits coverage to care from contracted doctors) during the period of 90 days after discharge, in addition to patients who survived 90 days after hospital discharge.

For both the training and replication data sets, we extracted (1) the data of all patients with HFx without hospital-acquired complications who were readmitted within 30 days of discharge and (2) an equal number of controls matched for age, gender, and race, who were not readmitted within 90 days of discharge. This 90-day window of no readmittance represents an episode of care proposed by CMS for patients with HFx [23], indicating that the controls are substantially free from complications that result in readmission during this period, thereby allowing an effective comparison with the cases.

The above inclusion and exclusion criteria for patients resulted in a training data set consisting of 16,886 patients (8443 cases and 8443 controls), and a replication data set consisting of 16,222 patients (8111 cases and 8111 controls) for a total of 33,108 patients with HFx (Multimedia Appendix 1). For each of the above patients, we extracted their status on 70 high-level comorbidities (Multimedia Appendix 2) as defined by hierarchical condition categories (HCCs) [24], which represent the range of conditions typically encountered in older adults. As our index condition was HFx, we excluded it as a comorbidity, resulting in the status of 69 HCC comorbidities across 33,108 patients with HFx in the Medicare database. This retrospective study was approved by the Institutional Review Board of the University of Texas Medical Branch. The Medicare data files used for the study were in the research identifiable format, and the records were anonymized and deidentified before the analysis. Therefore, the analysis of the data did not require informed consent. Furthermore, a data use agreement was completed, which met all CMS privacy and confidentiality requirements.

Analytical Methods Based on Research Questions

Which Combinations of Comorbidities Confer High Risk for Readmission in Patients With HFx?

To address this research question, we used a supervised combinatorial method to identify and replicate comorbidities that conferred high risk for readmission. Combinatorial methods have been used to analyze the prevalence of comorbidity combinations [19] and the risk of developing multimorbidities [25]. Here, we used the latter approach to identify which combinations of comorbidities confer significant risk for readmission. This analysis was performed first to base all subsequent analyses on only those comorbidities that were significant and replicated in another year.

We identified high-risk comorbidities in the training data set consisting of 16,886 patients (8443 cases and 8443 controls) by first removing all cases and controls that had none of the 69 comorbidities, resulting in 13,644 patients. Furthermore, similar to other studies on comorbidities [20], we removed 32 low-prevalence comorbidities that together occurred in less than 1% of the remaining patients (Multimedia Appendix 3), resulting in 13,512 patients in the training data set.

Next, we calculated the risk of remaining comorbidities across patients. As the patients had a median of 2 comorbidities in the HCC list, we measured the risk of all pairs of comorbidities using 2 tests. First, we used a *pairwise overall test* that measured the odds ratio (OR) of each pair of the 69 comorbidities compared with the rest of the patients and reported 95% confidence intervals. Second, we selected those pairs that were significant at $P<.05$ after correcting for multiple testing using the false discovery rate (FDR) method [26]. For each of the above comorbid pairs that were significant, we used a *pairwise directionality test* to determine the direction of their risk. Here, we conducted 2 tests: (1) A and B versus A and (2) A and B versus B, where A and B represent the sets of patients with comorbidities A and B, respectively. Within each test, we used the FDR to correct for multiple testing and considered $P<.05$ after adjustment to be significant.

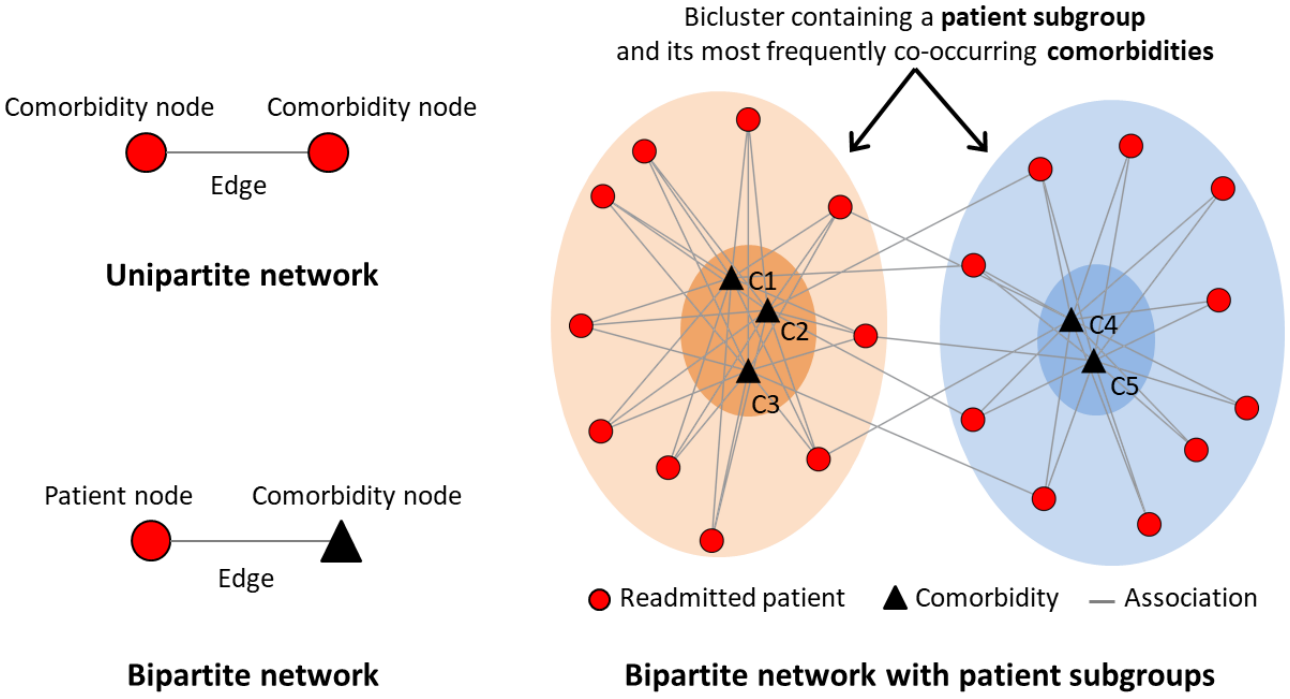
To test for replication of the significance and direction of the comorbidity pairs, we repeated the above analyses using the replication data set. As the patients in the test data also had a median of 2 comorbidities, we analyzed which of the significant pairs in the training data were also significant and had the same risk direction in the replication data set. Significant comorbidity pairs that had an identical direction of risk in each data set were selected for subsequent analyses. All tests of statistical significance were two-sided, and the analyses were performed using R version 3.6.1 (R Foundation for Statistical Computing; Multimedia Appendix 4).

How do High-Risk Comorbidities Co-Occur Within and Across Subgroups of Readmitted Patients With HFx?

To analyze how the above significant and replicated pairs co-occurred in readmitted patients with HFx, we used unsupervised bipartite networks. As shown in Figure 2, a network consists of nodes and edges; nodes represent one or more types of entities (eg, patients or comorbidities), and edges between the nodes represent a specific relationship between the entities. As shown in the upper left-hand part of Figure 2, a unipartite network has nodes that are of the same type (typically used to analyze co-occurrence of comorbidities [21]). In contrast, as shown in the lower left-hand part of Figure 2, a bipartite network has nodes that are of 2 types, and edges exist only between the 2 types, such as between patients (circles) and comorbidities (triangles). This quantitative and visual representation, which integrates patients and their comorbidities in a single representation, enables stakeholders to infer the mechanisms in each patient subgroup, a corner stone of precision medicine.

To analyze the data, we used the following steps: (1) represented the data as a bipartite network where nodes represented either patients or comorbidities, and the edges represented the presence or absence of a comorbidity; (2) identified patient subgroups and their most frequently co-occurring comorbidities using bicluster modularity [27,28] and tested its significance through comparisons with 1000 random permutations of the data; (3) used the Rand index (RI) [29] to measure the similarity of comorbidity co-occurrence between the training and replication data sets, and tested the RI significance; and (4) used the *ExplodeLayout* algorithm [30] to separate the biclusters, with the goal of reducing the visual overlap among them, thereby enhancing their comprehensibility.

Figure 2. The distinction between a unipartite network, a bipartite network, and how the latter can be used to identify biclusters of patients and comorbidities.



What is the Relationship Between Comorbidity Risk, Comorbidity Co-Occurrence, and Patient Subgroups?

CoRisk Network Analysis: Integration of Risk, Co-Occurrence, and Patient Subgroups

The results from the supervised risk analysis and the unsupervised bipartite network analysis were integrated into a single network visualization. This was achieved by representing the high-risk and replicated pairs and their direction (identified in research question 1) using a directed unipartite network, where nodes represented the comorbidities and directed edges represented the direction of that risk. This unipartite network was superimposed onto the bipartite network of readmitted patients with HFx and comorbidities (described in research question 2) resulting in a *co-occurrence risk* (CoRisk) network. We define this CoRisk network visualization as the merging of 2 networks: (1) a bipartite network consisting of nodes representing patients and comorbidities, with edges representing their pairwise relationship and (2) a comorbidity risk network consisting of weighted directed edges between the comorbidities representing the risk and direction of significant and replicated comorbidity pairs. This integration of the supervised and unsupervised analytical results was designed to enable clinician stakeholders to interpret the relationship between high-risk pairs of comorbidities, their co-occurrence, and patient subgroups.

Clinical Interpretation of CoRisk Network

The CoRisk network was presented to a stakeholder team specializing in geriatrics and hospital re-admission, in addition to a biostatistician, who together examined the clinical meaningfulness of the risk, co-occurrence, and patient subgroups. The stakeholders were asked to visually analyze the CoRisk network and use their domain knowledge to (1) infer the underlying process that precipitated re-admission and (2) provide corroborative evidence from published literature to support their inferences.

Results

In this section, we present the results of our analysis based on the 3 research questions:

Which Combinations of Comorbidities Confer High Risk for Readmission in Patients With HFx?

As shown in Table 1, the *pairwise overall test* identified 24 pairs (all rows shown in the table) that were significant in the training data set. Furthermore, the *pairwise directionality test* identified 10 pairs that were significant in both directions, 13 that were significant only in one direction, and 1 that was not significant in either direction (for clarity, only significant results are shown for the pairwise directionality test in Table 1).

Table 1. The 24 comorbidity pairs that had significantly higher risk for readmission in the training data set, of which 11 replicated (serial number pairs 1-11) in the test data by being significant in the same direction.

Serial number	Comorbidity pair		Pairwise overall test		Pairwise directionality test			
	A	B	(A&B) vs (A + B + not A & not B)		(A&B) vs A		(A&B) vs B	
			OR[a] (95% CI)	False discovery rate *P* valu	OR (95% CI)	False discovery rate *P* value	OR (95% CI)	False discovery rate *P* value
1	CHF[b]	COPD[c]	1.62 (1.40-1.89)	<.001	1.24 (1.05-1.47)	.019	1.36 (1.15-1.61)	.004
2	CHF	MCMCT[d]	1.81 (1.46-2.25)	<.001	1.38 (1.10-1.73)	.01	1.33 (1.05-1.69)	.03
3	CHF	Renal failure (I-V)	1.49 (1.31-1.70)	<.001	NS[e]	NS	1.34 (1.15-1.55)	.003
4	CHF	Stroke	1.99 (1.40-2.83)	.005	1.50 (1.05-2.14)	.04	NS	NS
5	Diabetes (without complications)	CHF	1.47 (1.25-1.73)	<.001	1.53 (1.28-1.83)	.000	NS	NS
6	Arrhythmias	Renal failure (I-V)	1.38 (1.20-1.60)	.001	NS	NS	1.20 (1.02-1.41)	.04
7	RF(I-V)	MCMCT	1.59 (1.28-1.96)	.001	1.37 (1.10-1.71)	.01	NS	NS
8	COPD	Renal failure (I-V)	1.46 (1.22-1.74)	.002	NS	NS	1.26 (1.04-1.52)	.03
9	Diabetes (without complications)	COPD	1.47 (1.21-1.77)	.004	1.49 (1.22-1.83)	.001	NS	NS
10	Stroke	Renal failure (I-V)	2.39 (1.55-3.69)	.004	NS	NS	2.04 (1.32-3.17)	.005
11	Vascular disease	MCMCT	2.03 (1.39-2.98)	.009	1.70 (1.13-2.53)	.02	NS	NS
12	CHF	Arrhythmias	1.46 (1.30-1.64)	<.001	NS	NS	1.25 (1.09-1.43)	.005
13	Arrhythmias	COPD	1.62 (1.38-1.89)	<.001	1.36 (1.15-1.62)	.002	1.34 (1.12-1.60)	.005
14	Arrhythmias	Stroke	2.23 (1.63-3.04)	<.001	1.85 (1.35-2.54)	.001	NS	NS
15	Stroke	COPD	3.18 (1.87-5.41)	.001	2.03 (1.15-3.60)	.02	2.56 (1.50-4.36)	.004
16	Arrhythmias	Hemiple-gia/hemiparesis	2.18 (1.51-3.16)	.002	1.80 (1.24-2.62)	.006	2.12 (1.38-3.25)	.004

Serial number	Comorbidity pair		Pairwise overall test		Pairwise directionality test			
	A	B	(A&B) vs (A + B + not A & not B)		(A&B) vs A		(A&B) vs B	
			OR[a] (95% CI)	False discovery rate *P* valu	OR (95% CI)	False discovery rate *P* value	OR (95% CI)	False discovery rate *P* value
17	Angina	Arrhythmias	1.85 (1.38-2.49)	.003	1.92 (1.38-2.68)	.001	1.53 (1.13-2.07)	.02
18	CHF	Hemiple-gia/hemiparesis	2.25 (1.49-3.40)	.005	1.70 (1.12-2.57)	.02	2.11 (1.33-3.34)	.005
19	Cardio-respirato-ry failure	CHF	1.65 (1.28-2.13)	.005	1.70 (1.25-2.31)	.003	NS	NS
20	Vascular disease	Renal failure (I-V)	1.63 (1.27-2.10)	.005	1.39 (1.04-1.85)	.04	1.40 (1.08-1.81)	.02
21	COPD	MCMCT	1.58 (1.24-2.03)	.01	NS	NS	NS	NS
22	Septicemia/shock	Renal failure (I-V)	2.51 (1.49-4.22)	.02	2.85 (1.47-5.52)	.006	2.14 (1.27-3.61)	.01
23	Intestinal obstruc-tion	Arrhythmias	2.36 (1.45-3.84)	.02	2.17 (1.25-3.77)	.01	1.94 (1.18-3.17)	.02
24	Stroke	Hemiple-gia/hemiparesis	1.78 (1.26-2.53)	.04	NS	NS	1.63 (1.08-2.46)	.03

[a]OR: odds ratio.

[b]CHF: congestive heart failure.

[c]COPD: chronic obstructive pulmonary disease.

[c]MCMCT: major complications of medical care and trauma.

[d]NS: not significant.

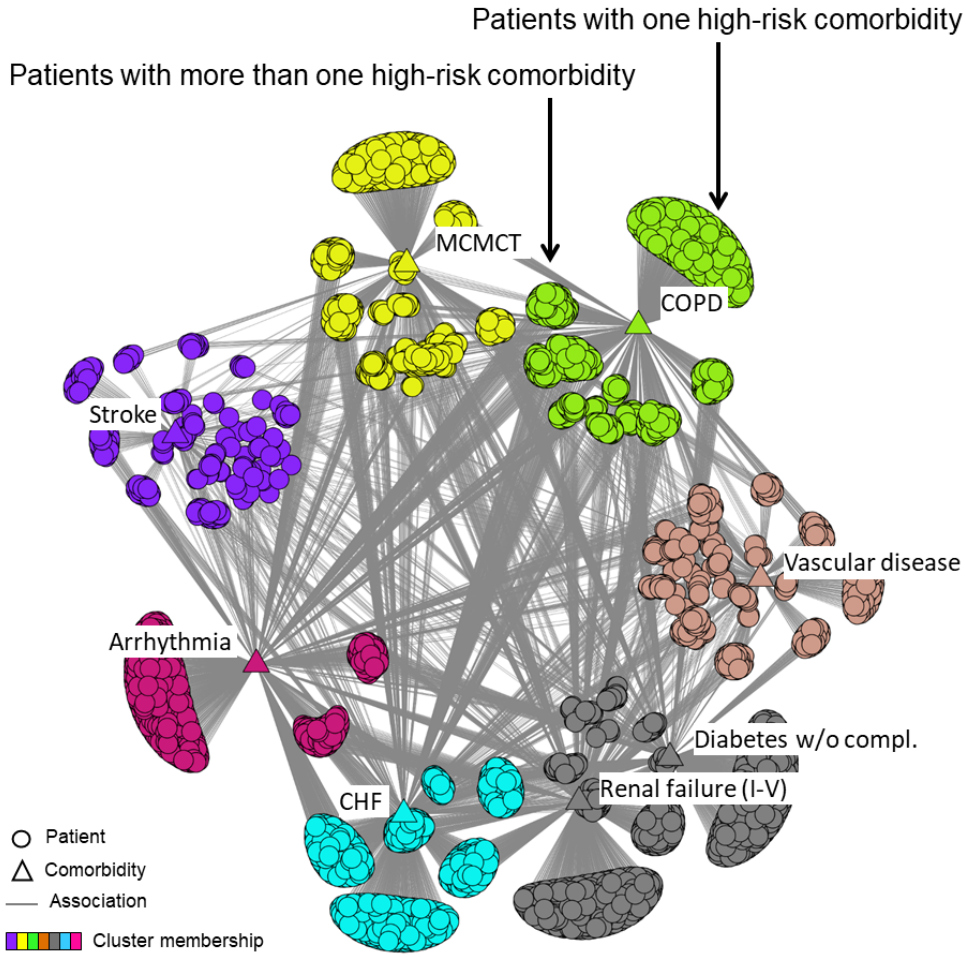
Next, we identified which of the above 24 pairs replicated by identifying comorbidity pairs that were identical in their significance and direction in the replication data set. As shown in Table 1, of the 24 pairs, 11 pairs (highlighted in blue and pink) replicated in the replication data set. Of these, 2 pairs (serial number pairs 1-2) were significant in both directions, and 9 pairs (serial number pairs 3-11) were significant only in one direction. The overlapping pairs resulted in 8 unique comorbidities: congestive heart failure (CHF), COPD, major complications of medical care and trauma (MCMCT), RF I-V, stroke, diabetes without complications (diabetes), arrhythmias, and vascular disease.

How Do High-Risk Comorbidities Co-Occur Within and Across Subgroups of Readmitted Patients With HFx?

Visualization

To comprehend how high-risk comorbidities co-occurred across patients, we conducted a bipartite network analysis. The nodes consisted of the 8 significant and replicated comorbidities implicated in risk for readmission from the above combinatorial analysis, and all readmitted patients with HFx with at least one of those comorbidities (n=6150). As shown in Figure 3, the bipartite network analysis revealed 7 biclusters of patients and high-risk comorbidities.

Figure 3. Bipartite network of significant and replicated comorbidities and re-admitted patients with HFx from the training data set.



Quantitative Verification and Layout Refinement

The network had a modularity of 0.440, which was significant (P<.001; Medicare=0.440; random mean 0.383 [0.002]) compared to 1000 random permutations of the network while preserving the network size (number of nodes) and network density (number of edges). The corresponding network generated from the replication data set also had 7 biclusters, a modularity of 0.444, which was also highly significant (P<.001; Medicare=0.444; random mean 0.379 [0.002]) compared to 1000 random permutations of the data while preserving network size and density.

Replication of Modularity and Comorbidity Co-Occurrence

The co-occurrence of comorbidities within and across clusters (as measured by the RI) between the training and replication data sets was significant (P=.02; Medicare=0.929; random mean 0.869 [0.027]), indicating a strong similar and significant co-occurrence pattern of comorbidities in the 2 networks. The training and test bipartite networks were therefore strongly biclustered (as measured by the similarly high biclustered modularity), highly significant (as measured by the permutation test), had a similar pattern of co-occurrence (as measured by the RI and its significance), and had the same number (7) of biclusters.

Although the above quantitative analysis revealed a significant and replicated overall clustered topology, a visual analysis of the network revealed 2 important patterns related to comorbidity co-occurrence and heterogeneity within patient subgroups:

Comorbidity Co-Occurrence

As shown in Figure 3, 6 comorbidities belonged to single-comorbidity biclusters, whereas 2 comorbidities co-occurred in the same cluster. This indicates that although many patients in one bicluster had comorbidities in another (as shown by the many edges between the clusters), the bicluster overlap in most cases was not strong enough to pull comorbidities into the same bicluster. One exception was the bicluster with RF and diabetes, where there were many patients with both, resulting in them being pulled together into the same bicluster.

Heterogeneity Within Patient Subgroups

As shown in Figure 3, each bicluster had a set of patients with only one comorbidity (in the outer side of the bicluster), and another set of patients with more than one comorbidity (in the inner side of the bicluster), revealing an additional level of heterogeneity within each bicluster. As shown in Table 2, the biclusters had different proportions of one or more comorbidities. For example, only 30% of patients in the arrhythmia bicluster had more than one high-risk comorbidity compared with 78% of patients in the vascular disease cluster. This bicluster-specific heterogeneity, as measured by the ratio

of patients with one to many comorbidities, was significantly different across the 7 biclusters (X^2_6=868.6; N=6150; P<.001).

The bipartite network analysis therefore not only revealed how the comorbidities co-occurred across patient subgroups but also the patient heterogeneity at the network-wide level and at the bicluster-specific level, revealing the real-world variations in the comorbidity profiles of patients with HFx.

Table 2. The number of patients with one or more comorbidities across the 7 biclusters (patients with one comorbidity in the RF and diabetes bicluster had either RF or diabetes).

Number of comorbidities	CHF[a], n (%)	Arrhythmia, n (%)	Stroke, n (%)	MCMCT[b], n (%)	COPD[c], n (%)	Vascular disease, n (%)	Renal failure and diabetes, n (%)	Total, n (%)
Comorbidities=1	536 (50)	545 (69.8)	37 (13.7)	337 (39.4)	510 (41.3)	114 (21.7)	1062 (75.32)	3141 (51.07)
Comorbidities>1	536 (50)	236 (30.2)	233 (86.3)	518 (60.6)	726 (58.7)	412 (78.3)	348 (24.7)	3009 (48.93)
Total	1072 (100)	781 (100)	270 (100)	855 (100)	1236 (100)	526 (100)	1410 (100)	6150 (100)

[a]CHF: congestive heart failure.

[b]MCMCT: major complications of medical care and trauma.

[c]COPD: chronic obstructive pulmonary disease.

What is the Relationship Between Comorbidity Risk, Comorbidity Co-Occurrence, and Patient Subgroups?

As shown in Figure 4, the CoRisk network revealed how the high-risk pairs were (1) related to each other, (2) their directionality, and (3) how they were related to the patient subgroups. This integrated network enabled stakeholders to identify 2 sets of comorbidities. The first set (diabetes and RF) consisted of comorbidities that can have multi-organ consequences and is therefore referred to as systemic diseases. In contrast, the second set (CHF, arrhythmia, stroke, MCMCT, COPD, and vascular disease) consisted of comorbidities that had mainly organ-specific consequences. For example, while cardiac arrhythmia could potentially have systemic consequences, this comorbidity is specific to the electrophysiological properties of the heart.

As the clinician stakeholders were most interested in the interrelationship of risk between multi-organ and organ-specific comorbidities, we bolded all the edges that started from an organ-specific comorbidity (CHF, arrhythmia, stroke, MCMCT, and COPD) and ended at a multi-organ comorbidity (RF or diabetes). As shown in Figure 4, all the remaining edges pointed toward RF and diabetes, forming an *asymmetrical hub* (more edges pointing in than those pointing out). This meant that the implicated pairs connecting the nodes had significantly higher risk compared with RF and diabetes alone, but not significantly higher risk compared with the other members of the pair. For example, the directed edge starting from COPD and pointing to diabetes indicated that patients with COPD and diabetes have a significantly higher risk compared with diabetes alone, but not a significantly higher risk compared with COPD alone.

The asymmetrical risk hubs of RF and diabetes suggested that because they have multi-organ consequences, their outcomes are largely chronic and therefore require considerable severity on their own before they become the sole risk factors for readmission (note that the HCC definition of RF has a wide range from Stage I to V, possibly resulting in several patients with RF being in the early stages). However, when they co-occur with an organ-specific disease such as CHF, arrhythmia, stroke, MCMCT, or COPD, it can exacerbate those pre-existing conditions leading to a significantly higher risk of readmission. This pattern of asymmetrical risks resulted in the following hypothesis for a 2-tiered comorbidity exacerbation risk model with significantly higher risk at each subsequent tier:

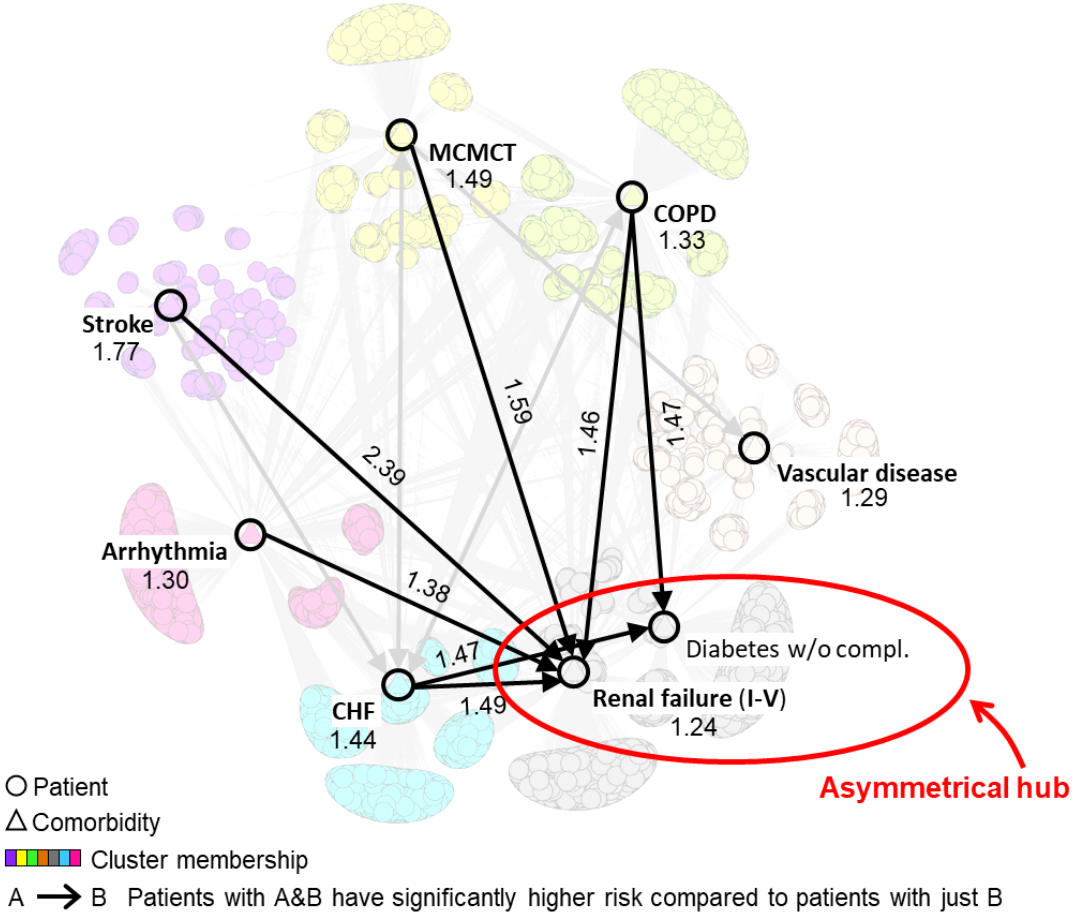
1. Tier 1 risk (only multi-organ comorbidities): RF, diabetes. This tier consists of patients in the RF-diabetes cluster.
2. Tier 2 risk (multi-organ plus organ-specific comorbidities): RF plus CHF or arrhythmia or stroke or MCMCT or COPD (patients in the inner part of the biclusters) in addition to patients with CHF, arrhythmia, stroke, MCMCT, or COPD (patients in the outer part of the respective biclusters).

Combining the above risk model with their own domain experience, the physician and the occupational therapist on the stakeholder team inferred hypotheses for the processes precipitating readmission in patients with HFx and provided corroborative evidence from the literature to support their inferences. They noted that when a patient is discharged from a hospital after an HFx surgery, the standard-of-care in generating discharge notes and order sets is focused on wound healing, postoperative delirium, mobilization, rehabilitation, and nutritional needs [31,32]. However, despite these guidelines, older adult patients, particularly those in skilled nursing facilities, regularly suffer from dehydration and malnutrition [33-37]. These conditions can worsen compromised renal function [38] as well as glycemic control in diabetics, ultimately triggering the deterioration of existing organ-specific comorbidities such as CHF and COPD [39-41]. Unfortunately, by the time symptoms of exacerbation in comorbidities are detected, the patient's health may have considerably declined, requiring urgent care, triggering an unplanned hospital readmission.

The nurse practitioner on the stakeholder team further stated that a contributing factor to the above cascade of events could be the lack of multidisciplinary care when patients with HFx are discharged after surgery. As stated in a recent review [42], HFx management "requires physicians to anticipate problems that may arise during recovery, whether the complications are from hip fracture and immobility, exacerbations of chronic

diseases, or problems with social and psychological support...it takes a team of dedicated professionals working together seamlessly to deliver care appropriate for patient goals, and to maximize recovery." In fact, although an increase in the number of registered nurses and multidisciplinary care teams has been associated with reduced 30-day readmission rates and improved health outcomes [43-45], such post–acute care is not yet widespread. The results of the analysis, combined with domain experience and corroborating evidence, enabled the clinician stakeholders to infer that the generation of discharge notes and order sets *before* discharge and the level of multidisciplinary care *after* discharge could be prime targets for reducing the risk of hospital readmission in specific subgroups of patients with HFx.

Figure 4. CoRisk network showing the integrated results from the supervised combinatorial analysis and the unsupervised bipartite network analysis. The numbers on the nodes refer to the odds ratios of comorbidities that were significantly associated to 30-day readmissions, the numbers on the edges refer to the ORs of pairs that were significant based on the pairwise overall test, and the direction of the edges represent the pairs that were significant and replicated in the same direction based on the pairwise directionality test. CHF: congestive heart failure; COPD: chronic obstructive pulmonary disease; w/o compl: without complications; MCMCT: major complications of medical care and trauma.



Discussion

Implications for Designing Targeted Interventions and Predictive Models

Our approach to integrate the results from supervised and unsupervised approaches into the CoRisk network helped to reveal (1) the overlap among the high-risk pairs, resulting in the stakeholders identifying the asymmetrical hub, and (2) the relationship of high-risk pairs to the network-wide and bicluster-specific patient heterogeneity. These results enabled the clinician stakeholders to infer hypotheses about the processes that precipitate readmission through a comorbidity exacerbation risk model. These results have the following implications for the design of interventions and predictive modeling.

Design of Postoperative Interventions

When a patient with HFx is discharged, the discharge notes and order sets could state which of the seven high-risk pre-existing comorbidities exist in the patient, with the respective recommendations for recognizing the early signs of the worsening of those comorbidities. For example, patients with RF should be monitored by rehabilitation or home health providers for urine output or weight gain, and those with diabetes should have more than usual monitoring of blood glucose during the convalescent period. However, patients with *both* RF and CHF should have their volume status more closely monitored, as they are more likely to develop an acute CHF exacerbation than patients with CHF alone. This is because patients with RF have a reduced ability to regulate the volume status and small fluctuations can precipitate acute CHF exacerbation, resulting in cardiorenal syndrome [46].

Furthermore, rehabilitation providers (physical therapists, physicians, registered nurses, and social workers) should be specifically trained to recognize and report changes in physical status, such as reduced oral intake, which might be an early warning of impending exacerbation of the specific comorbidities identified in the analysis. Finally, given the scarcity of rehab resources, clinicians could use the 2-tiered significant risk profile discussed above in triaging care, such as conducting more frequent evaluations of patients with HFx with COPD and RF compared with those with only RF. Future clinical trials could test whether improved discharge notes and order sets, in addition to early identification and treatment of worsening comorbidities through multidisciplinary team monitoring, can help to reduce the risk of readmission in patients with HFx.

Design of Preoperative Interventions

The results also suggest that combinations of high-risk comorbidities could be used to fine-tune the current criteria to select patients who should undergo HFx surgery. For example, certain comorbidity combinations could simply reflect the overall poor performance status of a patient, for whom postoperative interventions, no matter how robust, might end up being largely ineffective in preventing readmission. Future models could identify which subsets of patients have such *unmodifiable* readmission risks that outweigh the benefits of surgery and therefore could be better served with more conservative approaches.

Design of Predictive Models

The bipartite network analysis of patients and comorbidities showed significant and replicated heterogeneity among the readmitted patients based on their comorbidity profiles. However, current logistic regression models designed to predict readmission do not consider such heterogeneity in readmitted patients. For example, the regression model developed for CMS to predict readmission in arthroplasty or hip replacement patients [14] uses a single model to predict readmission for all patients. Although this model was an important advancement in predicting readmission in this population, it assumes that all patients can be modeled using a uniform set of coefficients for the same variables, an assumption that could conceal heterogeneity in readmitted patients and affect the accuracy of prediction in patient subgroups.

As stated by the biostatistician on the stakeholder team, a common approach to address such heterogeneity is to develop stratified regression models [47,48], one for each stratum of the population. The mathematical intuition underlying stratified regression models is that regression models can achieve a better fit to subsets of the data that are homogenous compared with a single regression model that is fitted to all of the data. For example, recognizing that races have different risks for developing type 2 diabetes, a recent study demonstrated that race-stratified regression models resulted in improved prediction accuracy for a racial subgroup [47]. However, such patient stratifications are typically selected based on an *a priori* understanding of the domain, which might miss important patterns in the data.

In contrast to the above approach of selecting patient subgroups, we believe our approach can enable the automatic identification of patient stratification that is data-driven and furthermore tested for significance and replicability, as we have demonstrated. Such information could then be used to develop stratified regression models to test whether they reveal heterogeneity in prediction accuracy for one or more patient subgroups. For example, stratified regression models could be developed and tested for each of the 7 clusters shown in Figure 3. Furthermore, given that each of the clusters had an outer subgroup (with only one high-risk comorbidity) and an inner subgroup (with more than one high-risk comorbidity), future regression models could also be targeted at each of these subgroups within biclusters, depending on their prevalence. Finally, each of the above regression models could test for interactions among the 11 high-risk comorbidity pairs shown in Table 1.

Improvements achieved through stratified regression models are dependent on a host of factors, including the degree of homogeneity in patient subgroups, the adequacy of sample size within those subgroups, and the tradeoff between prediction accuracy and model complexity. Future research should, therefore, determine whether stratified regression models based on automatically identified patient subgroups can produce more robust predictive models for hospital readmission.

Strengths and Limitations

The strength of this study is that we integrated the results from well-known methods with novel approaches, which together enabled a deeper understanding of the associations between risk, co-occurrence, and subgroups. This in turn led to insights related to targeted interventions (a critical goal of precision medicine), in addition to the design of predictive models. Furthermore, the analytical results were replicated in another year, demonstrating its generalizability. Critical to this process was the team-centered informatics approach [22] we pursued at each step of the project, which used intuitive visual analytical representations to span the disciplinary boundaries of clinicians and methodology stakeholders, enabling them to comprehend and address the complexity in a large data set.

A limitation of this study is that we tested the method on just one index condition, and our ongoing research [49] is testing the approach on other index conditions. Furthermore, the interpretability of the clusters could be enhanced by constructing additional figures wherein the patient nodes are colored based on covariates important to hospital readmission (eg, age, gender, race, length of hospital stay, and reason for readmission), in addition to determining which of them are significantly higher and lower across the clusters. Finally, the prevalence and severity of comorbidities may vary in patients receiving care in clinics, acute care hospitals, skilled nursing facilities, and nursing home settings. Therefore, future research should analyze whether the results vary across different care settings.

Fully cognizant that few data sets are without limitations, we consciously chose to analyze Medicare data because of its scale (enabling us to have adequate numbers of patients when analyzing patient heterogeneity), availability of data over multiple years (enabling us to test external replicability), and generalizability (enabling us to analyze data from patients and

hospitals across the United States). However, given that Medicare data are collected mainly for administrative purposes, it has well-known limitations, including the lack of test results, which could enable a finer understanding of the severity of comorbidities. Furthermore, comorbidities associated with mental health are known to be undercoded in the Medicare database, which could bias our results. Therefore, when clinical data across hospitals become available in future (eg, through the PCORnet [50] funded by the Patient-Centered Outcomes Research Institute [PCORI] and through the Accrual to Clinical Trials network [51] funded by the National Center for Advancing Translational Sciences [NCATS]), we intend to repeat our analysis using clinical data, but we fully realize that such data might have limitations that are yet unknown.

Methodologically, our approach of integrating supervised and unsupervised visual analytical approaches is just one of the many possible ways such integration can be achieved [52]. In our future research, we plan to explore other integration strategies with a specific focus on enabling clinician stakeholders to go beyond the analyses of prevalence and risk, enabling inferences for the underlying processes precipitating readmission. Such improvements in data and methods should enable discharge planners and providers in rehabilitation facilities to more accurately predict which patients will be readmitted and to select targeted interventions to reduce the risk of readmission and, consequently, the concomitant burden on patients and caregivers.

Acknowledgments

The authors thank James Goodwin, Allan Brasier, and Gautam Vallabha for their support and feedback. Funding for SB was provided in part by the Clinical and Translational Science Award (UL1 TR001439) from the NCATS, National Institutes of Health, the PCORI (ME-1511-33194), and the UTMB Claude D Pepper Older Americans Independence Center funded by NIA, National Institutes of Health (P30 AG024832); funding for SV was provided in part by the NLM, National Institutes of Health (R01 LM012095); funding for KB was provided in part by the National Science Foundation (DMR-1507371 and IOS-1546858); funding for MR and YK was provided in part by the NIDA, National Institutes of Health (R01 DA039192); funding for KO was provided in part by the NIA, National Institutes of Health (K07 AG064031), and by the CLDR, National Institutes of Health (P2CHD065702). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, PCORI, or NSF.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Number of patients based on the selection criteria.
[DOCX File , 47 KB - medinform_v8i10e13567_app1.docx]

Multimedia Appendix 2
The 70 HCC codes that represent the range of comorbidities typically encountered in the elderly.
[DOCX File , 18 KB - medinform_v8i10e13567_app2.docx]

Multimedia Appendix 3
Cumulative frequency of comorbidities in HFx patients. The analysis revealed that 32 comorbidities together accounted for <1% of the patients, and were removed from the data, with the remaining 37 comorbidities used for the subsequent analyses.
[PNG File , 32 KB - medinform_v8i10e13567_app3.PNG]

Multimedia Appendix 4
R (supported by the R Foundation for Statistical Computing) packages and functions used for the analyses.
[DOCX File , 13 KB - medinform_v8i10e13567_app4.docx]

References

1. Hip Fractures Among Older Adults. Centers for Disease Control and Prevention. URL: http://www.cdc.gov/homeandrecreationalsafety/falls/adulthipfx.html [accessed 2020-01-07]
2. French DD, Bass E, Bradham DD, Campbell RR, Rubenstein LZ. Rehospitalization after hip fracture: predictors and prognosis from a national veterans study. J Am Geriatr Soc 2008 Apr;56(4):705-710. [doi: 10.1111/j.1532-5415.2007.01479.x] [Medline: 18005354]
3. The Revolving Door: A Report on US Hospital Readmissions. The Robert Wood Johnson Foundation. 2013. URL: https://www.rwjf.org/en/library/research/2013/02/the-revolving-door--a-report-on-u-s--hospital-readmissions.html [accessed 2020-06-16]

4. Jencks SF, Williams MV, Coleman EA. Rehospitalizations among patients in the medicare fee-for-service program. N Engl J Med 2009 Apr 2;360(14):1418-1428. [doi: 10.1056/NEJMsa0803563] [Medline: 19339721]
5. Ashton CM, Del Junco DJ, Souchek J, Wray NP, Mansyur CL. The association between the quality of inpatient care and early readmission: a meta-analysis of the evidence. Med Care 1997 Oct;35(10):1044-1059. [doi: 10.1097/00005650-199710000-00006] [Medline: 9338530]
6. Fonarow GC, Konstam MA, Yancy CW. The hospital readmission reduction program is associated with fewer readmissions, more deaths: time to reconsider. J Am Coll Cardiol 2017 Oct 10;70(15):1931-1934 [FREE Full text] [doi: 10.1016/j.jacc.2017.08.046] [Medline: 28982507]
7. Wasfy JH, Zigler CM, Choirat C, Wang Y, Dominici F, Yeh RW. Readmission rates after passage of the hospital readmissions reduction program: a pre-post analysis. Ann Intern Med 2017 Mar 7;166(5):324-331 [FREE Full text] [doi: 10.7326/M16-0185] [Medline: 28024302]
8. Wadhera RK, Maddox KE, Wasfy JH, Haneuse S, Shen C, Yeh RW. Association of the hospital readmissions reduction program with mortality among medicare beneficiaries hospitalized for heart failure, acute myocardial infarction, and pneumonia. J Am Med Assoc 2018 Dec 25;320(24):2542-2552 [FREE Full text] [doi: 10.1001/jama.2018.19232] [Medline: 30575880]
9. Ody C, Msall L, Dafny LS, Grabowski DC, Cutler DM. Decreases in readmissions credited to medicare's program to reduce hospital readmissions have been overstated. Health Aff (Millwood) 2019 Jan;38(1):36-43. [doi: 10.1377/hlthaff.2018.05178] [Medline: 30615522]
10. Boockvar KS, Halm EA, Litke A, Silberzweig SB, McLaughlin M, Penrod JD, et al. Hospital readmissions after hospital discharge for hip fracture: surgical and nonsurgical causes and effect on outcomes. J Am Geriatr Soc 2003 Mar;51(3):399-403. [doi: 10.1046/j.1532-5415.2003.51115.x] [Medline: 12588585]
11. Pollock FH, Bethea A, Samanta D, Modak A, Maurer JP, Chumbe JT. Readmission within 30 days of discharge after hip fracture care. Orthopedics 2015 Jan;38(1):e7-13 [FREE Full text] [doi: 10.3928/01477447-20150105-53] [Medline: 25611424]
12. Härstedt M, Rogmark C, Sutton R, Melander O, Fedorowski A. Impact of comorbidity on 6-month hospital readmission and mortality after hip fracture surgery. Injury 2015 Apr;46(4):713-718. [doi: 10.1016/j.injury.2014.12.024] [Medline: 25627481]
13. Cram P, Lu X, Kaboli PJ, Vaughan-Sarrazin MS, Cai X, Wolf BR, et al. Clinical characteristics and outcomes of medicare patients undergoing total hip arthroplasty, 1991-2008. J Am Med Assoc 2011 Apr 20;305(15):1560-1567 [FREE Full text] [doi: 10.1001/jama.2011.478] [Medline: 21505134]
14. 2015 Procedure-specific Readmission Measures Updates and Specifications Report: Elective Primary Total Hip Arthroplasty and/or Total Knee Arthroplasty, and Isolated Coronary Artery Bypass Graft Surgery. Centers for Medicare and Medicaid Services. 2015. URL: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology.html [accessed 2020-05-26]
15. Aryal S, Diaz-Guzman E, Mannino D. Prevalence of COPD and comorbidity. Eur Respir Monogr 2013;59:1-12. [doi: 10.1183/1025448x.10011012]
16. Baty F, Putora PM, Isenring B, Blum T, Brutsche M. Comorbidities and burden of COPD: a population based case-control study. PLoS One 2013;8(5):e63285 [FREE Full text] [doi: 10.1371/journal.pone.0063285] [Medline: 23691009]
17. Moni MA, Liò P. Network-based analysis of comorbidities risk during an infection: SARS and HIV case studies. BMC Bioinformatics 2014 Oct 24;15:333 [FREE Full text] [doi: 10.1186/1471-2105-15-333] [Medline: 25344230]
18. Cramer AO, Waldorp LJ, van der Maas HL, Borsboom D. Comorbidity: a network perspective. Behav Brain Sci 2010 Jun;33(2-3):137-50; discussion 150. [doi: 10.1017/S0140525X09991567] [Medline: 20584369]
19. Lochner KA, Cox CS. Prevalence of multiple chronic conditions among medicare beneficiaries, United States, 2010. Prev Chronic Dis 2013 Apr 25;10:E61 [FREE Full text] [doi: 10.5888/pcd10.120137] [Medline: 23618541]
20. Islam MM, Valderas JM, Yen L, Dawda P, Jowsey T, McRae IS. Multimorbidity and comorbidity of chronic diseases among the senior Australians: prevalence and patterns. PLoS One 2014;9(1):e83783 [FREE Full text] [doi: 10.1371/journal.pone.0083783] [Medline: 24421905]
21. Folino F, Pizzuti C, Ventura M. A Comorbidity Network Approach to Predict Disease Risk. In: Proceedings of the International Conference on Information Technology in Bio- and Medical Informatics. 2010 Presented at: ITBAM'10; September 1-2, 2010; Bilbao, Spain p. 102-109. [doi: 10.1007/978-3-642-15020-3_10]
22. Bhavnani SK, Visweswaran S, Divekar R, Brasier AR. Towards team-centered informatics: accelerating innovation in multidisciplinary scientific teams through visual analytics. J Appl Behav Sci 2018 Nov 5;55(1):50-72. [doi: 10.1177/0021886318794606]
23. Report to the Congress: Medicare and the Health Care Delivery System. Chapter 3. Approaches to Bundle Payment for Post-Acute Care Washington. Medicare Payment Advisory Commission (MedPAC). 2013. URL: http://www.medpac.gov/docs/default-source/reports/jun13_entirereport.pdf?sfvrsn=0 [accessed 2020-06-16]
24. Pope G, Kautter J, Ingber M, Freeman S, Sekar R, Newhart C. Evaluation of the CMS-HCC Risk Adjustment Model. Centers for Medicare & Medicaid Services' Office of Research, Development, and Information. 2011. URL: https://www.

cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/downloads/evaluation_risk_adj_model_2011.pdf [accessed 2015-03-31]
25. St Sauver JL, Boyd CM, Grossardt BR, Bobo WV, Finney Rutten LJ, Roger VL, et al. Risk of developing multimorbidity across all ages in an historical cohort study: differences by sex and ethnicity. BMJ Open 2015 Feb 3;5(2):e006413 [FREE Full text] [doi: 10.1136/bmjopen-2014-006413] [Medline: 25649210]
26. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 2018 Dec 5;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
27. Treviño S, Nyberg A, del Genio CI, Bassler KE. Fast and accurate determination of modularity and its effect size. J Stat Mech 2015 Feb 3;2015(2):P02003. [doi: 10.1088/1742-5468/2015/02/p02003]
28. Chauhan R, Ravi J, Datta P, Chen T, Schnappinger D, Bassler KE, et al. Reconstruction and topological characterization of the sigma factor regulatory network of Mycobacterium tuberculosis. Nat Commun 2016 Mar 31;7:11062 [FREE Full text] [doi: 10.1038/ncomms11062] [Medline: 27029515]
29. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 1971 Dec;66(336):846. [doi: 10.2307/2284239]
30. Bhavnani SK, Chen T, Ayyaswamy A, Visweswaran S, Bellala G, Rohit D, et al. Enabling comprehension of patient subgroups and characteristics in large bipartite networks: implications for precision medicine. AMIA Jt Summits Transl Sci Proc 2017;2017:21-29 [FREE Full text] [Medline: 28815099]
31. Mak JC, Cameron ID, March LM, National Health and Medical Research Council. Evidence-based guidelines for the management of hip fractures in older persons: an update. Med J Aust 2010 Jan 4;192(1):37-41. [Medline: 20047547]
32. Rao S, Cherukuri M. Management of hip fracture: the family physician's role. Am Fam Physician 2006 Jun 15;73(12):2195-2200 [FREE Full text] [Medline: 16836036]
33. Mentes JC, Chang BL, Morris J. Keeping nursing home residents hydrated. West J Nurs Res 2006 Jun;28(4):392-406; discussion 407. [doi: 10.1177/0193945906286607] [Medline: 16672630]
34. Bennett JA. Dehydration: hazards and benefits. Geriatr Nurs 2000;21(2):84-88. [doi: 10.1067/mgn.2000.107135] [Medline: 10769332]
35. Lelovics Z. [Nutritional status and nutritional rehabilitation of elderly people living in long-term care institutions]. Orv Hetil 2009 Nov 1;150(44):2028-2036. [doi: 10.1556/OH.2009.28723] [Medline: 19861289]
36. Guigoz Y, Lauque S, Vellas BJ. Identifying the elderly at risk for malnutrition. The mini nutritional assessment. Clin Geriatr Med 2002 Nov;18(4):737-757. [doi: 10.1016/s0749-0690(02)00059-9] [Medline: 12608501]
37. Shipman D, Hooten J. Are nursing homes adequately staffed? The silent epidemic of malnutrition and dehydration in nursing home residents. Until mandatory staffing standards are created and enforced, residents are at risk. J Gerontol Nurs 2007 Jul;33(7):15-18. [doi: 10.3928/00989134-20070701-03] [Medline: 17672164]
38. Teixeira A, Trinquart L, Raphael M, Bastianic T, Chatellier G, Holstein J. Outcomes in older patients after surgical treatment for hip fracture: a new approach to characterise the link between readmissions and the surgical stay. Age Ageing 2009 Sep;38(5):584-589. [doi: 10.1093/ageing/afp124] [Medline: 19596738]
39. Ronco C, Haapio M, House AA, Anavekar N, Bellomo R. Cardiorenal syndrome. J Am Coll Cardiol 2008 Nov 4;52(19):1527-1539 [FREE Full text] [doi: 10.1016/j.jacc.2008.07.051] [Medline: 19007588]
40. Fabbian F, de Giorgi A, Manfredini F, Lamberti N, Forcellini S, Storari A, et al. Impact of renal dysfunction on in-hospital mortality of patients with severe chronic obstructive pulmonary disease: a single-center Italian study. Int Urol Nephrol 2016 Jul;48(7):1121-1127. [doi: 10.1007/s11255-016-1272-5] [Medline: 27020445]
41. Heyes GJ, Tucker A, Marley D, Foster A. Predictors for readmission up to 1 year following hip fracture. Arch Trauma Res 2015 Jun;4(2):e27123 [FREE Full text] [doi: 10.5812/atr.4(2)2015.27123] [Medline: 26101764]
42. Hung WW, Egol KA, Zuckerman JD, Siu AL. Hip fracture management: tailoring care for the older patient. J Am Med Assoc 2012 May 23;307(20):2185-2194. [doi: 10.1001/jama.2012.4842] [Medline: 22618926]
43. Forsythe L, Murray C, Shah B. Effectiveness of interprofessional care teams on reducing hospital readmissions in patients with heart failure: a systematic review. MedSurg Nurs 2018;27(3):- [FREE Full text]
44. Lasater KB, Mchugh MD. Nurse staffing and the work environment linked to readmissions among older adults following elective total hip and knee replacement. Int J Qual Health Care 2016 Apr;28(2):253-258 [FREE Full text] [doi: 10.1093/intqhc/mzw007] [Medline: 26843548]
45. Ouslander JG, Berenson RA. Reducing unnecessary hospitalizations of nursing home residents. N Engl J Med 2011 Sep 29;365(13):1165-1167. [doi: 10.1056/NEJMp1105449] [Medline: 21991889]
46. Silverberg D, Wexler D, Blum M, Schwartz D, Iaina A. The association between congestive heart failure and chronic renal disease. Curr Opin Nephrol Hypertens 2004 Mar;13(2):163-170. [doi: 10.1097/00041552-200403000-00004] [Medline: 15202610]
47. Lacy ME, Wellenius GA, Carnethon MR, Loucks EB, Carson AP, Luo X, et al. Racial differences in the performance of existing risk prediction models for incident type 2 diabetes: the cardia study. Diabetes Care 2016 Feb;39(2):285-291 [FREE Full text] [doi: 10.2337/dc15-0509] [Medline: 26628420]
48. Tan A, Kuo Y, Goodwin JS. Predicting life expectancy for community-dwelling older adults from medicare claims data. Am J Epidemiol 2013 Sep 15;178(6):974-983 [FREE Full text] [doi: 10.1093/aje/kwt054] [Medline: 23851579]

49. Bhavnani S, Lin Y, Chennuri L, Bores J, Chen C, Kuo Y. Identification, Replication, Visualization, and Interpretation of Patient Subgroups: Implications for Precision Medicine, and Predictive Modeling. In: AMIA Joint Summits on Translational Science Proceedings AMIA Summit on Translational Science. 2018 Presented at: AMIA'18; November 3-7, 2018; San Francisco, CA URL: https://informaticssummit2018.zerista.com/event/member/470507?embedded=1
50. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc 2014;21(4):578-582 [FREE Full text] [doi: 10.1136/amiajnl-2014-002747] [Medline: 24821743]
51. Amin W, Borromeo C, Saul M, Becich M, Visweswaran S. Informatics Synergies Between PaTH and ACT Networks. In: AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2015 Presented at: AMIA'15; November 12-16, 2015; Washington, DC URL: http://www.thevislab.com/lab/lib/exe/fetch.php?media=papers:2015_informatics_synergies_between_path_and_act_networks.pdf
52. Shneiderman B. Inventing discovery tools: combining information visualization with data mining. Inf Vis 2002 Mar;1(1):5-12. [doi: 10.1057/palgrave/ivs/9500006]

Abbreviations

CHF: congestive heart failure
CMS: Centers for Medicare & Medicaid Services
COPD: chronic obstructive pulmonary disease
CoRisk: co-occurrence risk
FDR: false discovery rate
HCCs: hierarchical condition categories
HFx: hip fracture
HRRP: Hospital Readmissions Reduction Program
MCMCT: major complications of medical care and trauma
MS-DRG: Medicare Severity-Diagnosis Related Group
NCATS: National Center for Advancing Translational Sciences
PCORI: Patient-Centered Outcomes Research Institute
RF: renal failure
RI: Rand index

Edited by G Eysenbach; submitted 31.01.19; peer-reviewed by E Massou, H Sievänen, V Giordano, G Heyes; comments to author 02.03.19; revised version received 08.10.19; accepted 16.12.19; published 26.10.20.

Please cite as:
Bhavnani SK, Dang B, Penton R, Visweswaran S, Bassler KE, Chen T, Raji M, Divekar R, Zuhour R, Karmarkar A, Kuo YF, Ottenbacher KJ
How High-Risk Comorbidities Co-Occur in Readmitted Patients With Hip Fracture: Big Data Visual Analytical Approach
JMIR Med Inform 2020;8(10):e13567
URL: https://medinform.jmir.org/2020/10/e13567
doi:10.2196/13567
PMID:33103657

Original Paper

Fueling Clinical and Translational Research in Appalachia: Informatics Platform Approach

Alfred A Cecchetti[1], MScIT, MS, PhD; Niharika Bhardwaj[1], MBBS, MS; Usha Murughiyan[1], MBBS; Gouthami Kothakapu[1], MSCS; Uma Sundaram[1], MD

Department of Clinical and Translational Science, Joan C. Edwards School of Medicine, Marshall University, Huntington, WV, United States

Corresponding Author:
Alfred A Cecchetti, MScIT, MS, PhD
Department of Clinical and Translational Science
Joan C Edwards School of Medicine
Marshall University
1600 Medical Center Drive
Huntington, WV, 25701
United States
Phone: 1 304 691 1585
Email: cecchetti@marshall.edu

Abstract

Background: The Appalachian population is distinct, not just culturally and geographically but also in its health care needs, facing the most health care disparities in the United States. To meet these unique demands, Appalachian medical centers need an arsenal of analytics and data science tools with the foundation of a centralized data warehouse to transform health care data into actionable clinical interventions. However, this is an especially challenging task given the fragmented state of medical data within Appalachia and the need for integration of other types of data such as environmental, social, and economic with medical data.

Objective: This paper aims to present the structure and process of the development of an integrated platform at a midlevel Appalachian academic medical center along with its initial uses.

Methods: The Appalachian Informatics Platform was developed by the Appalachian Clinical and Translational Science Institute's Division of Clinical Informatics and consists of 4 major components: a centralized clinical data warehouse, modeling (statistical and machine learning), visualization, and model evaluation. Data from different clinical systems, billing systems, and state- or national-level data sets were integrated into a centralized data warehouse. The platform supports research efforts by enabling curation and analysis of data using the different components, as appropriate.

Results: The Appalachian Informatics Platform is functional and has supported several research efforts since its implementation for a variety of purposes, such as increasing knowledge of the pathophysiology of diseases, risk identification, risk prediction, and health care resource utilization research and estimation of the economic impact of diseases.

Conclusions: The platform provides an inexpensive yet seamless way to translate clinical and translational research ideas into clinical applications for regions similar to Appalachia that have limited resources and a largely rural population.

(JMIR Med Inform 2020;8(10):e17962) doi:10.2196/17962

KEYWORDS

Appalachian region; medical informatics; health care disparities; electronic health records; data warehousing; data mining; data visualization; machine learning; data science

Introduction

Background: Unique Challenges in Appalachia

With regard to health care, Appalachia with its predominantly rural communities is known to have one of the worst outcomes in the United States [1]. This is especially true of southern and central rural Appalachia, which face some of the most severe health disparities in the nation [1]. Over the years, the gap in the overall health between Appalachia and the nation as a whole has continued to grow [2,3]. To close this gap, it is critical to identify the cause of these disparities and direct efforts toward developing necessary interventions to address them.

Such an effort necessitates the adoption of modern technologies such as a centralized research data warehouse to house all data necessary to obtain a comprehensive picture of the health of the Appalachian population before analysis to gain actionable insights can be performed. A centralized data warehouse, once considered strictly a business tool, has evolved into an important instrument for cost containment, tracking of patient outcome, providing clinical decision support at the point of care, improving prognostic accuracy, and facilitating research [4]. Thus, rural academic medical centers have moved toward implementing data warehouse systems that feed analytical systems for research needs [5]. This entails (1) the integration of data from different types of medical settings (ie, multi-institutional) such as hospitals, clinics, and specialty centers; (2) linkage of financial data with clinical data—a well-established practice proven to be pivotal to high-quality care and great economic outcomes [6,7]; and (3) integration of other determinants of health such as environmental [8], social [9], and spiritual factors [10] to create longitudinal health records across the care continuum.

However, there are challenges in creating a multi-institutional data warehouse [11]. The electronic health records (EHRs) do not easily interact with one another due to the use of nonstandard terminologies and difficulty in understanding the flow of information. In addition, significant differences exist between rural and urban health systems [12-16]. Unlike their urban counterparts, health care data in Appalachia are typically fragmented, existing in silos within dissimilar databases, registries, data collections, and departmental systems. With innovations in medical technology, the list of data sources continues to grow, producing unprecedented amounts of data from all aspects of care, including diagnosis, medication, procedures, laboratory test results, imaging data, and patient self-monitoring [17-21]. To complicate matters, the overall health and health behaviors of Appalachians are strongly affected by Appalachia's unique culture, geography, and health system issues [22-24]. Consequently, Appalachian academic medical centers face the complex challenge of collecting, organizing, standardizing, and analyzing these enormous quantities of heterogeneous data originating from a wide variety of sources to address the unmet needs of the population they serve.

Why an Informatics Platform?

Data integration and interoperability have been shown to be key to unlocking these data for data analytics, enabling the development of novel patient management strategies for rural hospitals [25,26] and translational research that leads to new approaches at the bedside for prevention, diagnosis, and treatment of disease, which are essential to improving the health of a population [27-29]. Data analytics, once the domain of the statistician, has now become an equal partner in clinical research and research operations [30,31]. Following the data explosion, data analytics increasingly involves the use of visual analytics tools such as Tableau (Tableau Software Inc) and Power BI (Microsoft Corp) to explore data easily and in a self-service fashion and to clearly and effectively communicate complex ideas [32], especially to those members of the medical community who might not have an intimate understanding of the underlying data. Furthermore, machine learning is gaining importance, especially in the area of predictive analytics, to improve the practice of medicine and to infer potentially innovative risk factors [28,33-35].

However, these applications (eg, data warehouse, data analytics, statistical analysis, machine learning, visual analytics) are generally uncoordinated without any overarching governance. Thus, we developed an informatics platform, that is, a suite of interconnected, coordinated applications hosted within an operational environment [36], called the Appalachian Informatics Platform, in West Virginia—the only state located entirely in Appalachia—that facilitates interoperable access to integrated information, data visualization, and data analytics, thereby functioning as an excellent basis for clinical and translational research to improve health care.

The goal of this study is to describe the structure and process of development of the Appalachian Informatics Platform and demonstrate its value in supporting clinical and translational research.

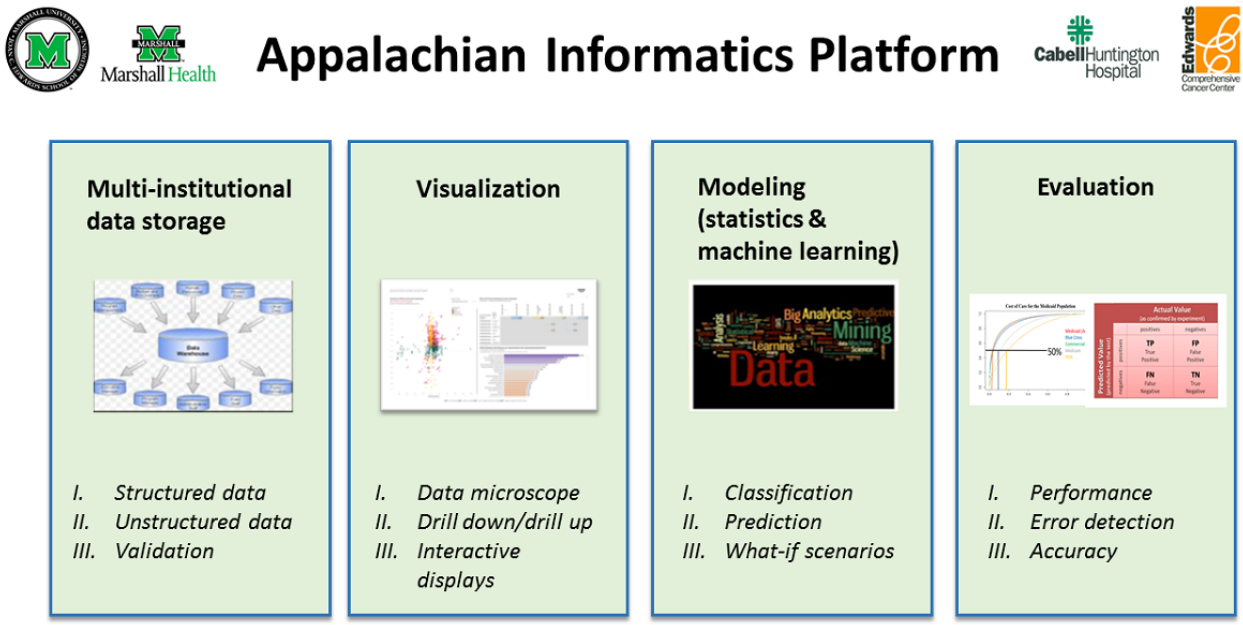
Methods

The Appalachian Informatics Platform (Figure 1) is composed of 4 major components: (1) multi-institutional data storage—clinical data warehouse (CDW); (2) modeling (statistical and machine learning); (3) visualization; and (4) evaluation. Each of these components is described in detail in separate sections.

The CDW forms an integral part of the Appalachian Informatics Platform. The Appalachian Informatics Platform, in addition to the CDW, contains embedded data analytics (modeling and evaluation) and interactive visualization tools (eg, Tableau [Tableau Software Inc], Power BI [Microsoft Corp]). Together, these enable the analysis of Appalachian health information to speed up the transition of translational research ideas into clinical practice.

The CDW serves as a secure source of quality data for descriptive, diagnostic, predictive, and prescriptive analytics for research and operational needs. The visual analytics tools enable an initial exploratory analysis of the processed data and the interactive presentation of analytical findings for further analysis and review. Depending on the use case, data can be analyzed using statistical modeling via external (eg, SPSS [IBM Corp], Stata [StataCorp]) or integrated (eg, R [R Foundation for Statistical Computing], Python [Python Software Foundation] in Structured Query Language [SQL]) applications or machine learning modeling. The performance of the resulting models was evaluated using appropriate metrics. Once trained and evaluated, machine learning models can be deployed and stored in the CDW for future use if needed. Furthermore, the stored machine learning models can be continuously evaluated and improved as more data are generated.

Figure 1. Appalachian informatics platform.



The informatics committee governs the access to and utilization of the Appalachian Informatics Platform and ensures adherence to security and privacy rules. In addition, team-building activities are also incorporated into our clinical informatics model to foster the development of an effective clinical informatics team.

Multi-Institutional Data Storage: Appalachian Clinical and Translational Science Institute-Clinical Data Warehouse

The Appalachian Clinical and Translational Science Institute (ACTSI)'s Division of Clinical Informatics solicited buy-in from different entities, namely, Cabell-Huntington Hospital (CHH), Edwards Comprehensive Cancer Institute (ECCC), Marshall Health (MH) practice plan, and Marshall University Joan C Edwards School of Medicine (MU JCESOM), to build the Appalachian Clinical and Translational Science Institute-Clinical Data Warehouse (ACTSI-CDW) in West Virginia. An agreement was created between these entities that provided access to both financial and clinical data.

The multi-institutional CDW contains more than 9 years of billing and clinical data. It comprises relational tables and dimension and fact tables (Online Analytical Processing [OLAP] cube), which enable secure data storage and data access. Designed from the start to facilitate information flow, the CDW can send out a stream of near real-time data that can be used for any authorized research purpose. Documentation includes a data dictionary and flowcharts. Flowcharts follow the patient from admission (or appointment, if outpatient) to discharge (or exit, if outpatient). The data dictionary contains the standardized and source field names, descriptions, and properties along with the associated metadata for the data contained within the data warehouse. For instance, (1) the entry of a patient into any medical service (admission or appointment) was combined with the single term *encounter* and (2) a higher level of precision was introduced by separating patient age into 2 variables, current age or the age when the procedure was performed.

The CDW process is based on an older data warehouse process developed at the University of Pittsburgh [37]. The process is as follows:

1. Data dictionaries are created by recording institutional source field names and field properties and linking them to the standardized CDW names and properties found within the CDW databases. Descriptions of each field (source and CDW) are included.
2. Individual institutional flowcharts show the workflow of the data and the location of the people responsible for the quality of the data, which are also used for quality control purposes.
3. At present, the CDW contains data from 6 institutional software packages hosted in various parts of the country (eg, Cerner data from Kansas City, Missouri; McKesson data from North Druid Hills, Georgia; etc). The data are exported in a standard format (ie, ASCII flat file, XML, etc) and transferred through secure file transfer protocol (eg, Cerberus [Cerberus, LLC]) to the CDW Development server.
4. The data are integrated into the Microsoft SQL databases using Microsoft SQL Server Integration Services (SSIS), a graphical tool that extracts, transforms, and loads (ETL) the data to target schemas that will be used to contain the target data objects: relational tables, dimensions, and cubes. ETL systems enable a smooth migration from one system to another irrespective of the underlying storage system.
5. Conformed dimensions were developed, and patient linkages using various methods (eg, simple heuristics) [38] were also available and made at this time.

6. At present, a transactional grain fact table has been developed, but other fact tables will be created as needed.
7. The CDW contains internal structured billing and EHR data (ie, demographics, encounter details, vitals, medications, procedures, diagnoses, orders, immunizations, laboratory and imaging results, date and time, payee, and provider). It also contains unstructured EHR data (eg, H&P, admission notes, discharge summaries, other clinical notes). These data are received from MH, CHH, and MU JCESOM's ECCC as well as from other outside institutions. In addition, non-EHR data are incorporated using REDCap.
8. Unstructured data are analyzed using text analytics tools, and classification variables based on text mining are incorporated into the CDW.
9. The data structure (OLAP cubes and relational tables), once checked and verified, is transferred from the secure development server to the secure production server for use.
10. Various security measures (eg, IP and password restrictions) are in place to prevent unauthorized use.
11. The CDW structure, which stores multi-institutional medical information, can now provide data for both operational and research analytical model development (statistical or machine learning) using very simple deidentified interfaces (eg, Excel [Microsoft Corp]) or more complex interactive tools (eg, R [R Foundation for Statistical Computing], Tableau [Tableau Software Inc], Power BI [Microsoft Corp], etc). Within the CDW, the data can be manipulated, cleaned, and prepared before the analysis as needed.
12. Structured and unstructured data currently exist within the CDW. Image and BioSample data will soon be incorporated (like the Pittsburgh model), but the full design has not been finalized yet. An *Honest Broker* person assumes control of sample shipping and receiving.
13. Standard Operation Procedures have been developed for administrative and technical areas.
14. The Health Insurance Portability and Accountability Act (HIPAA) guidelines are followed, and protocol to protect patient information has also been implemented.

The CDW is contained within a Microsoft SQL database that can interact with outside objects using other electronic methods such as SignalR, a software library for Microsoft ASP.NET that allows server code to send asynchronous notifications to client-side web applications and SqlDependency, an object that represents a query notification dependency between an application and an instance of SQL server. Objects such as these provide the ability for the data warehouse to interact in real time with the outside regional population using the newest technologies such as Microsoft Machine Learning Server with embedded R or Python procedure coding.

Data Validation

The information derived from multiple data sources can have inconsistencies and missing values because of their heterogeneous nature that needs to be corrected [39-42]. Thus, for each research study, clinical and translational researchers using the data warehouse are required to verify a random sample (calculated on the basis of the size of the study population) of all extracted study data are directly verified at the original data source to ensure data accuracy and validity. Identified errors or omissions are transmitted back to the host systems for correction or inclusion.

Augmenting the CDW Using REDCap

For certain studies, data available in the CDW may not be precise enough or include variables needed to perform this study. For such studies, data can be augmented using data capture tools. One such tool is the Research Electronic Data Capture, or REDCap, a workflow methodology and software solution designed for the rapid development and deployment of electronic data capture tools to support clinical and translational research [43-45].

Our institution has deployed and maintains 2 REDCap servers: secure (located under institutional firewall) and global (outside the firewall). The secure REDCap system is used for storing data considered protected health information (PHI) under HIPAA. The global system, on the other hand, is used to store deidentified or non-PHI data. These data are then transferred to and stored within the multi-institutional data warehouse. This method of augmenting the information pulled from the existing source systems provides research-grade data from outside sources that are normally not contained within a data warehouse.

Visualization

Visualization of information is an excellent method of providing knowledge that can be easily understood by any member of the health care discipline. Within the informatics platform, Tableau provides interactive drill-down and drill-up capabilities for specific projects.

Tableau is a visual analytics tool that provides an interactive method of exploring multidimensional data, optimized from the data warehouse and OLAP data sources. Tableau, using either indexed relational tables or a data cube, can perform associated operations such as slice, dice, roll-up, and drill-down on the data, providing detailed interactive visual overlays that range from the lowest grain of the data to high-level representations of the data. Tableau charts, graphs, filters, and maps can provide visualization of the various subgroups of interest using a storyboard approach that presents a specific question followed by an interactive dashboard that explores that question in detail. The use of visual elements such as logos, pictograms, icons, or pictures into the dashboards, in association with the subgroups, provides easy-to-reference image aids that provide clarity and understanding of complex information. The data warehouse provides the drill-down, drill-up and slice and dice capability, whereas the hub design connects both financial and clinical data to provide a full picture.

The developed interactive dashboards are securely shared with users within a department or a team, as needed, through the use of Tableau Server [46].

Modeling (Statistics and Machine Learning)

The modeling component of the informatics platform supports the construction of tailored regional models (statistical or machine learning) to understand and predict disease and other medical events within this region. EHR is primarily a billing system, research only being a secondary function and, thus, is heterogeneous, incomplete, and noisy [25], leading to

unrepresentative samples, selection bias, and misclassification [47]. During the modeling process, these issues are eliminated or minimized.

To assist in modeling, software packages such as Stata [StataCorp] and SPSS [IBM Corp] and embedded open-source machine learning programs (eg, R [R Foundation for Statistical Computing], Python [Python Software Foundation]) are used. This enables faster and easier development of classification, regression, and clustering algorithms for research use. In addition, we utilize products such as Microsoft's LINQ to electronically gather information and directly incorporate that information into the CDW.

Evaluation

During the modeling process, evaluation of the data set as it relates to the regional population is carried out. Local experts native to this region are asked to evaluate the model from a clinical as well as a financial standpoint. Poverty is endemic within the Appalachian population, and a model that suggests the use of a very expensive medication or procedure over an older but less expensive medication or procedure is unlikely to be used [48]. Thus, the model must take into account whether the patient has the means and access to the recommended medication or procedure [49]. In addition, the willingness of Appalachian medical institutions and health care providers to follow the model's suggestions must also be evaluated.

Once developed, the models were tuned and tested. Location, time of treatment, outside temperature, and other contributory factors available within the CDW were employed to fine-tune the models, as applicable. The performance of the models was measured using the R programming environment using measures such as area under curve, sensitivity, specificity, F_1 score, precision, recall, etc.

Security, Privacy, and the Informatics Committee

Data access and usage are permitted only as described in the mutual agreement between the 3 institutions and are subject to internal security and privacy rules. All data requests must follow the standard operating procedure built on the basis of mutual multi-institutional agreement. Foremost, the researcher must have appropriate credentials and authorization to be able to request for data. If the researcher is authorized to make requests, he or she must obtain the IRB approval for his or her proposed study and submit the IRB proposal and supporting documentation for review by the informatics committee. The informatics committee, independent of the IRB, reviews all requests for data from the data warehouse to ensure compliance with the agreement. If the research project is approved, the research team designated members are scheduled for the deidentified data extraction process.

Team Building

Integral to the informatics platform is team building that builds upon previous work [37]. To facilitate effective team meetings and interprofessional collaboration (local and global) without the need or expense of constant travel, a permanent clinical informatics conference room with a fixed connected computer, an uninterruptable power supply (UPS), a smart board, a camera, and a speaker system, along with a video conferencing system (Zoom) connectivity, was built. This ensures adequate communication among all those involved (ie, team members, users, leadership, etc) and access to resources that would otherwise be unavailable.

Results

Since the implementation of the platform, several studies have been conducted. Each study listed below was approved by the informatics committee, and the deidentified data and platform tools were made available securely to the research team.

To evaluate the functionality and value of this platform, we first analyzed the aggregated data of Medicaid-insured patients across different health systems using the interconnected applications within the platform for population health management. Relevant data were extracted from the CDW, followed by exploratory analysis using a Tableau dashboard. Due to the isolated nature of the study population, regional variables such as distance from the CHH and weather conditions (ie, temperature) were also included. Errors and missing values were identified using the dashboard, and data were subsequently cleaned and prepared. Using these clean data, the regional population was classified into 3 spend categories: *low cost*, *acute*, and *persistent* subgroups on the basis of the charges accrued. Next, the Charlson Comorbidity Index (CCI) was incorporated into the CDW to predict mortality risk within 1 year of hospitalization for patients with comorbid conditions within each spend category (Table 1) [50,51]. Of these categories, the persistent group had the largest percentage of patients with a high risk of mortality, followed by acute and low cost after excluding the deceased patients (persistent: 898/1247, 72.01%; acute: 2074/6946, 29.86%; low cost: 5130/102,814, 4.99%). The CCI was not very sensitive in predicting the risk of mortality but was very specific and accurate (sensitivity: 896/1512, 59.26%; specificity: 102,905/111,007, 92.7%; accuracy: 103,801/112,519, 92.25%). The effect of distance and weather on the CCI needs further investigation that is being conducted. Adjustments are being made to this standard national index to incorporate other Appalachian characteristics that could improve the sensitivity of this risk scoring system.

This way, the platform has been utilized for a variety of purposes such as increasing knowledge of the pathophysiology of diseases, risk identification, risk prediction, health care resource utilization research, and estimation of the economic impact of diseases to enable data-driven clinical decisions, leading to improved clinical outcomes. Textbox 1 contains a list of studies conducted so far.

Table 1. The 10-year mortality risk predicted using the Charlson Comorbidity Index.

Mortality risk	Deceased, n (%)	Alive, n (%)
High risk	896 (0.80)	8102 (7.20)
Low risk	616 (0.55)	102,905 (91.46)

Textbox 1. Studies conducted using the Appalachian Informatics Platform.

Diagnostic accuracy improvement studies

- Albumin Level as a Risk Marker and Predictor of Peripartum Cardiomyopathy [52]
- Clinical Determinants of Myocardial Injury, Detectable and Serial Troponin Levels Among Patients With Hypertensive Crisis [53]
- Is Fever a Red Flag for Secondary Bacterial Pneumonia During RSV Bronchiolitis [54]
- Metabolic Syndrome: Are Current Colon Cancer Screening Guidelines Enough in a Rural Population? [55]
- Utilization of Appalachian Clinical and Translational Science Institute Data Warehouse to More Accurately Predict Disease Processes Important for Central Appalachia [56]

Resource utilization and financial impact research studies

- Fueling Dementia Research in Appalachia via Appalachian Informatics Platform: A Longitudinal Study [57]
- Hospital Emergency Department Visits For Non-Traumatic Oral Health Conditions [58]

Studies to understand disease pathophysiology

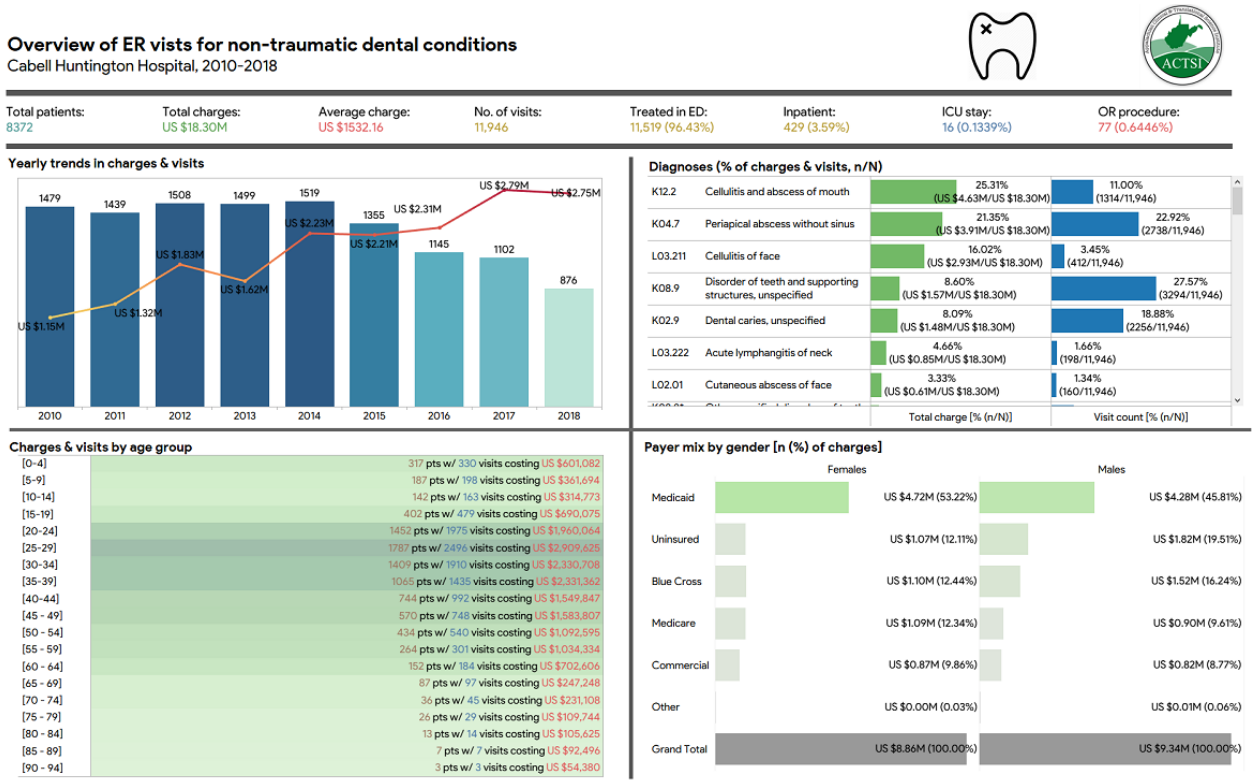
- Serum Calcium Homeostasis and Volume Dynamics in Alzheimer's Disease and Diabetes Mellitus-2 [59]

Five studies utilized the platform for risk identification and risk prediction to improve diagnostic accuracy [52-56]. Sundaram et al [56] demonstrated the value of ACTSI-CDW as a primary source to improve the diagnosis of metabolic syndrome, a diagnosis very relevant to the Central Appalachian population. The researchers discovered that utilizing billing codes alone severely underestimated the number of patients with metabolic syndrome by a factor of more than 10 as compared with looking at specific criteria that determine this diagnosis [56]. Another study assessed the relationship between metabolic syndrome and colorectal cancer and found that patients with metabolic syndrome, especially those with insulin resistance, were more likely to have colorectal cancer, indicating the probable need for earlier screening for colorectal cancer in these patients [55]. Elmore et al [54] examined the role of fever in predicting the development of secondary bacterial pneumonia in children with RSV and other viral illnesses. They found that febrile children were 2 to 8 times (RSV, 47/78 vs 27/100; other bronchiolitis, 54/83 vs 7/88) more likely to have secondary bacterial pneumonia compared with afebrile children and, thus, may need to be aggressively evaluated to enable early diagnosis and treatment [54]. Amro et al [52] studied the relationship between hypoalbuminemia and peripartum cardiomyopathy and noted that lower albumin levels were significantly associated with peripartum cardiomyopathy (P<.001; odds ratio 0.033, 95% CI 0.034-0.865) and could potentially be used as a risk marker for it. Acosta et al [53] used data from the ACTSI-CDW to identify risk factors (lower BMI, before CHF, and prior use of aspirin) that predict myocardial injury, detectable troponin, and increase in serial troponin levels in patients with hypertensive crisis.

Ferdjallah et al [59] analyzed the data from the ACTSI-CDW to understand how Alzheimer disease and diabetes mellitus affect serum calcium homeostasis and extracellular fluid volume. They observed that acute changes in serum calcium were significantly correlated with changes in extracellular fluid volume in both disease states [59].

The platform has also been applied in 2 studies to assess resource utilization (eg, emergency room, medications, etc) and the financial impact of the disease. For instance, Bhardwaj et al [57] utilized the platform to identify the problems associated with benzodiazepine use in geriatric patients within the health system, such as a higher number of emergency room visits and charges in geriatric patients with dementia plus at least one BZD prescription. In another study [58] that aimed to measure the volume and cost of emergency room use for these conditions and identify the factors that predict such use, the researchers built a dashboard (Figure 2) to easily explore and analyze relevant data on nontraumatic dental conditions that led to emergency room visits and to report the key findings of the study. The authors [58] observed that emergency room visits by uninsured patients were 4 times more likely and those by Medicaid insured 2 times more likely to be for dental problems than Medicare-insured patients.

Figure 2. Tableau dashboard displaying patterns and trends in charges for non-traumatic dental ER visits at Cabell Huntington Hospital between 2010 and 2018. ER: emergency room.



Discussion

Utility of the Appalachian Informatics Platform

The Appalachian Informatics Platform has supported several research projects involving the use of different components of the platform, depending on project needs. The studies described reported findings that are seldom reported in this region, enhanced our knowledge of pathophysiology and risk factors, and helped estimate and analyze resource utilization and economic burden of certain diseases within Appalachia using minimal resources (a small IT team and a relatively inexpensive platform).

Before the implementation of the platform, many research studies that followed the patient across multiple care settings or involved analysis of big data were not possible due to the unavailability of technical and economic resources owing to a lack of buy-in from rural health care organizations. As the data existed in silos, there was a lack of standardization and normalization, which resulted in major data inconsistencies. Studies conducted using these disjointed data sets often used unrepresentative small biased samples and had low statistical power and quality.

The introduction of the platform has helped address these issues. It is now easier to pinpoint and correct errors and/or missing values and understand the distribution of data using visual analysis tools. Further, the time needed to conduct these studies from start to finish has been greatly reduced owing to the availability of all applications necessary to complete the study within the platform. This has been specifically useful because many researchers do not have the technical skills needed to perform complex and advanced data analysis, especially on larger data sets.

The paper also revealed that national models do not necessarily perform well when applied to the Appalachian population. The Appalachian Informatics Platform allows for seamless integration of regional variables into the national model, which may improve the performance of these models. For each of the top 10 causes of death in West Virginia in 2017 per the Centers for Disease Control and Prevention [60], a machine learning algorithm was used to predict outcomes on a national level: heart disease [61,62], cancer [63,64], accidents [65,66], respiratory disease [67,68], stroke [69,70], diabetes [71,72], Alzheimer disease [73,74], pneumonia [75,76], kidney disease [77,78], and suicide [79,80]. Each of these cited models could be modified to fit the characteristics of the Appalachian population, especially those characteristics that make this region different in terms of geography, economy, education, and culture from the rest of the United States. The development of these regional models could help rural health general practitioners tackle complex medical conditions without the need for an expensive specialized health care provider nearby [46].

We hope that this paper will help other rural health care organizations, such as ours, that serve underserved populations realize the value and ease of using an informatics platform to conduct research and improving care for their patients despite limited resources.

Ongoing Projects and Future Directions

At present, a model that utilizes embedded data analytics to monitor the side effects of certain types of cancer by ingesting deidentified statements in the regional variety of English

language from patients within this region [81,82] is under development. This model could be used to analyze patient responses at a certain point in time for a cross-sectional study or continuously in real time for a long-term longitudinal study to identify the patients in need of care before their scheduled follow-up visit. The ongoing results from this model would be sent to their health care providers for appropriate actions. In case of an emergency, patient-designated community support networks such as religious or other support groups may be intimated to bring the patient to the emergency department so that the patient can receive timely care.

We plan to expand upon our unified informatics platform to integrate programming applications for the development of state-of-the-art applications targeted specifically toward the unmet health care needs of the Appalachian population.

Conclusions

This paper establishes the value of the Appalachian Informatics Platform in enabling seamless and secure data access, model development through an analytics engine to explore novel and unexpected hypotheses, and simple yet effective communication of all findings via interactive visualization.

The relatively inexpensive nature of such a platform coupled with its demonstrated advantages will hopefully encourage small and midsized rural academic centers, which traditionally have fewer resources than their urban counterparts, to adopt a research informatics platform within their institutions using the template described in this paper as a guide.

Acknowledgments

This work was supported by the National Institutes of Health grants DK-67420, DK-108054, and P20GM121299-01A1 and Veteran's Administration Merit Review grant BX003443-01 to US and UL1TR00011719 to PK.

Conflicts of Interest

None declared.

References

1. Krometis L, Gohlke J, Kolivras K, Satterwhite E, Marmagas SW, Marr LC. Environmental health disparities in the central Appalachian region of the United States. Rev Environ Health 2017 Sep 26;32(3):253-266. [doi: 10.1515/reveh-2017-0012] [Medline: 28682789]
2. Singh GK, Kogan MD, Slifkin RT. Widening disparities in infant mortality and life expectancy between Appalachia and the rest of the United States, 1990-2013. Health Aff (Millwood) 2017 Aug 1;36(8):1423-1432. [doi: 10.1377/hlthaff.2016.1571] [Medline: 28784735]
3. Marshall J, Thomas L, Lane N. Health disparities in Appalachia. Appalachian Regional Commission. 2017. URL: https://www.arc.gov/wp-content/uploads/2020/06/Health_Disparities_in_Appalachia_August_2017.pdf [accessed 2020-09-25]
4. Foran DJ, Chen W, Chu H, Sadimin E, Loh D, Riedlinger G, et al. Roadmap to a comprehensive clinical data warehouse for precision medicine applications in oncology. Cancer Inform 2017;16:1176935117694349. [doi: 10.1177/1176935117694349] [Medline: 28469389]
5. Kaufman A, Rhyne RL, Anastasoff J, Ronquillo F, Nixon M, Mishra S, et al. Health extension and clinical and translational science: an innovative strategy for community engagement. J Am Board Fam Med 2017 Jan 2;30(1):94-99. [doi: 10.3122/jabfm.2017.01.160119] [Medline: 28062823]
6. Roberts MS, Dreese EM, Hurley N, Zullo N, Peterson M. Blending administrative and clinical needs: the development of a referring physician database and automatic referral letter. Proc Annu Symp Comput Appl Med Care 1991:559-563 [FREE Full text] [Medline: 1807664]
7. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014;2:3. [doi: 10.1186/2047-2501-2-3] [Medline: 25825667]
8. Ahern MM, Hendryx M. Health disparities and environmental competence: a case study of appalachian coal mining. Environmental Justice 2008 Jun;1(2):81-86. [doi: 10.1089/env.2008.0511]
9. McCulloch B. The relationship of family proximity and social support to the mental health of older rural adults: The Appalachian context. Journal of Aging Studies 1995 Mar;9(1):65-81. [doi: 10.1016/0890-4065(95)90026-8]
10. Simpson MR, King MG. 'God brought all these churches together': issues in developing religion-health partnerships in an Appalachian community. Public Health Nurs 1999 Feb;16(1):41-49. [doi: 10.1046/j.1525-1446.1999.00041.x] [Medline: 10074821]
11. Holve E, Segal C, Lopez MH. Opportunities and challenges for comparative effectiveness research (CER) with electronic clinical data: a perspective from the EDM forum. Med Care 2012 Jul;50 Suppl:S11-S18. [doi: 10.1097/MLR.0b013e318258530f] [Medline: 22692252]
12. Rabinowitz HK, Paynter NP. MSJAMA. The rural vs urban practice decision. J Am Med Assoc 2002 Jan 2;287(1):113. [Medline: 11754723]

13. Anderson AE, Henry KA, Samadder NJ, Merrill RM, Kinney AY. Rural vs urban residence affects risk-appropriate colorectal cancer screening. Clin Gastroenterol Hepatol 2013 May;11(5):526-533 [FREE Full text] [doi: 10.1016/j.cgh.2012.11.025] [Medline: 23220166]
14. Reif S, Whetten K, Ostermann J, Raper JL. Characteristics of HIV-infected adults in the deep south and their utilization of mental health services: a rural vs. urban comparison. AIDS Care 2006;18 Suppl 1:S10-S17. [doi: 10.1080/09540120600838738] [Medline: 16938670]
15. Shubhakaran KP, Khichar RJ. Stroke Management Disparity in Urban Vs Rural Locations. American Academy of Neurology. 2018. URL: https://n.neurology.org/content/stroke-management-disparity-urban-vs-rural-locations [accessed 2020-09-25]
16. Newgard CD, Fu R, Bulger E, Hedges JR, Mann NC, Wright DA, et al. Evaluation of rural vs urban trauma patients served by 9-1-1 emergency medical services. J Am Med Assoc Surg 2017 Jan 1;152(1):11-18 [FREE Full text] [doi: 10.1001/jamasurg.2016.3329] [Medline: 27732713]
17. 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]
18. Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. IEEE Access 2017;5:8869-8879. [doi: 10.1109/ACCESS.2017.2694446]
19. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 May 2;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
20. Wang Y, Kung L, Byrd TA. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change 2018 Jan;126:3-13. [doi: 10.1016/j.techfore.2015.12.019]
21. Bhardwaj N, Wodajo B, Spano A, Neal S, Coustasse A. The impact of big data on chronic disease management. Health Care Manag (Frederick) 2018;37(1):90-98. [doi: 10.1097/HCM.0000000000000194] [Medline: 29266087]
22. Elam C. Culture, poverty and education in Appalachian Kentucky. Educ Culture 2002;18(1):10-13.
23. Coyne C, Demian-Popescu C, Friend D. Social and cultural factors influencing health in southern West Virginia: a qualitative study. Prev Chronic Dis 2006 Oct;3(4):A124 [FREE Full text] [Medline: 16978499]
24. Behringer B, Friedell GH. Appalachia: where place matters in health. Prev Chronic Dis 2006 Oct;3(4):A113 [FREE Full text] [Medline: 16978488]
25. Kim J, Ohsfeldt RL, Gamm LD, Radcliff TA, Jiang L. Culture, poverty and education in Appalachian Kentucky hospital characteristics are associated with readiness to attain stage 2 meaningful use of electronic health records. J Rural Health 2017 Jun;33(3):275-283. [doi: 10.1111/jrh.12193] [Medline: 27424940]
26. Mason P, Mayer R, Chien W, Monestime J. Overcoming Barriers to Implementing Electronic Health Records in Rural Primary Care Clinics. The Qualitative Report 2017;22(11):2943-2955 [FREE Full text]
27. Woolf SH. The meaning of translational research and why it matters. J Am Med Assoc 2008 Jan 9;299(2):211-213. [doi: 10.1001/jama.2007.26] [Medline: 18182604]
28. Karstoft K, Galatzer-Levy IR, Statnikov A, Li Z, Shalev AY, Members of Jerusalem Trauma OutreachPrevention Study (J-TOPS) group. Bridging a translational gap: using machine learning to improve the prediction of PTSD. BioMed Central Psychiatry 2015 Mar 16;15:30 [FREE Full text] [doi: 10.1186/s12888-015-0399-8] [Medline: 25886446]
29. Ethier J, Curcin V, Barton A, McGilchrist MM, Bastiaens H, Andreasson A, et al. Clinical data integration model. Core interoperability ontology for research using primary care data. Methods Inf Med 2015;54(1):16-23. [doi: 10.3414/ME13-02-0024] [Medline: 24954896]
30. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 2014 Jul;33(7):1123-1131. [doi: 10.1377/hlthaff.2014.0041] [Medline: 25006137]
31. Handelsman D. Applying Business Analytics to Optimize Clinical Research Operations. SAS Institute. 2012. URL: https://support.sas.com/resources/papers/proceedings12/171-2012.pdf [accessed 2020-09-28]
32. Simpao AF, Ahumada LM, Gálvez JA, Rehman MA. A review of analytics and clinical informatics in health care. J Med Syst 2014 Apr;38(4):45. [doi: 10.1007/s10916-014-0045-x] [Medline: 24696396]
33. Iwabuchi SJ, Liddle PF, Palaniyappan L. Clinical utility of machine-learning approaches in schizophrenia: improving diagnostic confidence for translational neuroimaging. Front Psychiatry 2013;4:95 [FREE Full text] [doi: 10.3389/fpsyt.2013.00095] [Medline: 24009589]
34. Ainali C. Machine learning for translational medicine. King's College London (University of London). 2013. URL: https://kclpure.kcl.ac.uk/portal/files/31802684/2013_Ainali_Chrysanthi_0829730_ethesis.pdf [accessed 2020-09-28]
35. Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc 2011;18(5):601-606 [FREE Full text] [doi: 10.1136/amiajnl-2011-000163] [Medline: 21508414]
36. Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P, et al. A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogenous clinical data. Med Care 2012 Jul;50 Suppl:S49-S59 [FREE Full text] [doi: 10.1097/MLR.0b013e318259c02b] [Medline: 22692259]
37. Cecchetti A, Parmanto B, Vecchio M, Ahmad S, Buch S, Zgheib NK, et al. Team building: electronic management-clinical translational research (eM-CTR) systems. Clin Transl Sci 2009 Dec;2(6):449-455 [FREE Full text] [doi: 10.1111/j.1752-8062.2009.00157.x] [Medline: 20443940]

38. Weber SC, Lowe H, Das A, Ferris T. A simple heuristic for blindfolded record linkage. J Am Med Inform Assoc 2012 Jun;19(e1):e157-e161 [FREE Full text] [doi: 10.1136/amiajnl-2011-000329] [Medline: 22298567]
39. Palma G. Electronic Health Records: The Good, the Bad and the Ugly. Becker's Health IT. 2013 Oct 14. URL: https://www.beckershospitalreview.com/healthcare-information-technology/electronic-health-records-the-good-the-bad-and-the-ugly.html [accessed 2020-09-25]
40. Kaufman KR, Hyler SE. Problems with the electronic medical record in clinical psychiatry: a hidden cost. J Psychiatr Pract 2005 May;11(3):200-204. [doi: 10.1097/00131746-200505000-00008] [Medline: 15920394]
41. Shin EY, Ochuko P, Bhatt K, Howard B, McGorisk G, Delaney L, et al. Errors in electronic health record-based data query of statin prescriptions in patients with coronary artery disease in a large, academic, multispecialty clinic practice. J Am Heart Assoc 2018 Apr 12;7(8). [doi: 10.1161/JAHA.117.007762] [Medline: 29650707]
42. Goodloe R, Farber-Eger E, Boston J, Crawford DC, Bush WS. Reducing clinical noise for body mass index measures due to unit and transcription errors in the electronic health record. American Medical Informatics Association Jt Summits Transl Sci Proc 2017 Jul 26;2017:102-111 [FREE Full text] [Medline: 28815116]
43. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap): a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
44. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
45. Harris PA. Research electronic data capture (REDCap) - planning, collecting and managing data for clinical and translational research. BioMed Central Bioinformatics 2012 Jul 31;13(S12). [doi: 10.1186/1471-2105-13-s12-a15]
46. Cecchetti AA. Why Introduce Machine Learning To Rural Health Care? Marshall Journal of Medicine 2018 Apr;4(2) [FREE Full text] [doi: 10.18590/mjm.2018.vol4.iss2.2]
47. McDonald HI, Shaw C, Thomas SL, Mansfield KE, Tomlinson LA, Nitsch D. Methodological challenges when carrying out research on CKD and AKI using routine electronic health records. Kidney Int 2016 Nov;90(5):943-949 [FREE Full text] [doi: 10.1016/j.kint.2016.04.010] [Medline: 27317356]
48. Pierce C, Scherra E. The Challenges of Data Collection in Rural Dwelling Samples. OJRNHC 2004 Dec;4(2):25-30. [doi: 10.14574/ojrnhc.v4i2.197]
49. Verby JE. Patients' and physicians' views of health in a rural area. Acad Med 1989 Nov;64(11):665-666. [doi: 10.1097/00001888-198911000-00009] [Medline: 2803426]
50. Hendryx M, Ahern MM, Nurkiewicz TR. Hospitalization patterns associated with Appalachian coal mining. J Toxicol Environ Health A 2007 Dec;70(24):2064-2070. [doi: 10.1080/15287390701601236] [Medline: 18049995]
51. Ortmeyer CE, Costello J, Morgan WK, Swecker S, Peterson M. The mortality of Appalachian coal miners, 1963 to 1971. Arch Environ Health 1974 Aug;29(2):67-72. [doi: 10.1080/00039896.1974.10666535] [Medline: 4835173]
52. Amro A, Baez GA, Koromia GA, Bhardwaj N, Aguilar R, El-Hamdani M, et al. Albumin level as a risk marker and predictor of peripartum cardiomyopathy. Journal of the American College of Cardiology 2019 Mar;73(9):835 [FREE Full text] [doi: 10.1016/s0735-1097(19)31442-1]
53. Acosta G, Amro A, Aguilar R, Abusnina W, Bhardwaj N, Koromia GA, et al. Clinical determinants of myocardial injury, detectable and serial troponin levels among patients with hypertensive crisis. Cureus 2020 Jan 27;12(1):e6787 [FREE Full text] [doi: 10.7759/cureus.6787] [Medline: 32140347]
54. Elmore D, Yaslam B, Putty K, Magrane T, Abadir A, Bhatt S, et al. Is Fever a Red Flag for Bacterial Pneumonia in Children With Viral Bronchiolitis? Glob Pediatr Health 2019;6:2333794X19868660. [doi: 10.1177/2333794X19868660] [Medline: 31431903]
55. Bhardwaj N, Sundaram S, Carter L, Cecchetti A, Sundaram U. Tu1801 metabolic syndrome: are current colon cancer screening guidelines enough in a rural population? Gastroenterology 2020 May;158(6):S-1167 [FREE Full text] [doi: 10.1016/s0016-5085(20)33589-7]
56. Sundaram S, Bhardwaj N, Gress T, Cecchetti A. Utilization of Appalachian Clinical and Translational Science Institute Data Warehouse to more accurately predict disease processes important for central Appalachia. In: 12th Annual CCTS Spring Conference.: University of Kentucky; 2017 Presented at: CCTS'17; March 30, 2017; Lexington, KY URL: https://www.ccts.uky.edu/media/913
57. Bhardwaj N, Cecchetti AA, Murughiyan U, Neitch S. Analysis of Benzodiazepine prescription practices in elderly Appalachians with dementia via the Appalachian informatics platform: longitudinal study. J Med Internet Res Med Inform 2020 Aug 4;8(8):e18389 [FREE Full text] [doi: 10.2196/18389] [Medline: 32749226]
58. Khanna R, Gress T, Cecchetti A. Hospital Emergency Department Visits For Non-Traumatic Oral Health Conditions. National Oral Health Conference. 2018. URL: http://www.nationaloralhealthconference.com/pdfs/2018-poster-abstracts.pdf [accessed 2020-09-29]
59. Ferdjallah M, Driscoll H. Serum Calcium Homeostasis and Volume Dynamics in Alzheimer's Disease and Diabetes Mellitus-2. The Health Science Center 32nd Annual Research Day at Marshall University. 2020. URL: https://jcesom.marshall.edu/media/58548/9110_researchsyllabus_2020.pdf [accessed 2020-09-29]

60. Stats of the State of West Virginia. Centers for Disease Control and Prevention. 2017. URL: https://www.cdc.gov/nchs/pressroom/states/westvirginia/westvirginia.htm [accessed 2020-09-25]
61. Patil P, Kinariwala S. Automated Diagnosis of Heart Disease using Random Forest Algorithm. GitHub. 2017. URL: https://github.com/mbbrigitte/Predicting_heart_disease_UCI/blob/master/heartdisease_UCI.Rmd [accessed 2020-09-25]
62. Sreejith S, Rahul S, Jisha R. A real time patient monitoring system for heart disease prediction using random forest algorithm. In: Advances in Signal Processing and Intelligent Recognition Systems. Cham, UK: Springer; Dec 25, 2015:485-500.
63. Tanaka T, Voigt MD. Decision tree analysis to stratify risk of de novo non-melanoma skin cancer following liver transplantation. J Cancer Res Clin Oncol 2018 Mar;144(3):607-615. [doi: 10.1007/s00432-018-2589-5] [Medline: 29362916]
64. Paxton R, Zhang L, Wei C, Price D, Zhang F, Courneya KS, et al. An exploratory decision tree analysis to predict physical activity compliance rates in breast cancer survivors. Ethn Health 2019 Oct;24(7):754-766. [doi: 10.1080/13557858.2017.1378805] [Medline: 28922931]
65. Moreno HA. Predicting car accidents in Barcelona using a Random Forest model. Universitat Politècnica de Catalunya. 2017 Jan. URL: http://hdl.handle.net/2117/100298 [accessed 2020-09-25]
66. Xiaohui J. Forecast model of road traffic accidents based on LS-SVM with grey correlation analysis. Appl Res Comput 2016;3:038 [FREE Full text]
67. Khatri KL, Tamil LS. Early detection of peak demand days of chronic respiratory diseases emergency department visits using artificial neural networks. IEEE J Biomed Health Inform 2018 Jan;22(1):285-290. [doi: 10.1109/jbhi.2017.2698418]
68. Oijuela-Canon AD, Gomez-Cajas DF, Sepulveda-Sepulveda A. Respiratory Diseases Discrimination Based on Acoustic Lung Signals and Neural Networks. In: 20th Symposium on Signal Processing, Images and Computer Vision. 2015 Presented at: STSIVA'15; September 2-4, 2015; Bogota, Colombia. [doi: 10.1109/stsiva.2015.7330461]
69. McKinley R, Häni L, Gralla J, El-Koussy M, Bauer S, Arnold M, et al. Fully automated stroke tissue estimation using random forest classifiers (FASTER). J Cereb Blood Flow Metab 2017 Aug;37(8):2728-2741 [FREE Full text] [doi: 10.1177/0271678X16674221] [Medline: 27798267]
70. Chen L, Bentley P, Rueckert D. A novel framework for sub-acute stroke lesion segmentation based on random forest. In: Ischemic Stroke Lesion Segmentation. 2015 Presented at: ISLES'15; October 5, 2015; Munich, Germany p. 17-20 URL: http://www.isles-challenge.org/ISLES2015/pdf/20150930_ISLES2015_Proceedings.pdf#page=17
71. Xu W, Zhang J, Zhang Q, Wei X. Risk Prediction of Type II Diabetes Based on Random Forest Model. In: Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics. 2017 Presented at: AEEICB'17; February 27-28, 2017; Chennai, India. [doi: 10.1109/aeeicb.2017.7972337]
72. Shukla N, Arora M. International Journal of Computer Sciences and Engineering 2016;4(7):101-104.
73. Basaia S, Agosta F, Wagner L, Canu E, Magnani G, Santangelo R, Alzheimer's Disease Neuroimaging Initiative. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. Neuroimage Clin 2019;21:101645 [FREE Full text] [doi: 10.1016/j.nicl.2018.101645] [Medline: 30584016]
74. Lu D, Popuri K, Ding GW, Balachandar R, Beg MF, Alzheimer's Disease Neuroimaging Initiative. Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural mr and FDG-PET images. Sci Rep 2018 Apr 9;8(1):5697 [FREE Full text] [doi: 10.1038/s41598-018-22871-z] [Medline: 29632364]
75. Wiemken TL, Furmanek SP, Mattingly WA, Guinn BE, Cavallazzi R, Fernandez-Botran R, et al. Predicting 30-day mortality in hospitalized patients with community-acquired pneumonia using statistical and machine learning approaches. The University of Louisville Journal of Respiratory Infections 2017;1(3) [FREE Full text] [doi: 10.18297/jri/vol1/iss3/10/]
76. Menéndez Villanueva R. [The diagnostic evaluation of rapid sputum technics for Pneumococcus in community-acquired pneumonia. The usefulness of Bayes theorem for clinical application]. Arch Bronconeumol 1995;31(7):317-322. [Medline: 8777525]
77. B.V R, Sriraam N, Geetha M. Classification of non-chronic and chronic kidney disease using SVM neural networks. 2017 Dec 31;7(1.3):191-194 [FREE Full text] [doi: 10.14419/ijet.v7i1.3.10669]
78. Annapoorani J, Gnanaselvam C. Enhancing prediction accuracy of chronic kidney disease using neural networks. Automation and Autonomous System 2018;10(1):10-15. [doi: 10.36039/AA012018003]
79. Ayat S, Farahani HA, Aghamohamadi M, Alian M, Aghamohamadi S, Kazemi Z. A comparison of artificial neural networks learning algorithms in predicting tendency for suicide. 2012 Jul 26;23(5):1381-1386 [FREE Full text] [doi: 10.1007/s00521-012-1086-z]
80. Bhat H, Goldman-Mellor S. Predicting adolescent suicide attempts with neural networks. arXiv 2017 Dec 01:- epub ahead of print [FREE Full text]
81. Wolfram W, Christian D. Appalachian speech. Center for Applied Linguistics. 1976. URL: http://files.eric.ed.gov/fulltext/ED130511.pdf [accessed 2020-09-25]
82. Luhman R. Appalachian english stereotypes: language attitudes in Kentucky. Lang Soc 2008 Dec 18;19(3):331-348. [doi: 10.1017/s0047404500014548]

Abbreviations

ACTSI: Appalachian Clinical and Translational Science Institute

CCI: Charlson Comorbidity Index
CDW: clinical data warehouse
CHH: Cabell-Huntington Hospital
ECCC: Edwards Comprehensive Cancer Institute
EHR: electronic health record
ETL: extract, transform, and load
HIPAA: Health Insurance Portability and Accountability Act
MH: Marshall Health
MU JCESOM: Marshall University Joan C Edwards School of Medicine
OLAP: Online Analytical Processing
PHI: protected health information
SQL: structured query language

Edited by G Eysenbach; submitted 24.01.20; peer-reviewed by T Muto, B Fan, N Mohammad Gholi Mezerji; comments to author 12.05.20; revised version received 24.07.20; accepted 27.07.20; published 14.10.20.

Please cite as:
Cecchetti AA, Bhardwaj N, Murughiyan U, Kothakapu G, Sundaram U
Fueling Clinical and Translational Research in Appalachia: Informatics Platform Approach
JMIR Med Inform 2020;8(10):e17962
URL: http://medinform.jmir.org/2020/10/e17962/
doi:10.2196/17962
PMID:33052114

Original Paper

Phenotypically Similar Rare Disease Identification from an Integrative Knowledge Graph for Data Harmonization: Preliminary Study

Qian Zhu[1], PhD; Dac-Trung Nguyen[1], MS; Gioconda Alyea[2], MS, MD; Karen Hanson[2], MS, MBA; Eric Sid[3], MD, MHA; Anne Pariser[3], MD

[1]Division of Pre-Clinical Innovation, National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Rockville, MD, United States

[2]ICF International Inc, Rockville, MD, United States

[3]Office of Rare Diseases Research (ORDR), National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Bethesda, MD, United States

Corresponding Author:
Qian Zhu, PhD
Division of Pre-Clinical Innovation
National Center for Advancing Translational Sciences (NCATS)
National Institutes of Health (NIH)
9800 Medical Center Drive
Rockville, MD, 20850
United States
Phone: 1 301 4807841
Email: qian.zhu@nih.gov

Abstract

Background: Although many efforts have been made to develop comprehensive disease resources that capture rare disease information for the purpose of clinical decision making and education, there is no standardized protocol for defining and harmonizing rare diseases across multiple resources. This introduces data redundancy and inconsistency that may ultimately increase confusion and difficulty for the wide use of these resources. To overcome such encumbrances, we report our preliminary study to identify phenotypical similarity among genetic and rare diseases (GARD) that are presenting similar clinical manifestations, and support further data harmonization.

Objective: To support rare disease data harmonization, we aim to systematically identify phenotypically similar GARD diseases from a disease-oriented integrative knowledge graph and determine their similarity types.

Methods: We identified phenotypically similar GARD diseases programmatically with 2 methods: (1) We measured disease similarity by comparing disease mappings between GARD and other rare disease resources, incorporating manual assessment; 2) we derived clinical manifestations presenting among sibling diseases from disease classifications and prioritized the identified similar diseases based on their phenotypes and genotypes.

Results: For disease similarity comparison, approximately 87% (341/392) identified, phenotypically similar disease pairs were validated; 80% (271/392) of these disease pairs were accurately identified as phenotypically similar based on similarity score. The evaluation result shows a high precision (94%) and a satisfactory quality (86% F measure). By deriving phenotypical similarity from Monarch Disease Ontology (MONDO) and Orphanet disease classification trees, we identified a total of 360 disease pairs with at least 1 shared clinical phenotype and gene, which were applied for prioritizing clinical relevance. A total of 662 phenotypically similar disease pairs were identified and will be applied for GARD data harmonization.

Conclusions: We successfully identified phenotypically similar rare diseases among the GARD diseases via 2 approaches, disease mapping comparison and phenotypical similarity derivation from disease classification systems. The results will not only direct GARD data harmonization in expanding translational science research but will also accelerate data transparency and consistency across different disease resources and terminologies, helping to build a robust and up-to-date knowledge resource on rare diseases.

(JMIR Med Inform 2020;8(10):e18395) doi:10.2196/18395

KEYWORDS

GARD; rare diseases; phenotypical similarity; data harmonization

Introduction

A rare disease in the United States is defined by the 1983 Orphan Drug Act as a condition that affects fewer than 200,000 people [1], whereas the analogous legislation introduced in the European Union in 2000 considers a disease to be rare when it affects fewer than 1 in 2,000 people [2]. In comparison to common diseases, health care providers are challenged by a lack of familiarity with diagnosing and treating rare diseases, which can lead to missed, delayed, or inaccurate diagnoses even when an approved, effective therapy is available [3]. Improved understanding and recognition of rare diseases are key for accurate and timely diagnosis, and this relies on broad dissemination of and access to knowledge about rare diseases [4]. A huge amount of effort has been made to develop rare disease resources for patients, families, and clinicians, such as the Genetic and Rare Diseases Information Center (GARD) [5], Orphanet [6], and Monarch Disease Ontology (MONDO) [7]; however, disparate data and incomplete data harmonization are still major barriers to improved coordination across specialists, leading to inefficiencies and delays in diagnosis, care, and treatment. This is exemplified by the difficulty faced in accurately answering the question, *how many total rare diseases are there?* A recent report by Haendel et al [8], after an examination of multiple rare resources, concluded that "there are total of 10,393 rare diseases in MONDO…the majority, 6370 rare diseases, are presented in three or more resources, whereas 4023 are unique to one source." The fact that more than one-third of rare diseases are unique to 1 source highlights a reality that those resources continue to use their own disease definitions or harmonization rules to develop their rare disease vocabularies. Insufficient effort put toward data harmonization ultimately leads to redundancy in categorization efforts and a resulting inconsistency of rare disease representation globally.

The goal of data harmonization is to improve the compatibility of data collected from independent sources (horizontally) in order to better understand disease etiology from different angles, which may forward the discovery of therapeutic approaches for rare diseases. For each individual source, data harmonization is crucial to better represent and organize data for supporting data harmonization horizontally. Current data harmonization efforts are primarily aligning standard nomenclatures or human efforts to translate specific medical and clinical features into a standardized and sharable format. For instance, Pontikos et al [9] introduced Phenooplis, an open platform for the harmonization and analysis of genetic and phenotypic data that harmonize phenotypes with the help of Human Phenotype Ontology (HPO). The International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) invited the cancer-genomics and bioinformatics communities to work together to identify the best pipelines for the detection of mutations in DNA-sequencing reads for cancer genomes in order to facilitate the harmonization of mutation-calling procedures among institutions [10,11]. Orphanet and OMIM (Online Mendelian Inheritance of Man) heavily relied on human efforts for their data curation and harmonization [12,13]. To avoid cumbersome human efforts and a lack of rare disease standards in this study, we proposed to systematically identify phenotypically similar rare diseases from GARD and determine their similarity types, including duplicate diseases, sibling diseases, and subtypes for supporting rare disease data harmonization.

Rare disease designations are often in conflict across different datasets due to the differing statutory requirements used in defining a rare disease in different countries, and as such, useful methods to improve interoperability across these broad terminologies and standards are required. With the aim of eliminating data redundancy and inconsistency across different resources, improving data interoperability, and facilitating data harmonization, the implementation of a knowledge graph is capable of semantically organizing and integrating complex networks of data into one collection. Knowledge graphs have been widely applied in the medical domain and in the rare disease field. For instance, Reumann et al [14] reported their solution for cognitive differential diagnosis (DDx) in rare diseases based on knowledge graph technology that incorporates data from ICD-10, DOID, medDRA, PubMed, Wikipedia, Orphanet, the CDC, and anonymized patient data. Li et al [15] presented their work to develop a rare disease classification algorithm established on a knowledge graph. Sosa et al [16] applied a knowledge graph–embedding method that explicitly models the uncertainty associated with literature-derived relationships and uses link prediction to generate drug repurposing hypotheses for rare diseases. In this study, we accessed data from an integrative knowledge graph that we developed from our previous study [17] with a variety of rare disease-related resources for phenotypical similarity identification among GARD diseases.

In this study, we report our preliminary work to identify phenotypically similar GARD diseases from an integrative knowledge graph using 2 approaches: (1) disease mapping comparison, and (2) phenotypical similarity derivation from disease classification systems. This effort will not only direct GARD data harmonization but will also support data harmonization across different resources, and eventually support clinical decision making. Phenotypically similar GARD diseases applied in this study specifically refer to disease subtypes and sibling diseases that share similar clinical manifestations. For example, 2 GARD diseases of "lactate dehydrogenase deficiency" and "lactate dehydrogenase A deficiency" are subtypes, and they have similar phenotypical profiles.

Background and Materials

Rare Disease Resources

The Genetic and Rare Diseases Information Center (GARD) is a program managed by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH). Since 2003, GARD has provided the public with access to current, reliable, and easy-to-understand information about

rare and genetic diseases [5]. As part of the data harmonization effort toward furthering the development of the GARD, we harmonized GARD diseases according to their phenotypical similarity in this study. To fulfill this task, we assessed phenotypical similarity among GARD diseases by leveraging several well-known disease resources, including Orphanet, OMIM, MONDO, the HPO, and the UMLS (Unified Medical Language System), owing to their complementary focus and coverage. We briefly describe these applied resources below.

Orphanet is an EU resource that focuses on gathering and improving knowledge on rare diseases [6]. Rare diseases in the Orphanet, depending on their clinical presentation, are included in as many classifications as needed. The Orphanet classification is organized according to three hierarchical levels: group of disorders, disorder, and subtype of a disorder. The disorder level is designated as the main topologic level for each clinical entity characterized by a set of homogeneous phenotypic abnormalities and evolution, allowing for a definitive clinical diagnosis [18,19].

OMIM (Online Mendelian Inheritance in Man**)** is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. It contains information on all known mendelian disorders and over 15,000 genes. OMIM focuses on the relationship between phenotype and genotype [20].

MONDO (Monarch Disease Ontology) aims to harmonize disease definitions across the world. It is a semi-automatically constructed ontology that merges multiple disease resources to yield a coherent merged ontology. One feature of the MONDO is that it curates precise 1-to-1 equivalence axioms connecting to other resources, validated by OWL reasoning [7]. MONDO provides a hierarchical structure that can be used for classification or for rolling up diseases to higher-level groupings.

The Human Phenotype Ontology (HPO) provides a standardized vocabulary of phenotypic abnormalities encountered in human disease. HPO currently contains over 13,000 terms and over 156,000 annotations to hereditary diseases [21].

The Unified Medical Language System (UMLS) is a terminology integration system developed at the National Library of Medicine (NLM). The UMLS Metathesaurus integrates more than 160 biomedical vocabularies. Synonymous terms from the various source vocabularies are grouped into one concept [22].

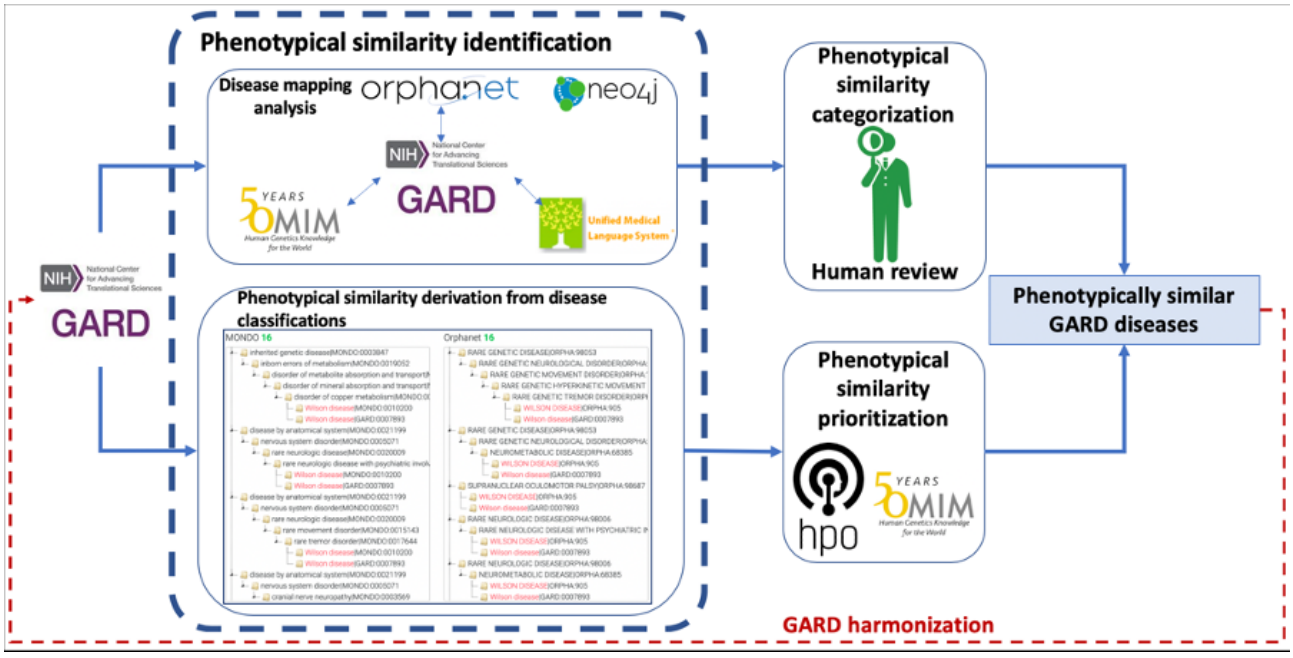
An Integrative Knowledge Graph

We previously developed an integrative knowledge graph with 34 different biomedical data resources at the time of writing, including the aforementioned resources. This graph database is hosted in Neo4j and is publicly accessible without login credentials [17]. In this study, we accessed this knowledge graph to obtain data from the aforementioned resources and applied it for the measurement of phenotypical similarity among GARD diseases.

Methods

In this study, we aimed to identify phenotypical similarity among rare diseases to support data harmonization and data interoperability with existing standardized terminologies and ontologies. We designed two complementary approaches: (1) analysis of disease mappings to Orphanet, OMIM, and the UMLS to measure phenotypical similarity among GARD diseases; (2) prioritizing phenotypical similarity derived from MONDO and Orphanet disease classification systems with shared phenotypes from the HPO and genes from OMIM. The architecture of this study is shown in Figure 1.

Figure 1. The architecture of phenotypically similar GARD disease identification.



Phenotypical Similarity Identification Based on GARD Disease Mappings

In order to identify phenotypical similarity, we computed disease similarity among disease mappings between GARD diseases and disease concepts from Orphanet, OMIM, and the UMLS, which offer a wide spectrum of characteristics of rare diseases—in Orphanet, diseases are defined upon their clinical presentation; in OMIM, disease definition is based on genetic etiology; in UMLS, a broader biomedical definition of diseases is offered.

Disease Mapping Retrieval from the Knowledge Graph

We obtained disease mappings from the aforementioned knowledge graph. There are 2 ways to retrieve disease mappings for GARD diseases from the knowledge graph: (1) by developing mappings based on specific edge properties; for instance, 2 concepts with the same concept names are mapped via one edge property of "N_Name"; (2) by extracting mappings directly from GARD disease nodes, which store GARD-curated external mappings to Orphanet, OMIM, and the UMLS. To ensure mapping quality, we performed the second approach by accessing 1 node property of I_CODE and storing external mappings for each GARD disease node. For instance, 3 external mappings, including "OMIM:603358," "ORPHANET:53693," and "UMLS:C1864002" for the GARD disease of "GRACILE SYNDROME(GARD:0000001)," are stored in its property of "I_CODE" and can be retrieved by executing the following Cypher Query 1 [23], which is Neo4j's graph query language that allows users to store and retrieve data from the graph database:

Cypher Query 1. match P = (n:S_GARD[a]) where any (x in n.I_CODE where x= "GARD:0000001") return n.I_CODE

[a]S_GARD referring to GARD data

We executed the Cypher Queries listed in Table 1 to retrieve disease mappings for GARD diseases. Each GARD disease obtains zero to multiple mappings accordingly. For instance, "Gracile Syndrome (GARD:0000001)" has the 3 disease mappings described above; however, "Acalvaria (GARD:0000361)" only has 1 mapping, "ORPHANET:945." To ensure that each GARD disease was associated with at least 1 mapping for similarity measurement, we excluded 1498 GARD diseases with no mappings to any of these 3 resources.

Table 1. Disease mapping extraction from the Neo4j knowledge graph.

Disease mappings	Cypher Queries
GARD2Orphanet	match P = (n:S_GARD) where any (x in n.I_CODE where x=~ 'ORPHA.*')return distinct n.I_CODE
GARD2OMIM	match P = (n:S_GARD) where any (x in n.I_CODE where x=~ 'OMIM.*')return distinct n.I_CODE
GARD2UMLS	match P = (n:S_GARD) where any (x in n.I_CODE where x=~ UMLS.*')return distinct n.I_CODE

Calculating Similarity to Prioritize Phenotypical Similarity of GARD Disease Pairs

In order to compare phenotypical similarity among the GARD diseases based on their similarity, we enumerated all mappings obtained for 5236 GARD diseases and ended with a total of 9672 mappings. For each GARD disease, we generated fingerprints based on those mappings. One disease mapping corresponding to one binary fingerprint, with presence denoted as 1 and absence denoted as 0. To this end, each GARD disease was represented as a vector of 9672 bits. Then, we calculated cosine similarity [24] for each pair of GARD diseases based on their fingerprints. For those disease pairs without any shared mappings, which means their similarity score equals 0, we excluded them for manual similarity identification.

Phenotypically Similar GARD Disease Identification

To determine the phenotypical similarity of GARD diseases, our subject matter experts (GA, KH, and ES) manually evaluated the prioritized disease pairs based on their similarity scores generated from the above step. The manual validation was not only attempting to examine the accuracy of computational results to establish business rules for further GARD data harmonization, but also to validate correctness and coverage of the GARD-curated external mappings.

The manual review process consisted of 3 steps: (1) categorizing GARD disease pairs to phenotypical similarity types, namely "Duplicates," "Subtypes," "Siblings," and "Unrelated;" (2) researching the latest epidemiology studies (eg, PubMed articles, trusted resources) for each disease if applicable, to re-evaluate the qualification of RARE disease based on the US definition of rare disease [1]; (3) documenting the decision-making process for future reference. As an example demonstrating this review process, "Testicular Cancer (GARD:0007746)" and "Testicular germ cell tumor (GARD:0013047)," with a similarity score of 0.71, were initially grouped as subtypes. However, researching the latest epidemiological data for testicular cancer uncovered that "in 2017, there were an estimated 269,769 men living with testicular cancer in the United States" [25]; this indicates that the prevalence rate of testicular cancer does not meet (ie, is higher than) the US definition of rare diseases, and so it was marked to "Retire."

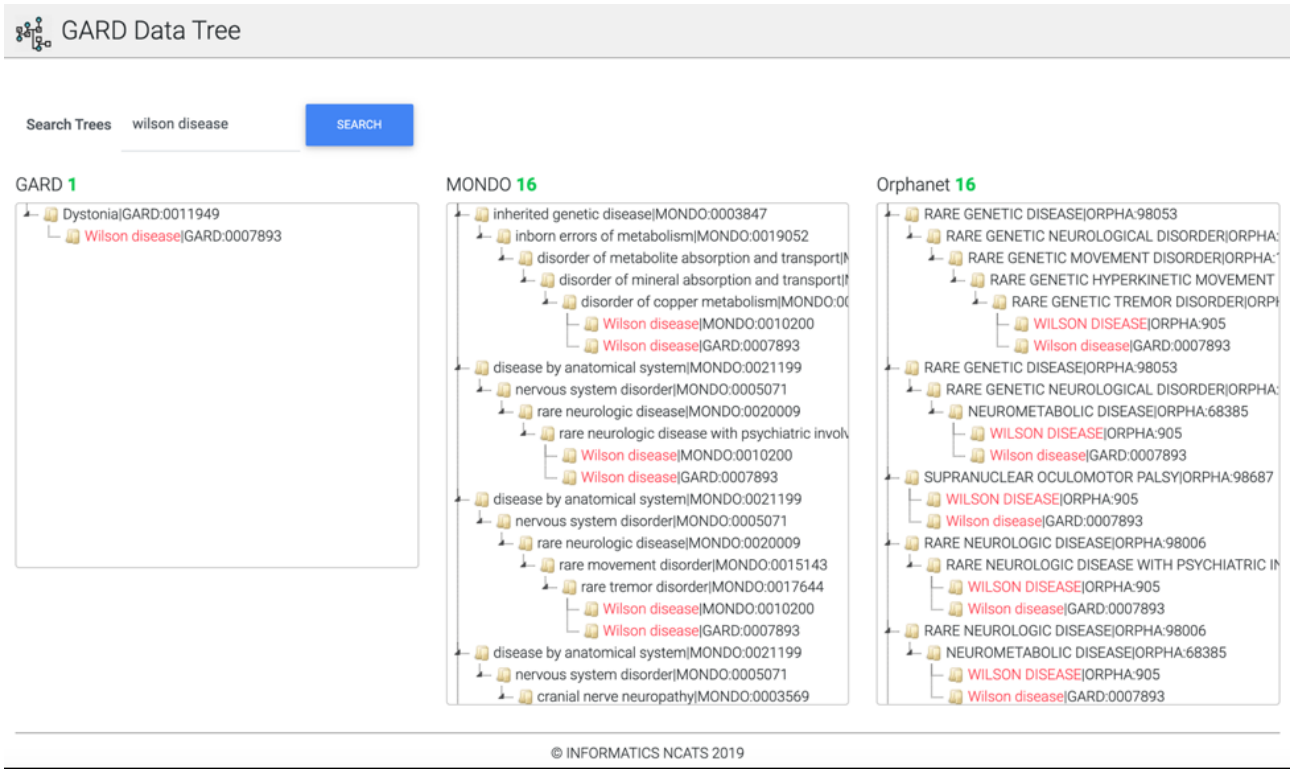
In this context, we defined *precision* as the fraction between the number of correctly identified phenotypically similar disease pairs based on manual evaluation and the total number of similar disease pairs identified; we defined *recall* as the fraction between the number of correctly identified phenotypically similar disease pairs and the total number of similar disease pairs; we defined *F measure* as the balanced harmonic mean of the precision and recall. We computed precision, recall, and F measure to measure the performance of this approach.

Phenotypical Similarity Derivation from Disease Classification Systems

Diseases from the same disease category exhibit a high phenotypic homogeneity [26]; we assume that phenotypical

similarity is evidently presenting among sibling diseases, which share the same parent diseases in disease classification systems. To further prove our assumption by assessing 3 disease classification systems, including GARD, MONDO, and Orphanet, we developed a web application to search and review a specific disease term presenting in these 3 disease trees to perform a comparison. This web application is publicly accessible [27]. Figure 2 shows one screenshot of the search results for "Wilson disease." MONDO and Orphanet have more refined and complete disease classifications than the GARD, which enables phenotypical similarity identification for GARD diseases.

Figure 2. Disease tree visualization via the GARD Data Tree web tool.

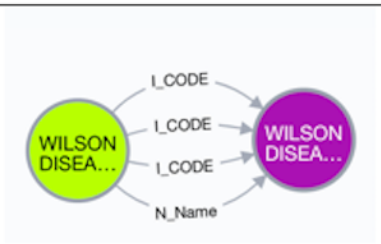


Retrieving Phenotypically Similar GARD Diseases

With the help of the GARD Data Tree web tool, we were able to form a process of deriving phenotypical similarity among GARD diseases in 3 steps: (1) mapping GARD diseases to MONDO and Orphanet; (2) extracting all sibling diseases of the mapped MONDO and Orphanet diseases from their disease trees; and (3) mapping the retrieved sibling diseases back to the GARD. The GARD diseases retrieved from the third step should be phenotypically similar to the query GARD disease from the first step. We further validated them by leveraging their associated phenotypes and genotypes.

These 3 steps can be formalized in Cypher Queries accordingly; examples are shown in Figure 3. After obtaining mappings between GARD and Orphanet/MONDO by executing Cypher Query 1 shown in Figure 3, we searched parent diseases of those mapped MONDO and Orphanet diseases. Cypher Query 2 is an example of extracting Orphanet parent diseases for the Orphanet concept "Wilson Disease (ORPHA:905)," which is mapped to "GARD:0007893" from Cypher Query 1. Cypher Query 3 demonstrates a process that extracts all child diseases for 1 Orphanet parent disease, "SUPRANUCLEAR EYE MOVEMENT DISORDER (ORPHANET:98687)," which is 1 parent node returned from Cypher Query 2, and maps those child Orphanet diseases to GARD diseases. In order to identify the most phenotypically similar GARD diseases obtained from Cypher Query 3 to the inquiry disease "Wilson Disease (GARD:0007893)," we prioritized similarity based on their associated phenotypes and genes.

Figure 3. Cypher query examples for extracting phenotypically similar GARD diseases by navigating Orphanet disease classification systems.

Cypher Queries	Explanations	Results
Cypher Query 1. match p=(n:S_GARD)<--(d:DATA) where d.gard_id = 'GARD:0007893' with n match p=(n)-[:I_CODE|:N_Name]-(m:S_ORDO_ORPHANET) where not m:TRANSIENT return p	Retrieving mappings to Orphanet for GARD diseases	
Cypher Query 2. match p=(m:S_ORDO_ORPHANET)-[:R_subClassOf]->(m1:S_ORDO_ORPHANET) where any(x in m.I_CODE where x= 'ORPHA:905') AND not m:TRANSIENT and not m1:TRANSIENT return p	Retrieving patent diseases for the Orphanet disease ('ORPHA:905') from Query 1	
Cypher Query 3. match p = (o:S_ORDO_ORPHANET)<-[:R_subClassOf]-(i:S_ORDO_ORPHANET) where o.I_CODE = 'ORPHA:98687' with i, o match p1 = (i)-[:I_CODE|:N_Name]-(n:S_GARD) return p1, i, o	Retrieving GARD disease mapped to the child Orphanet diseases for each of the sixteen parent Orphanet Diseases from Query 2	

Prioritizing Phenotypically Similar GARD Diseases Based on Phenotypes and Genotypes

Given the fact that a majority of rare diseases are genetic in origin and that clinical phenotypes are one of the red flags increasing rare disease attentiveness in clinical practice [28], we developed a protocol for prioritizing phenotypical similarity based on phenotypes and genotypes. We collected phenotypes from the HPO and genes from OMIM from our knowledge graph, for those similar GARD disease pairs identified from the above step. The number of phenotypes and genes shared by each pair of phenotypically similar GARD diseases was applied for prioritization.

Results

Results of Disease Mapping Analysis

Disease Concept Retrieval

We extracted disease mappings between GARD and Orphanet, OMIM, and the UMLS from our Neo4j knowledge graph. The retrieval results are shown in Table 2.

Table 2. Results of disease mapping retrieval from Neo4j graph.

Types of mapping	Number of mappings
GARD2Orphanet	2,869
GARD2OMIM	3,500
GARD2UMLS	3,584

Disease Similarity Calculation

We enumerated disease pairs for 5236 GARD diseases with disease mappings and calculated cosine similarity for those GARD pairs. After excluding those disease pairs with similarity equaling 0, 392 diseases pairs remained. Table 3 summarizes the results of the similarity calculation.

Table 3. Similarity calculation results for disease pairs (n=392).

Similarity scores	Number of disease pairs
1	34
0.5 <= Similarity < 1	264
0 < Similarity < 0.5	94

Evaluation and Disease Similarity Identification

Our subject matter experts manually reviewed these 392 disease pairs and assigned their similarity types accordingly. Table 4 shows their review results.

Of the 392 disease pairs, 341 (87%) were identified and categorized as phenotypically similar, corresponding to the categories “Duplicated,” “Siblings,” and “Subtypes.” Of those 341 disease pairs, 271 disease pairs (80%) with similarity scores greater than 0.5 were verified as phenotypically similar. However, 34 disease pairs were determined to be “Unrelated,” and another 17 disease pairs were “Ungrouped;” this needs further discussion, and so we excluded the latter group for calculations of precision, recall, and F measure.

Table 4. Manual review results for the disease pairs (n=392); precision=94%, recall=79%, F measure=86%.

Variables	Phenotypical similarity types				
	Duplicated	Siblings	Subtypes	Unrelated	Ungrouped
Number of disease pairs, n					
Phenotypically similar (n=341)	105	117	119	N/A[a]	N/A
Not phenotypically similar (n=51)	N/A	N/A	N/A	34	17
Similarity scores, n					
Phenotypically similar (n=341)					
0.7≥Score≥1 (n=95)	47	21	27	N/A	N/A
0.5≥Score≥0.7 (n=176)	42	81	53	N/A	N/A
Score>0.5 (n=70)	16	15	39	N/A	N/A
Not phenotypically similar (n=51)					
0.7≥Score≥1 (n=16)	N/A	N/A	N/A	8	8
0.5≥Score≥0.7 (n=15)	N/A	N/A	N/A	8	7
Score>0.5 (n=20)	N/A	N/A	N/A	18	2

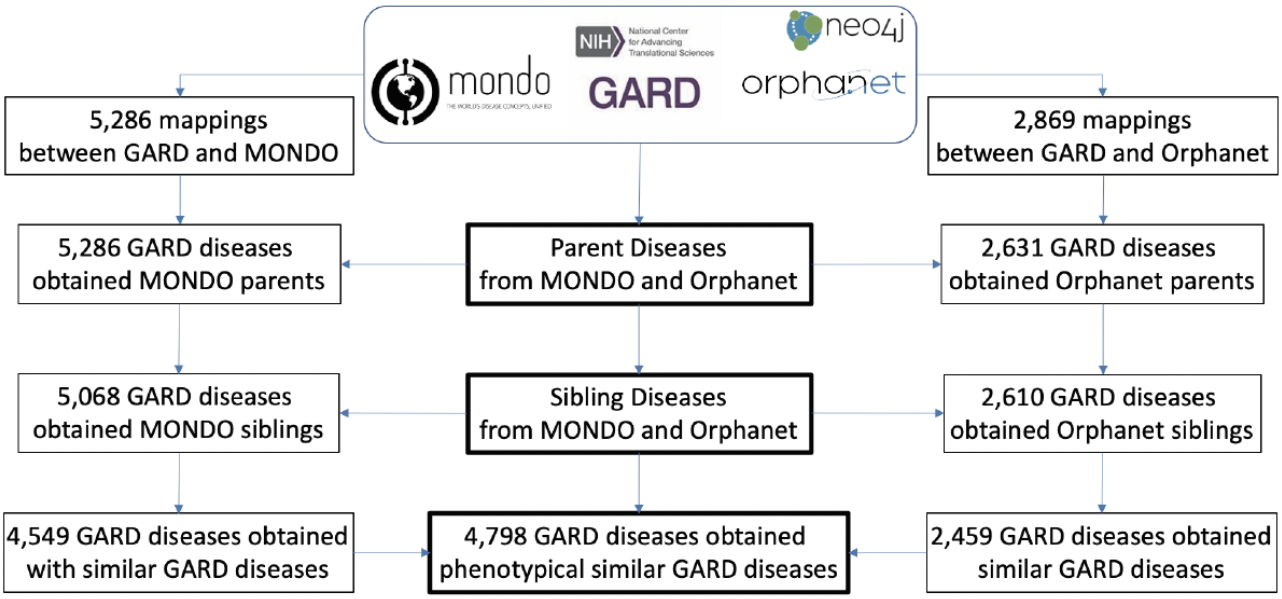
[a]N/A: not applicable.

Results of Phenotypical Similarity Derivation from Disease Classification Systems

Based on the above analysis, 53 GARD diseases were marked for retirement. Of the remaining of 5955 GARD diseases, 4798 GARD diseases obtained 1 or more phenotypically similar GARD disease(s) from this step. The stepwise results are shown in Figure 4.

Of 5286 GARD diseases mapped to one of 21,823 MONDO diseases with parent diseases, 4549 GARD diseases obtained phenotypically similar GARD diseases via MONDO sibling disease mappings. Of 2631 GARD diseases mapped to one of 7024 Orphanet diseases with parent diseases, 2459 GARD diseases obtained phenotypically similar GARD diseases via Orphanet sibling disease mappings. By combining these 2 lists of mappings, 4798 GARD diseases obtained phenotypically similar diseases. We paired these 4798 GARD diseases with identified phenotypically similar diseases and ended with unique 241,604 GARD disease pairs.

Figure 4. Results of phenotypically similar GARD disease retrieval based on MONDO and Orphanet disease classifications.



Phenotypically Similar Disease Prioritization Based on Phenotypes and Genotypes

Of the 241,604 disease pairs identified for these 4798 GARD diseases, 84,054 disease pairs shared at least 1 phenotype and 396 disease pairs shared at least 1 gene. By combing these 2 sets, there are 360 GARD disease pairs with at least 1 shared phenotype and gene. As all of those disease pairs were extracted from sibling diseases presenting in the MONDO and Orphanet, these 360 disease pairs were consequently grouped as "Siblings" with different degrees of phenotypical similarity based on the number of their shared phenotypes and genes.

By combining 341 disease pairs identified from the step of disease mapping analysis, 662 disease pairs showed phenotypical similarity. It is worth noting that there are 39 overlaps between these 2 sets. Based on the manual evaluation shown in Table 4, these 39 pairs consist of 25 disease pairs that are sibling diseases, 7 disease pairs that are subtypes, 2 pairs that are duplicates, and 5 pairs that are unrelated diseases.

Discussion

In this study, we identified and prioritized phenotypical similarity among GARD diseases by comparing disease similarity and deriving phenotypical similarity from disease classification systems. As a proof-of-concept, we demonstrated the usefulness of the identified phenotypically similar disease pairs to support data harmonization for GARD. By incorporating these identified similar diseases, GARD will have the capability of supporting education and clinical decision making; for instance, GARD can provide more complementary information not only for the inquiry disease but also for phenotypically similar diseases.

There are many different rare disease resources available, and each of them has their own strength and focus. OMIM classifies diseases based on their genetic cause, Orphanet defines rare diseases based on phenotypical characteristics, and UMLS incorporates biomedical vocabulary and standards to define their disease concepts. Given the complementary definition of disease concepts from these 3 resources, we employed their mappings to the GARD diseases for disease similarity comparison. Of the 392 disease pairs, 271 disease pairs (80%) with similarity scores greater than 0.5 were successfully validated as clinically relevant by our genetic specialists. Besides these true positives, feedback from our subject matter experts on the false positives [ie, 16 disease pairs (~4%) with similarity scores greater than 0.5 were manually determined as irrelevant] and false negatives [ie, 70 disease pairs (~18%) with similarity scores less than 0.5 were manually determined as relevant] illustrates that it is important to accurately capture the latest information in regard to disease mappings across different resources, and to incorporate human interpretations. For example, "Spondylothoracic dysostosis (GARD:0006798)" and "Spondylocostal dysostosis 1 (GARD:0010726)" share 3 of the same mappings, "ORPHA:2311," "UMLS:C0265343," and "OMIM:277300," so their similarity score equals 1.0, indicating that they should be highly similar. However, our experts marked them as "Unrelated" due to the fact that these 2 conditions were grouped together in the past (both were previously referred to as Jarcho-Levin syndrome); they are considered as distinct conditions now, according to references from GHR (Genetic Home Reference) [29,30]. Berdon et al [31] also discussed the

clinical and radiological distinction between these 2 diseases. Another example is "Hunter Carpenter Macdonald syndrome (GARD:0002751)" and "Infantile neuroaxonal dystrophy (GARD:0003957)," which have a similarity score of 0.35, indicated they should be less relevant. However, it was marked as relevant by our experts given that PLA2G6-associated neurodegeneration (PLAN) comprises a continuum of 3 phenotypes with overlapping clinical and radiologic features for these 2 diseases, and similar evidence can be found at Orphanet [32] that reveals that Hunter-Carpenter-McDonald syndrome has been moved to "Infantile neuroaxonal dystrophy." In comparison of the total 13,705,230 GARD disease pairs, there are only 392 disease pairs with similarity scores greater than 0, which might direct the extension in 2 ways. First, 3 selected resources might not be comprehensive enough to cover all GARD diseases for disease similarity comparison based on their disease mappings. Therefore, we plan to extend our work with additional rare disease resources, such as MONDO, Disease Ontology, NCI Thesaurus, etc. Second, external disease mappings curated by GARD are accurate but might be incomplete due to cumbersome human effort. Thus, we will extend the disease mappings by inferring new associations via network analysis from the Neo4j knowledge graph.

Phenotypical similarity derivation from disease classifications resulted in 360 disease pairs shared with at least 1 phenotype and gene, and they are grouped as sibling diseases. Among 241,604 disease pairs retrieved from the disease classification trees, there are 84,054 disease pairs that share at least 1 phenotype and 396 disease pairs that share at least 1 gene. Compared to the number of disease pairs with shared phenotypes, a relatively small number of disease pairs shared at least 1 gene; we are planning to obtain more genes for GARD diseases from other resources, including DisGeNet [33] and ClinVar [34]. Given the success we gained from this study in identifying phenotypical similarity derived from sibling diseases from disease classifications, we propose to extend this work with subtype diseases (ie, parent diseases and child diseases) by mining disease classifications. Once we have GARD diseases that we are able to assign to those relevant categories, we will develop our own disease classification system, which will not only define more accurate disease definitions and relationships among those diseases but will also serve as a unique, rare disease resource in the United States.

By combining 2 sets generated by our 2 approaches, we identified 662 phenotypically similar disease pairs and mapped them to 4 phenotypical similarity types, namely, "Duplicates," "Subtypes," "Siblings," and "Unrelated," which will be applied to direct GARD data harmonization. To be specific, for "Duplicate" disease pairs, we will select and keep primary diseases in the GARD database; "Siblings" and "Subtypes" will direct GARD disease classification regeneration; for "Unrelated" diseases, we will keep these 2 diseases separately in the GARD database.

By comparing these 2 sets, there are 39 overlapped disease pairs. These 39 disease pairs were grouped as "Siblings" by the second approach of disease classification derivation. However, based on the evaluation result (Table 4) from the first approach of disease mapping analysis, of these 39 disease pairs, 25 disease pairs were grouped as "Siblings," 7 pairs were grouped as "Subtypes," 2 pairs were grouped as "Duplicated," and 5 pairs were grouped as "Unrelated." For instance, "Malignant hyperthermia" and "King Denborough syndrome" are classified as sibling diseases by the second approach, since they are siblings in Orphanet, which groups them under the same disease parent class of "Rare Disease With Malignant Hyperthermia (ORPHA:466658)." However, they are determined as different diseases by our subject-matter experts, and the same statement has been made in the GARD page for "King-Denborough syndrome (GARD:0008433)," claiming that "King-Denborough syndrome is a congenital myopathy associated with susceptibility to malignant hyperthermia (GARD:0006964)" [35]. Such discrepancies occurring across different resources unveiled from this study illustrate that there is an urgent need to propose a standard protocol for guiding data harmonization in the rare disease field globally. Regardless of phenotypical similarity types, the process our subject-matter experts took in the evaluation step is crucial to re-evaluate rare diseases with the latest prevalence data, which is one critical step to determine their eligibility of RARE. For instance, there are more than 200,000 individuals in the United States who are affected with familial Alzheimer disease (GARD:0000632) [36,37]; thus, the prevalence rate of this disease does not meet the criteria of the United States' rare disease definition, so it will be retired from the GARD database.

Conclusion

In this paper, we report our recent effort at identifying phenotypical similarity among rare diseases by leveraging disease mappings among various resources and disease classifications. This effort will not only direct further GARD data harmonization but will also highlight the value of cross-resource collaboration. We propose to extend this work with more rare disease resources at the NIH or outside the NIH for the improved assembly of information for rare diseases in order to better disseminate information to patients and health care providers.

Acknowledgments

This research was supported by the Intramural research program of the NCATS, NIH.

The authors thank Tongan Zhao from the Division of Pre-Clinical Innovation (DPI) at NCATS, who developed the GARD Data Tree web tool ; Michelle Snyder, an operational program manager of the GARD from ICF International Inc, who supported the manual review and provided valuable discussion; and Jim Dickens, a program manager from the Office of Rare Diseases Research (ORDR) at NCATS, who helped to arranged discussion meetings and participated in valuable discussions.

Authors' Contributions

The work was conceived by QZ, who also designed and performed the experiments and wrote all the source code and the manuscript. TN supported data extraction from the Neo4j database and participated in the project discussion. GA and KH, as the subject matter experts, manually reviewed and evaluated the results and provided valuable insights. ES participated in the project discussion and helped with the manual review. AP, as the Director of the Office of Rare Diseases Research (ORDR) at NCATS, supported this work and participated in valuable discussions. All authors read, edited, and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Boat T, Field M. Rare diseases and orphan products: Accelerating research and development. Washington, DC: National Academies Press; 2011:A.
2. Groft S, Posada de la Paz M. Rare Diseases: Joining Mainstream Research and Treatment Based on Reliable Epidemiological Data. Adv Exp Med Biol 2017;1031:3-21. [doi: 10.1007/978-3-319-67144-4_1] [Medline: 29214563]
3. RARE AND ULTRA-RARE DISEASES. URL: https://bit.ly/2SbQifn [accessed 2020-09-24]
4. The Golbal Challenge of Rare Disease Diagnosis. URL: https://sc8-cms-shire-com.shirecontent.com/-/media/shire/shireglobal/shirecom/pdffiles/patient/shire-diagnosis-initiative-hcp-leaflet.pdf [accessed 2020-09-24]
5. Genetic and Rare Disease Information Center. URL: https://rarediseases.info.nih.gov/ [accessed 2020-09-24]
6. Weinreich S, Mangon R, Sikkens J, Teeuw M, Cornel M. [Orphanet: a European database for rare diseases]. Ned Tijdschr Geneeskd 2008 Mar 01;152(9):518-519. [Medline: 18389888]
7. Mondo Disease Ontology. URL: http://obofoundry.org/ontology/mondo.html [accessed 2020-09-24]
8. Haendel M, Vasilevsky N, Unni D, Bologa C, Harris N, Rehm H, et al. How many rare diseases are there? Nat Rev Drug Discov 2019 Nov 5;19(2):77-78. [doi: 10.1038/d41573-019-00180-y]
9. Pontikos N, Yu J, Moghul I, Withington L, Blanco-Kelly F, Vulliamy T. Phenopolis: an open platform for harmonization and analysis of genetic and phenotypic data. Bioinformatics 2017;33(15):a. [doi: 10.1093/bioinformatics/btx147]
10. Siu LL, Lawler M, Haussler D, Knoppers BM, Lewin J, Vis DJ, et al. Facilitating a culture of responsible and effective sharing of cancer genome data. Nat Med 2016 May 05;22(5):464-471 [FREE Full text] [doi: 10.1038/nm.4089] [Medline: 27149219]
11. Boutros PC, Ewing AD, Ellrott K, Norman TC, Dang KK, Hu Y, et al. Global optimization of somatic variant identification in cancer genomes with a global community challenge. Nat Genet 2014 Mar 27;46(4):318-319. [doi: 10.1038/ng.2932]
12. About Orphanet. URL: https://www.orpha.net/consor/cgi-bin/Education_AboutOrphanet.php?lng=EN [accessed 2020-09-24]
13. OMIM Frequently Asked Questions (FAQs). URL: https://www.omim.org/help/faq [accessed 2020-09-24]
14. Reumann M, Giovannini A, Nadworny B, Auer C, Girardi I, Marchiori C. Cognitive DDx Assistant in Rare Diseases. In: Annu Int Conf IEEE Eng Med Biol Soc. 2018 Jul Presented at: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 18-21 July 2018; Honolulu, HI, USA p. 3244-3247. [doi: 10.1109/EMBC.2018.8513041]
15. Li X, Wang Y, Wang D, Yuan W, Peng D, Mei Q. Improving rare disease classification using imperfect knowledge graph. BMC Med Inform Decis Mak 2019 Dec 5;19(S5). [doi: 10.1186/s12911-019-0938-1]
16. Sosa D, Derry A, Guo M, Wei E, Brinton C, Altman R. A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases. bioRxiv 2019:727925. [doi: 10.1142/9789811215636_0041]
17. Disease oriented integrative knowledge graph. URL: https://disease.ncats.io/browser/ [accessed 2020-09-24]
18. Orphanet rare disease classification. URL: https://www.orpha.net/consor/cgi-bin/Disease_Classif.php?lng=EN [accessed 2020-09-24]
19. Procedural document: Orphanet nomenclature and classification of rare diseases. URL: https://www.orpha.net/orphacom/cahiers/docs/GB/eproc_disease_inventory_R1_Nom_Dis_EP_04.pdf [accessed 2020-09-24]
20. Hamosh A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research 2004 Dec 17;33(Database issue):D514-D517. [doi: 10.1093/nar/gki033]
21. Robinson P, Mundlos S. The human phenotype ontology. Clinical genetics 2010;77(6):525-534. [doi: 10.1111/j.1399-0004.2010.01436.x]
22. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research 2004 Jan 01;32(90001):267D-2270. [doi: 10.1093/nar/gkh061]
23. Cypher Query Language. URL: https://neo4j.com/developer/cypher-query-language/ [accessed 2020-09-24]
24. Cosine Similarity. URL: https://en.wikipedia.org/wiki/Cosine_similarity [accessed 2020-09-24]
25. Cancer Stat Facts: Testicular Cancer From NCI SEER. URL: https://seer.cancer.gov/statfacts/html/testis.html [accessed 2020-09-24]

26. Hoehndorf R, Schofield PN, Gkoutos GV. Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases. Sci Rep 2015 Jun 08;5:10888 [FREE Full text] [doi: 10.1038/srep10888] [Medline: 26051359]
27. GARD Data Tree. URL: https://tripod.nih.gov/gardtree/ [accessed 2020-09-23]
28. FAQs About Rare Diseases. URL: https://rarediseases.info.nih.gov/diseases/pages/31/faqs-about-rare-diseases [accessed 2020-09-24]
29. Spondylocostal dysostosis. URL: https://ghr.nlm.nih.gov/condition/spondylocostal-dysostosis [accessed 2020-09-24]
30. Spondylothoracic dysostosis. URL: https://ghr.nlm.nih.gov/condition/spondylothoracic-dysostosis [accessed 2020-09-24]
31. Berdon WE, Lampl BS, Cornier AS, Ramirez N, Turnpenny PD, Vitale MG, et al. Clinical and radiological distinction between spondylothoracic dysostosis (Lavy-Moseley syndrome) and spondylocostal dysostosis (Jarcho-Levin syndrome). Pediatr Radiol 2010 Dec 22;41(3):384-388. [doi: 10.1007/s00247-010-1928-8]
32. Hunter-Carpenter-McDonald syndrome. URL: https://bit.ly/3cLGC51 [accessed 2020-09-24]
33. Piñero J, Bravo , 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 2016 Oct 19;45(D1):D833-D839. [doi: 10.1093/nar/gkw943]
34. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 2015 Nov 17;44(D1):D862-D868. [doi: 10.1093/nar/gkv1222]
35. King Denborough syndrome. URL: https://bit.ly/36iRohM [accessed 2020-09-24]
36. Mayeux R, Stern Y. Epidemiology of Alzheimer Disease. Cold Spring Harbor Perspectives in Medicine 2012 Apr 10;2(8):a006239-a006239. [doi: 10.1101/cshperspect.a006239]
37. Bird T. Alzheimer disease overview. GeneReviews®Internet: University of Washington, Seattle; 2018.

Abbreviations

GARD: genetic and rare diseases
HPO: Human Phenotype Ontology
MONDO: Monarch Disease Ontology
OMIM: Online Mendelian Inheritance in Man
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 24.02.20; peer-reviewed by N Mohammad Gholi Mezerji, J Yang; comments to author 09.06.20; revised version received 02.08.20; accepted 19.08.20; published 02.10.20.

Please cite as:
Zhu Q, Nguyen DT, Alyea G, Hanson K, Sid E, Pariser A
Phenotypically Similar Rare Disease Identification from an Integrative Knowledge Graph for Data Harmonization: Preliminary Study
JMIR Med Inform 2020;8(10):e18395
URL: https://medinform.jmir.org/2020/10/e18395
doi:10.2196/18395
PMID:33006565

Original Paper

Building a Pharmacogenomics Knowledge Model Toward Precision Medicine: Case Study in Melanoma

Hongyu Kang[1,2], MSc; Jiao Li[1], PhD; Meng Wu[1], MSc; Liu Shen[1], MSc; Li Hou[1], PhD

[1]Institute of Medical Information &Library, Chinese Academy of Medical Sciences/Peking Union Medical College, Beijing, China

[2]Department of Biomedical Engineering, School of Life Science, Beijing Institute of Technology, Beijing, China

Corresponding Author:
Li Hou, PhD
Institute of Medical Information &Library
Chinese Academy of Medical Sciences/Peking Union Medical College
3 Yabao Road, Chaoyang District
Beijing
China
Phone: 86 18910120178
Email: hou.li@imicams.ac.cn

Abstract

Background: Many drugs do not work the same way for everyone owing to distinctions in their genes. Pharmacogenomics (PGx) aims to understand how genetic variants influence drug efficacy and toxicity. It is often considered one of the most actionable areas of the personalized medicine paradigm. However, little prior work has included in-depth explorations and descriptions of drug usage, dosage adjustment, and so on.

Objective: We present a pharmacogenomics knowledge model to discover the hidden relationships between PGx entities such as drugs, genes, and diseases, especially details in precise medication.

Methods: PGx open data such as DrugBank and RxNorm were integrated in this study, as well as drug labels published by the US Food and Drug Administration. We annotated 190 drug labels manually for entities and relationships. Based on the annotation results, we trained 3 different natural language processing models to complete entity recognition. Finally, the pharmacogenomics knowledge model was described in detail.

Results: In entity recognition tasks, the Bidirectional Encoder Representations from Transformers–conditional random field model achieved better performance with micro-F1 score of 85.12%. The pharmacogenomics knowledge model in our study included 5 semantic types: drug, gene, disease, precise medication (population, daily dose, dose form, frequency, etc), and adverse reaction. Meanwhile, 26 semantic relationships were defined in detail. Taking melanoma caused by a *BRAF* gene mutation into consideration, the pharmacogenomics knowledge model covered 7 related drugs and 4846 triples were established in this case. All the corpora, relationship definitions, and triples were made publically available.

Conclusions: We highlighted the pharmacogenomics knowledge model as a scalable framework for clinicians and clinical pharmacists to adjust drug dosage according to patient-specific genetic variation, and for pharmaceutical researchers to develop new drugs. In the future, a series of other antitumor drugs and automatic relation extractions will be taken into consideration to further enhance our framework with more PGx linked data.

(JMIR Med Inform 2020;8(10):e20291) doi:10.2196/20291

KEYWORDS

pharmacogenomics; knowledge model; BERT–CRF model; named entity recognition; melanoma

Introduction

Pharmacogenomics

The field of pharmacogenomics (PGx) has developed rapidly since the initial scientific discoveries of genetic characteristics affecting individual response to drugs or other agents [1]. Through these years of development, PGx aims at understanding how genetic variants influence drug efficacy and toxicity. It combines pharmacology (the science of drugs) and genomics (the study of genes and their functions), and is certain to improve new drug development and precision medication. Such studies can reveal how genetic variation across individuals affects a

drug's pharmacokinetics and pharmacodynamics [2]. Many drugs do not work the same way for everyone. Consequently, PGx is often considered one of the most actionable areas of the personalized medicine paradigm [3].

As of June 2019, more than 190 drugs [4] approved by the US Food and Drug Administration (FDA) clearly stated on in their medical specifications that they need to be deployed with greater precision based on individual genotype. The introduction of targeted drugs and targeted therapies provides a more feasible and effective way for cancer treatment, improves drug efficacy, and reduces adverse reactions. Therefore, studies of new therapies related to PGx such as drug combinations and new drug discoveries [5] have become increasingly popular. A typical case of repurposing drugs is afatinib (40 mg q.d.), which was introduced [6] for treating lung cancer after *NGR1* gene fusion.

Named Entity Recognition

Named entity recognition (NER) is a basic tool for natural language processing (NLP) tasks such as information extraction, question answering system, syntactic analysis, and machine translation. Its main goal is identifying entities with specific meaning in the text, mainly including people's names, place names, organization names, proper nouns, etc. It is the foundation of identifying semantic relationships between entities and filling a knowledge base.

The common statistical models of NER mainly include the Hidden Markov Model [7] and the conditional random field (CRF) [8]. In recent years, neural network deep learning methods based on the development of word vector technology, such as the convolutional neural network (CNN) [9] and the recurrent neural network (RNN), have made a great breakthrough in the field of NLP. After that, long short-term memory (LSTM) [10] added a memory cell to RNN, to overcome the problem of gradient explosion and gradient disappearance. Bidirectional RNN [11] adopts a double-layer RNN structure, which can collect forward and backward information at the same time.

In 2018, Devlin et al [12] from Google AI Language proposed the Bidirectional Encoder Representations from Transformers (BERT) which provided outstanding performance in 11 NLP tasks, opening a new era for NLP. Similar to the general pretraining 2-stage training method, BERT uses the language model for pretraining as the first stage. In the second stage, it fine-tunes for downstream tasks, and achieves the best results in multiple NLP tasks. The BERT–CRF model [13] and multilingual BERT model [14] were trained on different languages such as Portuguese and the F1 score was ultimately improved. Today, the BERT model has also been applied in biomedical research. BERT-based models were investigated for their effectiveness in biomedical and clinical entity normalization, and achieved state-of-the-art performance on large-scale electronic health record notes [15] and online corpus [16]. The BioBERT model [17] for biomedical text mining tasks and the ClinicalBERT [18] for clinical notes were also introduced and outperformed previous models.

Biomedical Knowledge Representation

The Knowledge Representation Model can be understood as a structured set of directed graphs, in which the nodes of the graph represent entities or concepts, while the edges represent the semantic relationship between entities or concepts. During the development of the knowledge representation, sematic networks, ontology, and knowledge graphs/models are most commonly used in the field of biomedical science.

A semantic network [19], or frame network, is a knowledge base that represents semantic relations between concepts in a network.

An ontology is a formal explicit description of concepts in a domain, properties of each concept, various features and attributes, and restrictions on these properties [20]. The Drug Target Ontology [21] provided a framework and formal classification, which included related information between protein, gene, protein domain, binding site, small-molecule drug, mechanism of action, and many other types of information. Dumontier and Villanuevarosales [22] constructed a lightweight ontology, Pharmacogenomics Ontology, based on Pharmacogenomics Knowledge Base (PharmGKB) data, which contains 40 core concepts, involving phenotype, genotype, and drug therapy.

A knowledge graph/model emphasizes data cleaning and knowledge fusion, and its essence is a semantic network, which allows access to knowledge inference. Since this concept was put forward by Google in 2012 [23], researchers have conducted a series of discussions and research aimed at intelligent retrieval. High-quality heterogeneous graphs such as the Safe Medicine Recommendation (SMR) [24] and KnowLife [25] contain entities and relationships between disease, medicine, patient, gene, organ, and other biomedical entities constructed by bridging electronic medical records, ICD-9, DrugBank, electronic health record [26], and other databases, which leads to more hidden relationships.

Above all, the knowledge graph/model technology provides a means to extract structured knowledge from massive texts and images. It has broad applications in biomedical field and can promote intelligent semantic retrieval, medical questions and answers, clinical decision support, and many other scenarios.

Related Works

With the rapid growth and accumulation of massive PGx data, there is an increasing need for scientific data collecting, organizing, modeling, and mining. These data reflect a hierarchy of relationships and detailed information between biomedical entities. Currently, the semantic types and relationships involved in PGx knowledge representation are usually limited to drug, gene, and disease.

Drug–Gene Target Treatment

Drug2Gene [27] was a knowledge base combining information on compound, drug, gene, and protein from 19 publicly available databases. Sun et al [28] designed a computational workflow to construct drug-target networks including drugs, genes, and diseases from different knowledge bases.

Drug–Gene–Drug Interaction

Bo et al [29] extracted drug–gene–drug interactions from biomedical literature using the bidirectional LSTM (Bi-LSTM) model by combining biomedical resources with lexical information and entity position information. Coulet et al [30] instantiated a description logics knowledge base to identify gene variant–drug response associations.

Drug–Gene–Phenotype Relationship

Dalleau et al [31] assembled a set of linked PGx data from 6 distinct resources such as DisGeNET [32] and ClinVar [33].

Disease–Chemical–Gene Relationship

Kim et al developed DigSee [34] for disease–gene relationships and DigChem [35] for disease–gene–chemical relationships from biomedical literature abstracts at a PubMed scale.

However, there currently exist no in-depth explorations and descriptions of personalized medication, such as drug usage, dosage adjustment, and applicable population. Therefore, there is significance in applying the knowledge model to the field of PGx in further study, which will assist clinicians and clinical pharmacists in precise medication.

Objective

In this study, we proposed the following 2 objects:

1. We aimed to present a pharmacogenomics knowledge model consisting of 5 semantic types related to PGx and precision medication, and also give definitions of relationships between these entities. The model mostly focuses on anticancer drugs, drug usage, and adjustments of daily dosage.
2. We aimed to semiautomatically construct PGx corpora, which are relatively rare in the existing research, and make them open access. The NLP algorithms for PGx NER were also trained for facilitating corpus annotation.

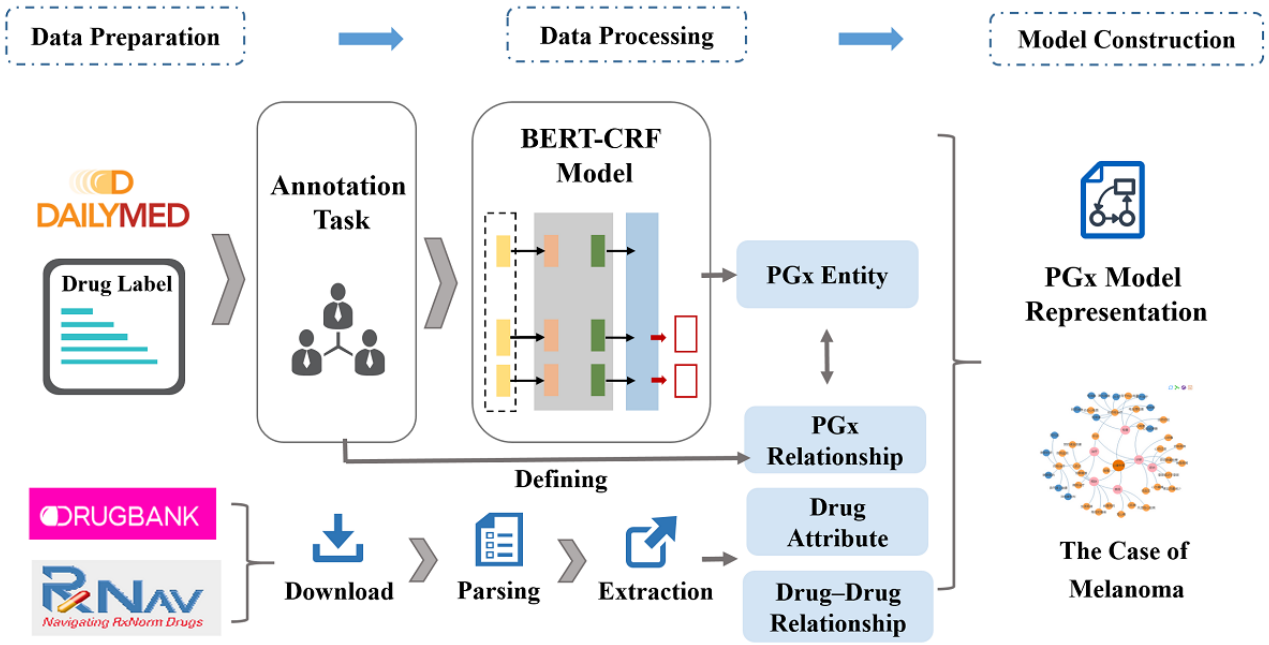
Methods

Study Steps

There are 3 main steps in our study (Figure 1).

1. Data preparation: Data related to PGx were collected from DailyMed, DrugBank, and RxNorm.
2. Data processing: Manual annotation for PGx entities and relationships were applied to drug labels in PDF/XML format from DailyMed. The BERT–CRF model were trained for entity recognition in this study. Data from DrugBank and RxNorm were also downloaded, parsed, and extracted for more drug attributes and relationships.
3. Model construction: The PGx knowledge model was described in this aspect based on the entities and relationships extraction. Melanoma was also used as an example to verify the accuracy and validity of our model.

Figure 1. The framework of our study.



Data Preparation

Data related to PGx need to be collected and integrated in this study, which are currently stored in DrugBank, PharmGKB, Comparative Toxicogenomics Database (CTD), RxNorm, and other databases. Based on the pharmacogenomics knowledge model built in our study, we chose the following 3 data sources to accomplish data crawling and data preparation.

DailyMed

The text of drug labels was obtained from DailyMed, which is a free drug information resource [36] provided by the US National Library of Medicine (NLM). It consists of digitized versions of drug labels as submitted to the US FDA. DailyMed was of special interest because of its comprehensive coverage, open availability, and the package inserts' combination of format consistency and rich detail. Drug labels in DailyMed give a

detailed description of drugs' indications and usage, adverse reaction, and applicable population, especially the dosage, dose form, and dosage adjustment. We downloaded 4067 drug labels randomly for pretraining tasks and 190 drug labels in the table of PGx biomarkers for annotation tasks.

DrugBank

DrugBank is a unique bioinformatics and cheminformatics resource that combines detailed drug (ie, chemical, pharmacological, and pharmaceutical) data with comprehensive drug target (ie, sequence, structure, and pathway) information [37] provided by the University of Alberta. The latest release of DrugBank (version 5.1.4, released July 2, 2019) was parsed in this paper for drug attributes such as drug name, description, chemical formula, molecular weight, drug approval status, and so on.

RxNorm

RxNorm [38] provides a suite of standards for clinical drugs in the form of "Ingredient–Strength–Dose Form–Brand name," and is designed by NLM for the electronic exchange of clinical health information. Several attributes and drug–drug interactions of precise medication were selected from RxNorm, such as daily dose, dose form, and frequency as attributes, and has_dose_form, dose_form_of as relationships.

Annotation Task

We recruited 3 annotators, all of whom had a medical training background and curation experience. Each drug label was annotated independently by 2 annotators (ie, double annotation). Differences were resolved by a third and senior annotator. Besides this, we measured agreement of relationship annotations using the *F* score to assess consistency.

Because all 190 drug labels in the FDA table of PGx biomarkers [4] are in PDF format, the annotator needed to convert all of them into an editable format such as .txt (Notepad or other word processors) or .doc/.docx (Microsoft Word) before annotation.

The main tasks involved in the annotation stage were the recognition of semantic types and semantic relationships from drug labels sections, including "Indications and Usage," "Dosage and Administration," "Use in Specific Populations," "Warnings and Precautions," and "Adverse Reactions." For semantic types, different highlighted colors represented different entities according to the frame of the PGx knowledge model. In this work, drug was annotated in yellow, gene was annotated in red, disease was annotated in gray, dosage and dose form were annotated in green, adverse reaction was annotated in purple, and population was annotated in blue. For semantic relationships, the more important and difficult section, annotators read the drug labels and recorded the relation descriptions between diseases and drugs, diseases and genes, diseases and diseases, drugs and genes, drugs and drugs, and drugs and dosage manually. This formed the basis of relationship definition in the follow-up work. Before annotation, we also indicated the annotation guidelines, see in Figure 2.

An example of drug label annotation is shown in Figure 3. Finally, all the annotated semantic types and relationships were recorded in a structured database designed in advance.

Figure 2. Annotation guidelines.

1. Annotate diseases/symptoms treated directly by drugs.

Incorrect Annotation	dilatrate-SR sustained release capsules are indicated for the prevention of angina pectoris due to coronary artery disease.
Correct Annotation	dilatrate-SR sustained release capsules are indicated for the prevention of angina pectoris due to coronary artery disease.

2. Annotate all the conditions for dosage adjustment

Incorrect Annotation	The recommended dose of Diazepam rectal gel is 0.2-0.5 mg/kg depending on age.
Correct Annotation	The recommended dose of Diazepam rectal gel is 0.2-0.5 mg/kg depending on age, for 0.5 mg/kg to 2 through 5, 0.3 mg/kg to 6 through 11, 0.3 mg/kg to 12 and older.

3. Annotate dosage adjustment caused by gene mutation/adverse reaction/population difference.

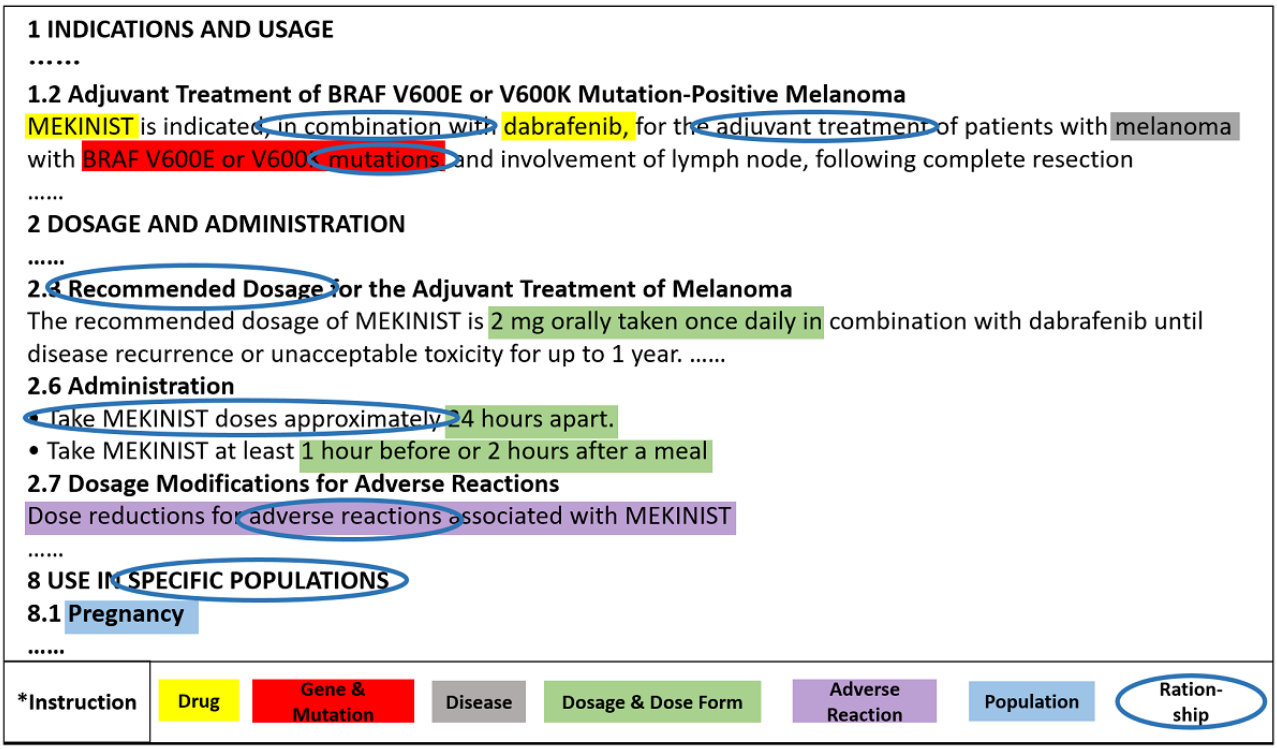
4. Do not annotate the maximum and minimum dosage of a drug unless the adjustment is affected by other conditions mentioned in 3.

Incorrect Annotation	In clinical trials, immediate-release oral isosorbide dinitrate has been administered in a variety of regimens, with total daily doses ranging from 30 to 480 mg. Do not exceed 160 mg (4 capsules) per day.
Correct Annotation	In clinical trials, immediate-release oral isosorbide dinitrate has been administered in a variety of regimens, with total daily doses ranging from 30 to 480 mg. Do not exceed 160 mg (4 capsules) per day.

5. Annotate the targeted gene, not all the genes related to the drug.

Incorrect Annotation	Carisoprodol is metabolized in the liver by CYP2C19 to form meprobamate.
Correct Annotation	Carisoprodol is metabolized in the liver by CYP2C19 to form meprobamate.

Figure 3. Annotation example of MEKINIST.



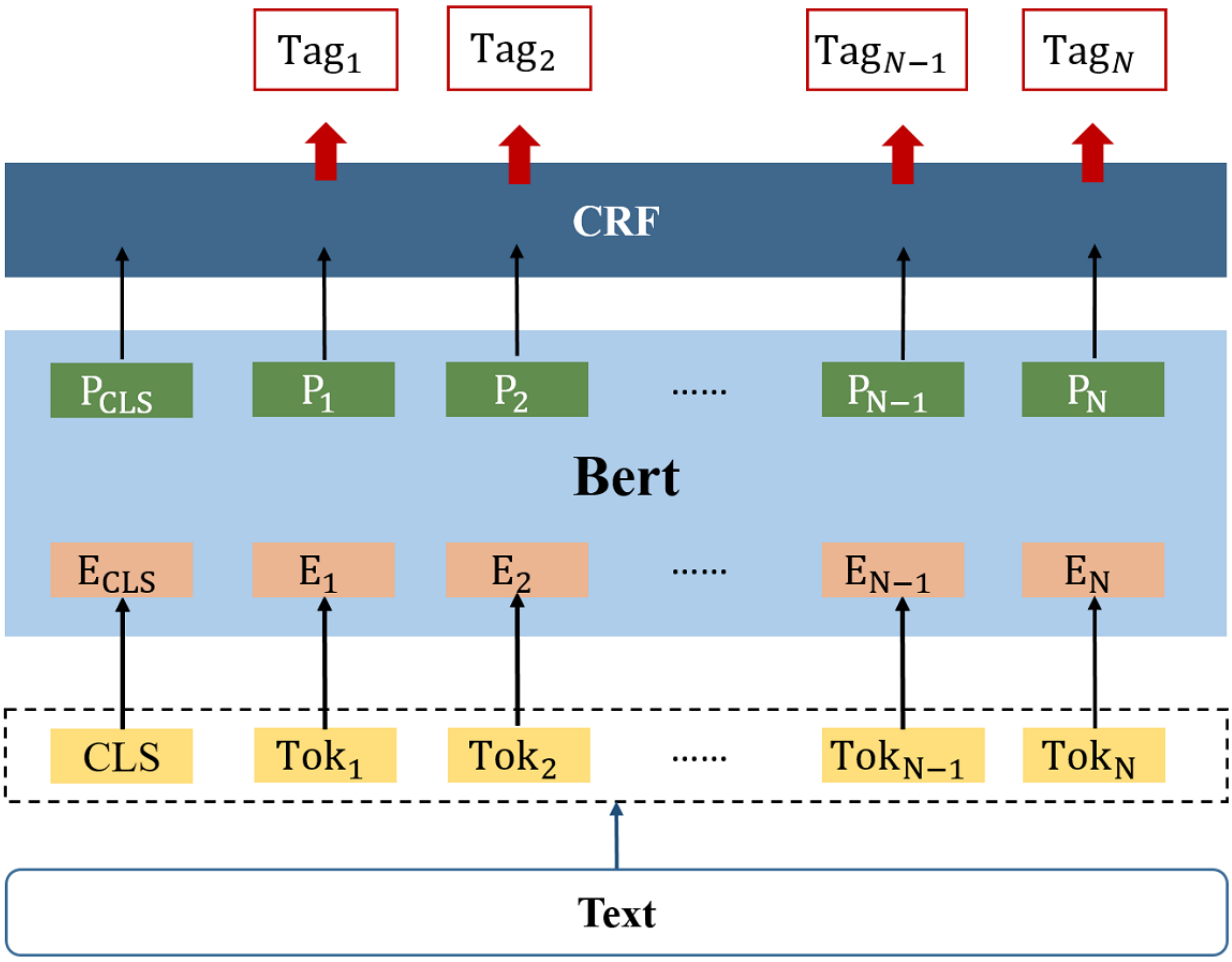
BERT–CRF for NER

After the annotation of entities, we applied the BERT–CRF model for NER. The CRF model and BERT–Bi-LSTM–CRF model were also trained in our study as a comparison.

The BERT–CRF architecture was composed of 4 sections: the input layer, the pretraining model, the full connection layer, and the CRF layer, which assigns a tag to each word based on its context in the output (Figure 4). We feed a sentence to the architecture to obtain contextual BERT embedding for each word as $\{Tok_1,...,Tok_N\}$ The context could be captured via many attention heads in each of its layers as well. These embeddings were then transported to a CRF layer to obtain the tag as $\{Tag_1,...,Tag_N\}$ for each word block.

The BERT-Base Multilingual, which has 110M parameters, was used in this NER task. We set the training batch size to 32, the max_seq to 80, and the learning rate to 0.00001. A total of 10 epochs were trained in each iteration to ensure model convergence. Other parameters related to BERT are set to default values. The dropout rate was set to 0.9 in fully connected layers to prevent over fitting. The transfer matrix in CRF is also left for the model to learn. The transfer matrix in the CRF layer was learned by the model itself. Importantly, the Bi-LSTM layer was added in this architecture before feeding the tweet-level representation into the CRF layer, to compare the performance between BERT–CRF with Bi-LSTM and without Bi-LSTM.

Figure 4. BERT–CRF architecture. BERT: Bidirectional Encoder Representations from Transformers; CRF: Conditional Random Field.



Model Representation

We extended the semantic types of our model from 3 common types of drug, gene, and disease to 5 types: drug, gene (gene name, gene mutation), disease (disease name, position, etc), precise medication (population, daily dose, dose form, frequency, take time for, take with a meal or not, etc), and adverse reaction.

All the semantic types and attributes covered in pharmacogenomics knowledge model are shown in Table 1.

The entities model in pharmacogenomics knowledge model was defined and EID represented the unique identifier for entities

Entity={EID*,TERM*,Source,SEMANTICType*} (1)

The relationships model in pharmacogenomics knowledge model was defined and RID represented the unique identifier for relationships

Relation={RID*,Relationship*,Domain*,

Range*,Definition,TreeNumber*} (2)

The whole pharmacogenomics knowledge model can be represented as the risk factors of precision medication for cancers. In this model, disease (C, especially for cancer in this paper) is usually caused by gene mutations (G), which decided the target drug (Dr) for treatment.

Dr = F(C,G) (3)

During treatment, routine dosage/dose form (Ds) has been already offered by the FDA drug labels. However, it differs when the patient has an adverse reaction (A) or the disease occurs in special groups (P) such as pregnancy, lactation, pediatric, geriatric. Assuming that the 4 factors are independent in some cases, each factor can effect dosage/dose form separately.

Ds = F(Dr,G,A,P) (4)

Above all, gene mutation, disease, adverse reaction, and patient populations are the risk factors in pharmacogenomics knowledge model of drugs to be used, and suitable dosage and dose form especially.

Dr, Ds=F(C,G,A,P) (5)

Table 1. Semantic types and attributes in the knowledge model.

Semantic Type	Entity/Attribute
Drug	Drug Name, Description, Chemical Formula, Molecular Weight, Drug Approval Status, CAS[a], UNII[b], Pharmacology Indication
Gene	Gene name, Mutation
Disease	Disease Name, Position
Adverse Reaction	N/A[c]
Population	Pediatric Use Population, Applicable Population, Gender, Age, Race
Drug Use	Daily dose, Dose form, Frequency, Take time for, Take with a meal or not, etc

[a]CAS: Chemical Abstracts Service Number.

[b]UNII: Unique Ingredient Identifier.

[c]N/A: not available.

Results

Data Set Overview

In this paper, we have collected 4067 drug labels in XML format downloaded from DailyMed as pretraining data for the BERT–CRF architecture, and 190 drug labels after annotation for model representation in which 90% (n=171) form the training set and 10% (n=19) form the test set, randomly assigned. Statistics-annotated corpus are presented in Table 2. Besides, the number of unique unigrams were 2216 in the training set and 829 in the test set; the number of unique bigrams were 120,705 in the training set and 18,851 in the test set.

Table 2. Number of entities in training and test sets.

Entity	Number of entities in the training set	Number of entities in the test set
Drug	76	31
Gene	60	26
Disease	94	33
Body_Part	23	7
Daily_Dose	99	27
Dose_Form	16	8
Frequency	32	12
Adverse_Reaction	372	77

Performance of Named Entity Recognition

Three basic models are compared, with the specific results shown in Table 3 in which minor averaging for the F1 score was used. The BERT–CRF model achieved better performance than the other 2 models in this task. In some recent studies, the full connectivity layer was done by the Bi-LSTM layer, which ultimately resulted in the BERT–Bi-LSTM–CRF model. However, the BERT–Bi-LSTM–CRF model presented a more complex structure and slower training speed than BERT–CRF. Besides this, there was a little difference of 2% between these 2 models, so BERT–CRF was selected in our study. The BERT–CRF model showed a high F1 score in drug, dose form, and body part, but a low F1 score in daily dose and disease, shown in Table 4. However, these performances were only for the PGx corpus built semiautomatically in this work, and the 3 basic models may present different results in other studies with large-scale corpora.

Table 3. Performance of the models.

Model	Precision (%)	Recall (%)	F1 (%)
CRF[a]	88.03	73.57	80.16
BERT–CRF[b]	85.12	85.12	85.12
BERT–Bi-LSTM–CRF[c]	85.22	81.00	83.05

[a]CRF: Conditional Random Field.

[b]BERT: Bidirectional Encoder Representations from Transformers

[c]Bi-LSTM: Bidirectional Long Short-Term Memory.

Table 4. Performance of the semantic type.

Semantic type	F1		
	CRF[a] (%)	BERT–Bi-LSTM–CRF[b,c] (%)	BERT–CRF (%)
Drug	94.12	94.12	100.00
Gene	66.67	80.00	71.43
Disease	61.54	66.67	57.14
Body_Part	57.14	57.15	85.71
Daily_Dose	31.58	31.58	42.11
Dose_Form	100.00	100.00	100.00
Frequency	62.50	75.00	75.00
Adverse Reaction	68.15	79.00	73.74

[a]CRF: Conditional Random Field.

[b]BERT: Bidirectional Encoder Representations from Transformers

[c]Bi-LSTM: Bidirectional Long Short-Term Memory.

Semantic Relationships Extraction

Because this study required a high accuracy of relationship extraction, we adopted a manual method in this task. Descriptions of semantic relationships were normalized at the same time during annotation, such as “in combination with” = “synergized by,” “recommended dosage” = “routine dosage.” The normalized descriptions are presented in Table 5. The other expressions in drug labels were stored as synonyms in our study at the same time. In order to make the pharmacogenomics knowledge model be more portable, several semantic relationships were extended, such as “is biomarker-efficacy of,” “is biomarker-prognosis of.”

In the end, 26 kinds of semantic relationships were extracted, and the consistency of the entity relationship annotation was 78.55%. Among them, there were 14 first-level semantic relationships and 12 second-level semantic relationships. Each kind of semantic relationships has been defined in detail, as shown in the accessory document.

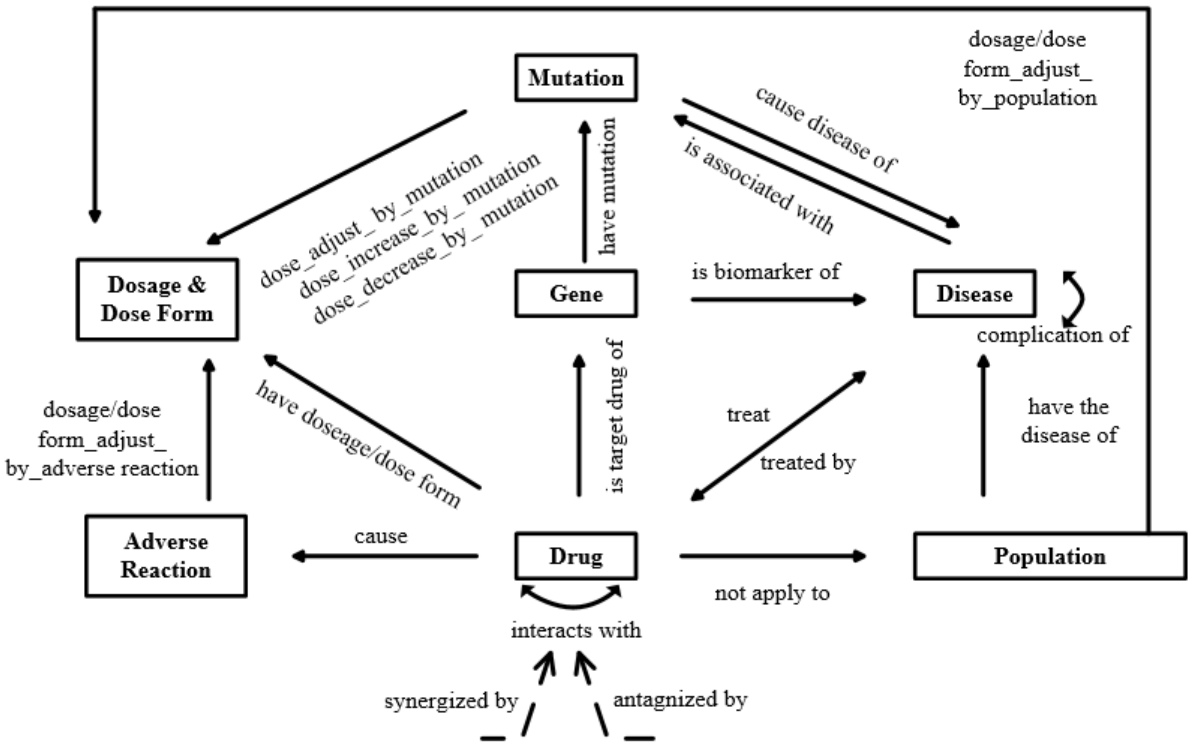
Table 5. Examples of semantic relationship–normalized description.

Normalized description	Expressions in drug labels
Treats	for the prevention of, for relief of the signs and symptoms, for the treatment of, for the prevention of, as monotherapy of
Synergized by	in combination with, coadministered with
Antagonized by	avoid concurrent administration of, avoid concomitant use of
Have dosage	total daily doses, recommended dosage
Have mutation	with *** mutation, the presence of *** mutation, be homozygous for

Pharmacogenomics Knowledge Model

Based on the entity recognition and relationship definitions mentioned above, the pharmacogenomics knowledge model is presented as Figure 5.

Figure 5. Overview of pharmacogenomics knowledge model.



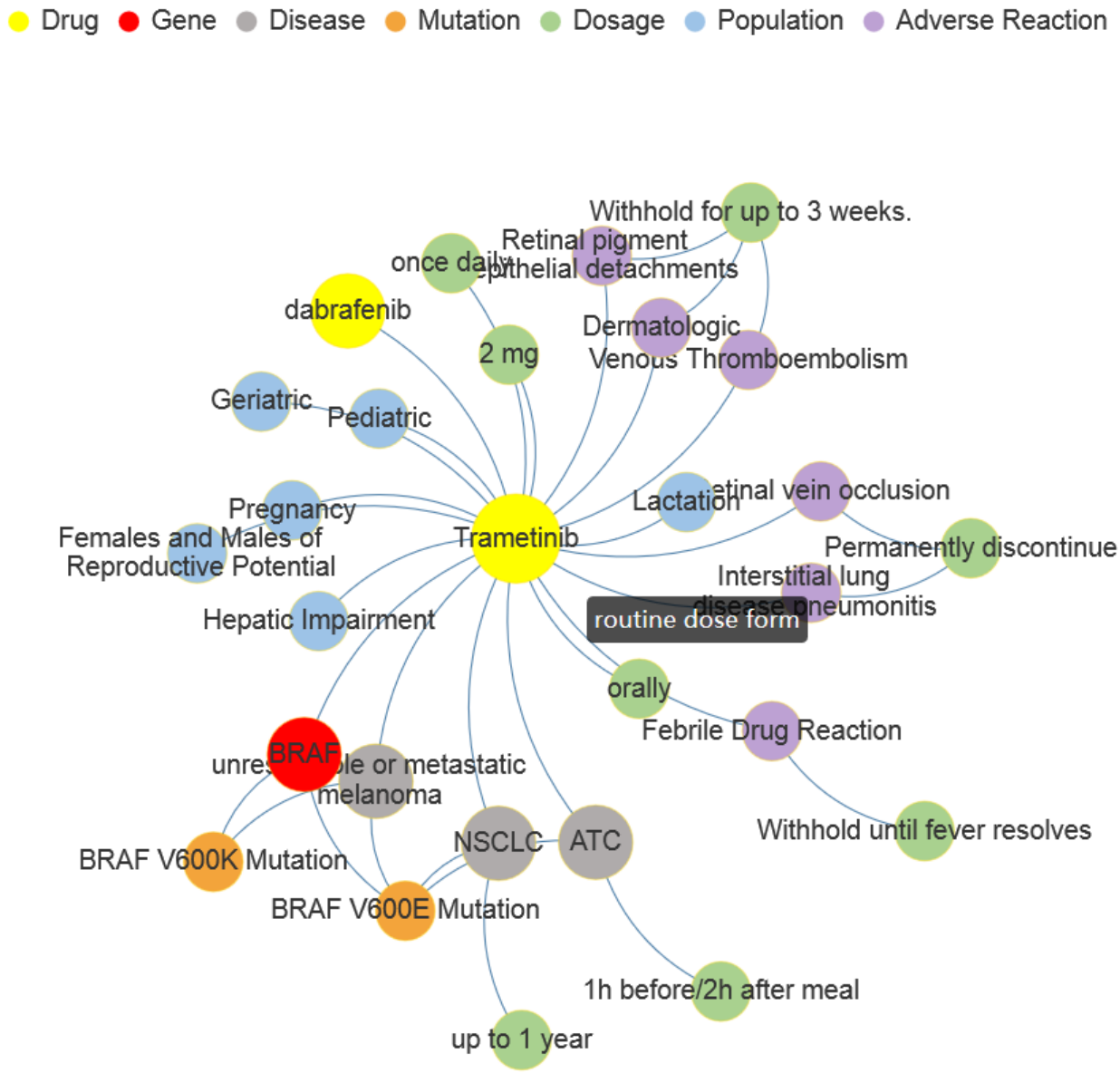
The Case of Melanoma

Melanoma is a malignant neoplasm derived from cells that are capable of forming melanin, which may occur in the skin of any part of body. It frequently metastasizes widely, and the regional lymph nodes, liver, lungs, and brain are likely to be involved. The incidence of malignant skin melanomas is rising rapidly in all parts of the world. Therefore, melanoma, which is caused by *BRAF* gene mutation, was taken as an example to verify our model.

Seven drugs were included in the cases: binimetinib, cobimetinib, dabrafenib, encorafenib, nivolumab, trametinib, and vemurafenib. Most were newly indicated for the treatment of unresectable or metastatic melanoma with *BRAF* V600E or V600K mutations, as detected by FDA-approved tests in 2018. Among them, dabrafenib, encorafenib, and vemurafenib are targeted drugs for *BRAF* gene mutations.

By researching the 7 drugs, 4846 triples were established in the pharmacogenomics knowledge model of melanoma, among them 4713 triples were drug–drug relationships, 41 were drug–adverse reaction, 30 were drug–dosage, 24 were adverse reaction–dosage, 22 were drug–disease, 7 were drug–gene, 4 were drug–population, 2 were gene–mutation, and 3 were gene–disease. An example of data visualization of trametinib can be seen in Figure 6. Relationships can be displayed when the mouse hovers over the joint(s).

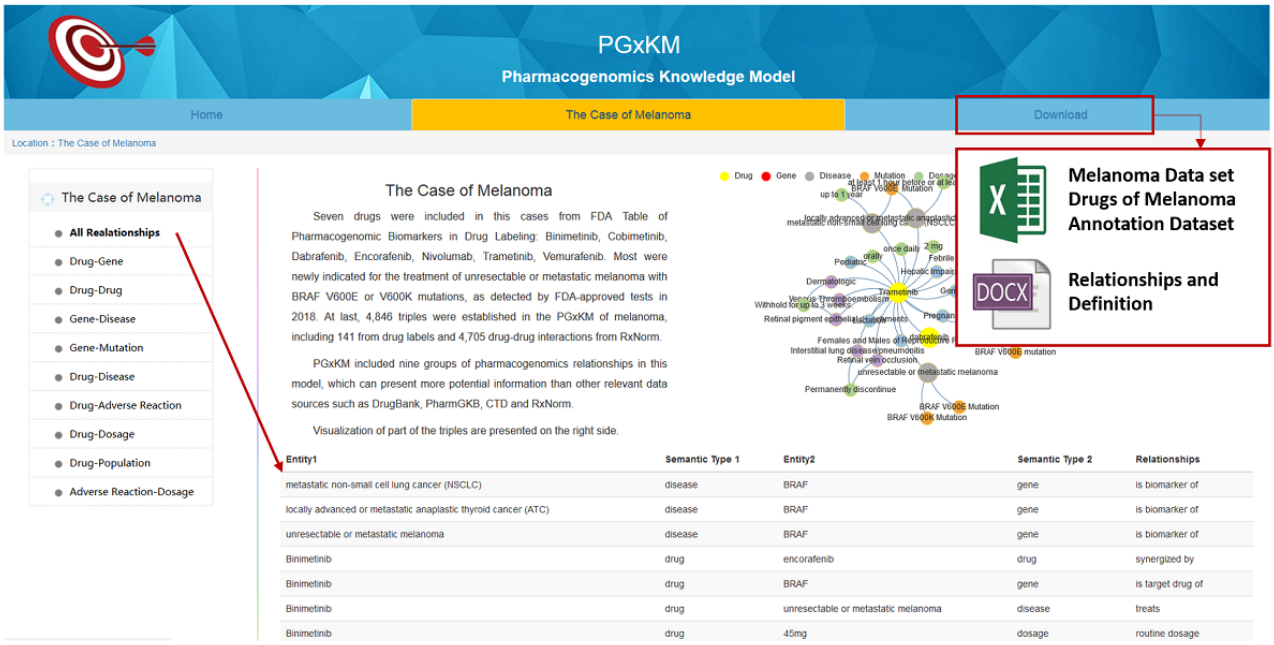
Figure 6. An example of pharmacogenomics knowledge model data visualization.



Data Set Access

We provided a user-friendly interface [39] that enables users to access the pharmacogenomics knowledge model data set (Figure 7). In the “Home” page, users can learn basic information and purpose of this knowledge model. On “The Case of Melanoma” page, users can obtain all the triples in melanoma cases and browse the triples by different groups of relationships. Visualization of the triples are presented as well. On the “Download” page, users can download the melanoma data set, drug attribute data set, and annotated data set in Microsoft Excel format, as well as the relationships and definition document in Microsoft Word format for the user’s convenience.

Figure 7. User interface of pharmacogenomics knowledge model data set.



Entity1	Semantic Type 1	Entity2	Semantic Type 2	Relationships
metastatic non-small cell lung cancer (NSCLC)	disease	BRAF	gene	is biomarker of
locally advanced or metastatic anaplastic thyroid cancer (ATC)	disease	BRAF	gene	is biomarker of
unresectable or metastatic melanoma	disease	BRAF	gene	is biomarker of
Binimetinib	drug	encorafenib	drug	synergized by
Binimetinib	drug	BRAF	gene	is target drug of
Binimetinib	drug	unresectable or metastatic melanoma	disease	treats
Binimetinib	drug	45mg	dosage	routine dosage

Discussion

Potential Relationships in Pharmacogenomics Knowledge Model

The pharmacogenomics knowledge model constructed in this paper reveals hidden relationships between drug, gene, disease, precise medication, and adverse reaction. Trametinib is used as an example, which is a kinase inhibitor indicated as a single agent for the treatment of BRAF-inhibitor treatment-naïve patients with unresectable or metastatic melanoma with *BRAF* V600E or V600K mutations as detected by an FDA-approved test. The recommended dosage is 2 mg orally once daily, and should be taken at least 1 hour before or at least 2 hours after a meal. However, we recognized from pharmacogenomics knowledge model that more careful attention should be paid to dosing schedules, when medication experience changes or other side effects occur. That is to say, trametinib needs to be stopped permanently in case of fever or interstitial lung disease, taken 1-2 hours before meals in case of metastatic thyroid cancer, and once a day in case of liver injury.

Comparison With Relevant Data Sources

The pharmacogenomics knowledge model included 9 groups of PGx relationships in this model, which can present more potential information than other relevant data sources such as DrugBank, PharmGKB, CTD, and RxNorm, as shown in Table 6.

Table 6. Comparison between pharmacogenomics data sources.

Relationships	DrugBank	PharmGKB[d]	CTD[e]	RxNorm[f]	PGxKM[g]
Drug–Gene	√[a]	√	√	—	√
Drug–Drug	√	√*[b]	—	—	√
Gene–Disease	—[c]		√	—	√
Gene–Mutation	—	√	—	—	√
Drug–Disease	√	√*	√	—	√
Drug–Adverse Reaction	—	—	—	—	√
Drug–Dosage	√	—	—	√	√
Drug–Population	—	—	—	—	√
Adverse Reaction–Dosage	—	—	—	—	√

[a]Have structured data and can be downloaded in the web set.

[b]Have information (unstructured data) for such relationships in the web set.

[c]Have no information for such relationships in the web set.

[d]PharmGKB: Pharmacogenomics Knowledge Base.

[e]CTD: Comparative Toxicogenomics Database.

[f]RxNorm: drug data interaction standard in American Clinical Information System

[g]PGxKM: pharmacogenomics knowledge model.

Limitations and Future Studies

However, there are still some limitations in our study. First, this study aimed to build a pharmacogenomics knowledge model and semiautomatically annotate the corpus using the existing NLP tools. However, we did not validate the feasibility of NLP tools or compare the NLP performance using a benchmark data set, such as clinical records from the Third i2b2 Workshop on NLP Challenges [40] or LabeledIn [41], of labeled indications for human drugs. Our future research will explore BERT–CRF model verification on other standard drug corporas. Second, relation extraction was manually done by the 3 annotators which will place restrictions on the application of pharmacogenomics knowledge model, and an evaluation of automatic relation extraction will be conducted in the future. Common relation extraction methods such as CNN, LSTM, and BERT method will be used to improve extraction efficiency.

In future studies, we also plan to do the following jobs to improve our research. First, a series of other antitumor drugs will be taken into consideration to fill up our framework, such as ceritinib and afatinib for non–small-cell lung cancer. Second, linked data can also be extended to other sources, such as CTD, PharmGKB, and DisGeNET. We hope that this knowledge model for PGx interactions could serve as a framework and a resource for future drug research and development.

Conclusions

A pharmacogenomics knowledge model was constructed for precision medication in our research, which reflected the multidimensional relationships between drug, gene, disease, as well as relationships from gene to drug to dosage or frequency associations. Extraction task for PGx entities has been done using the BERT–CRF model with F1 score of 85.12%. Our pharmacogenomics knowledge model contained 5 semantic types (drug, gene, disease, precise medication, and adverse reaction) and 26 semantic relationships had been defined in detail. Using melanoma caused by *BRAF* gene mutation as an example, we verified the feasibility of this model using the FDA's drug labels and relevant linked data. Finally, we highlighted this knowledge model as a scalable framework for clinicians and clinical pharmacists to adjust drug dosage according to patient-specific genetic variation, and to support pharmaceutical researchers during new drug discoveries.

Acknowledgments

This work is supported by the Special Research Fund for Central Universities-Peking Union Medical College (Grant No. 3332020049), the National Key Research and Development Program of China (Grant No. 2016YFC0901901), the National Natural Science Foundation of China (Grant No. 81601573), National Engineering Laboratory for Internet Medical Systems and Applications (Grant No. NELIMSA2018P02), the Key Laboratory of Knowledge Technology for Medical Integrative Publishing of China, the program of China Knowledge Center for Engineering Sciences and Technology (Medical Knowledge Service System; Grant No. CKCEST-2019-1-10).

Authors' Contributions

HK designed the model, performed the experiments, and wrote this paper. The study was originally conceived of by JL, who also improved the experiments and made modifications to this paper. HK, MW, and LS designed the annotation framework, made the rules of annotation, and analyzed the results. LH guided the study and made modifications to this paper. All the authors wrote and revised the manuscript, and all the authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Prasad K. Role of regulatory agencies in translating pharmacogenetics to the clinics. Clin Cases Miner Bone Metab 2009 Jan;6(1):29-34 [FREE Full text] [Medline: 22461095]
2. Wheeler HE, Maitland ML, Dolan ME, Cox NJ, Ratain MJ. Cancer pharmacogenomics: strategies and challenges. Nat Rev Genet 2012 Nov 27;14(1):23-34. [doi: 10.1038/nrg3352]
3. Scott SA. Clinical Pharmacogenomics: Opportunities and Challenges at Point of Care. Clin Pharmacol Ther 2012 Dec 05;93(1):33-35. [doi: 10.1038/clpt.2012.196]
4. Table of Pharmacogenomic Biomarkers in Drug Labeling. URL: https://www.fda.gov/drugs/science-research-drugs/table-pharmacogenomic-biomarkers-drug-labeling [accessed 2020-04-27]
5. Hida T, Nokihara H, Kondo M, Kim YH, Azuma K, Seto T, et al. Alectinib versus crizotinib in patients with ALK -positive non-small-cell lung cancer (J-ALEX): an open-label, randomised phase 3 trial. The Lancet 2017 Jul;390(10089):29-39. [doi: 10.1016/s0140-6736(17)30565-2]
6. Gay ND, Wang Y, Beadling C, Warrick A, Neff T, Corless CL, et al. Durable Response to Afatinib in Lung Adenocarcinoma Harboring NRG1 Gene Fusions. Journal of Thoracic Oncology 2017 Aug;12(8):e107-e110. [doi: 10.1016/j.jtho.2017.04.025]
7. Zhou G, Zhang J, Su J, Shen D, Tan C. Recognizing names in biomedical texts: a machine learning approach. Bioinformatics 2004 May 01;20(7):1178-1190. [doi: 10.1093/bioinformatics/bth060] [Medline: 14871877]
8. Lafferty J, Mccallum A, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. 2001 Jun Presented at: 18th International Conference on Machine Learning; June 28 to July 1; San Francisco p. 282-289.
9. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research 2011;12:2493-2537 [FREE Full text]
10. Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. 2014 Feb 5. URL: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43905.pdf [accessed 2020-04-05]
11. Schuster M, Paliwal K. Bidirectional recurrent neural networks. IEEE Trans. Signal Process 1997;45(11):2673-2681. [doi: 10.1109/78.650093]
12. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2018 Oct 11. URL: https://www.aclweb.org/anthology/N19-1423.pdf [accessed 2020-04-05]
13. Souza F, Nogueira R, Lotufo R. Portuguese named entity recognition using BERT-CRF. arXiv. 2019 Sep 23. URL: http://arXiv.org/abs/1909.10649 [accessed 2020-04-05]
14. Moon T, Awasthy P, Ni J. Towards lingua Franca named entity recognition with BERT. 2019 Nov 19. URL: http://arXiv.org/abs/1912.01389 [accessed 2020-04-05]
15. Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)-Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. JMIR Med Inform 2019 Sep 12;7(3):e14830 [FREE Full text] [doi: 10.2196/14830] [Medline: 31516126]
16. Ji Z, Wei Q, Xu H. Bert-based ranking for biomedical entity normalization. 2019 Aug. URL: https://arxiv.org/abs/1908.03548 [accessed 2020-04-05]
17. Lee J, Yoon W, Kim S. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. 2019 Jan 25. URL: https://arxiv.org/abs/1901.08746 [accessed 2020-04-05]
18. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. 2019 Apr. URL: https://arxiv.org/abs/1904.05342 [accessed 2020-04-05]
19. Percha B, Altman RB. A global network of biomedical relationships derived from text. Bioinformatics 2018 Aug 01;34(15):2614-2624 [FREE Full text] [doi: 10.1093/bioinformatics/bty114] [Medline: 29490008]
20. Ontology Development 101: A guide to creating your first ontology. CiteSeerX. URL: https://bit.ly/3j6mM5H [accessed 2020-04-16]
21. Lin Y, Mehta S, Küçük-McGinty H, Turner JP, Vidovic D, Forlin M, et al. Drug target ontology to classify and integrate drug discovery data. J Biomed Semantics 2017 Nov 09;8(1):50 [FREE Full text] [doi: 10.1186/s13326-017-0161-x] [Medline: 29122012]

22. Dumontier M, Villanueva-Rosales N. Towards pharmacogenomics knowledge discovery with the semantic web. Brief Bioinform 2009 Mar;10(2):153-163. [doi: 10.1093/bib/bbn056] [Medline: 19240125]
23. Introducing the Knowledge Graph: Things, Not Strings. URL: http://googleblog.blogspot.be/2012/05/introducing-knowledge-graph-things-not.html [accessed 2020-04-16]
24. Meng W, Liu M, Liu J. Safe medicine recommendation via medical knowledge graph embedding. 2017 Oct 16. URL: http://arXiv.org/abs/1710.05980 [accessed 2020-04-05]
25. Ernst P, Siu A, Weikum G. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinformatics 2015 May 14;16:157 [FREE Full text] [doi: 10.1186/s12859-015-0549-5] [Medline: 25971816]
26. Ruan T, Wang M, Sun J, Wang T, Zeng L, Yin Y, et al. An automatic approach for constructing a knowledge base of symptoms in Chinese. J Biomed Semantics 2017 Sep 20;8(Suppl 1):33 [FREE Full text] [doi: 10.1186/s13326-017-0145-x] [Medline: 29297414]
27. Roider HG, Pavlova N, Kirov I, Slavov S, Slavov T, Uzunov Z, et al. Drug2Gene: an exhaustive resource to explore effectively the drug-target relation network. BMC Bioinformatics 2014 Mar 11;15(1):68 [FREE Full text] [doi: 10.1186/1471-2105-15-68] [Medline: 24618344]
28. Sun J, Wu Y, Xu H, Zhao Z. DTome: a web-based tool for drug-target interactome construction. BMC Bioinformatics 2012 Jun 11;13 Suppl 9:S7 [FREE Full text] [doi: 10.1186/1471-2105-13-S9-S7] [Medline: 22901092]
29. Xu B, Shi X, Zhao Z, Zheng W. Leveraging Biomedical Resources in Bi-LSTM for Drug-Drug Interaction Extraction. IEEE Access 2018 Jun;6:33432-33439. [doi: 10.1109/access.2018.2845840]
30. Coulet A, Smaïl-Tabbone M, Napoli A, Devignes MD. Ontology-based knowledge discovery in pharmacogenomics. Adv Exp Med Biol 2011;696:357-366. [doi: 10.1007/978-1-4419-7046-6_36] [Medline: 21431576]
31. Dalleau K, Marzougui Y, Da Silva S, Ringot P, Ndiaye NC, Coulet A. Learning from biomedical linked data to suggest valid pharmacogenes. J Biomed Semantics 2017 Apr 20;8(1):16 [FREE Full text] [doi: 10.1186/s13326-017-0125-1] [Medline: 28427468]
32. DisGNET. URL: https://www.disgenet.org/ [accessed 2020-09-08]
33. ClinVar. URL: https://www.clinicalgenome.org/data-sharing/clinvar/ [accessed 2020-09-08]
34. Kim J, Kim J, Lee H. An analysis of disease-gene relationship from Medline abstracts by DigSee. Sci Rep 2017 Jan 05;7(1):40154 [FREE Full text] [doi: 10.1038/srep40154] [Medline: 28054646]
35. Kim J, Kim J, Lee H. DigChem: Identification of disease-gene-chemical relationships from Medline abstracts. PLoS Comput Biol 2019 May 15;15(5):e1007022 [FREE Full text] [doi: 10.1371/journal.pcbi.1007022] [Medline: 31091224]
36. About DaliyMed. URL: https://dailymed.nlm.nih.gov/dailymed/about-dailymed.cfm [accessed 2020-04-18]
37. About DrugBank. URL: https://www.drugbank.ca/about [accessed 2020-04-18]
38. Liu S, Wei Ma, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. IT Prof 2005 Sep;7(5):17-23. [doi: 10.1109/MITP.2005.122]
39. Pharmacogenomics knowledge model. URL: http://www.phoc.org.cn/PGxKM/ [accessed 2020-04-27]
40. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010 Sep 01;17(5):514-518 [FREE Full text] [doi: 10.1136/jamia.2010.003947] [Medline: 20819854]
41. Khare R, Li J, Lu Z. LabeledIn: cataloging labeled indications for human drugs. J Biomed Inform 2014 Dec;52:448-456 [FREE Full text] [doi: 10.1016/j.jbi.2014.08.004] [Medline: 25220766]

Abbreviations

ATC: Anaplastic thyroid cancer
BERT: Bidirectional Encoder Representations from Transformers
Bi-LSTM: bidirectional long short-term memory
CRF: conditional random field
CTD: the Comparative Toxicogenomics Database
FDA: the US Food and Drug Administration
NLM: the US National Library of Medicine
PGx: pharmacogenomics
PharmGKB: Pharmacogenomics Knowledge Base

Edited by C Lovis; submitted 15.05.20; peer-reviewed by Z He, C Friedrich; comments to author 21.06.20; revised version received 11.08.20; accepted 13.09.20; published 21.10.20.

Please cite as:
Kang H, Li J, Wu M, Shen L, Hou L
Building a Pharmacogenomics Knowledge Model Toward Precision Medicine: Case Study in Melanoma
JMIR Med Inform 2020;8(10):e20291
URL: http://medinform.jmir.org/2020/10/e20291/
doi:10.2196/20291
PMID:33084582

Original Paper

BeyondSilos, a Telehealth-Enhanced Integrated Care Model in the Domiciliary Setting for Older Patients: Observational Prospective Cohort Study for Effectiveness and Cost-Effectiveness Assessments

Jordi Piera-Jiménez[1,2], MSc; Signe Daugbjerg[3], PhD; Panagiotis Stafylas[4], MD, PhD; Ingo Meyer[5], MSc; Sonja Müller[6], MSc; Leo Lewis[7], MSc; Paolo da Col[8], MD; Frans Folkvord[1,9], PhD; Francisco Lupiáñez-Villanueva[1,10], PhD

[1]Open Evidence Research Group, Universitat Oberta de Catalunya, Barcelona, Spain
[2]Department of Research & Development, Badalona Serveis Assistencials, Badalona, Spain
[3]Graduate School of Health Economics and Management, Università Cattolica del Sacro Cuore, Roma, Italy
[4]Medical Research & Innovation (HEALTHINK), Thessaloniki, Greece
[5]PMV Research Group, Universität zu Köln, Köln, Germany
[6]Empirica Gesellschaft für Kommunikations und Technologieforschung GmbH, Bonn, Germany
[7]International Foundation for Integrated Care, Oxford, United Kingdom
[8]IGEA Hospital Trieste, Trieste, Italy
[9]Tilburg School of Humanities and Digital Sciences, Tilburg University, Tilburg, Netherlands
[10]Department of Information and Communication Sciences, Universitat Oberta de Catalunya, Barcelona, Spain

Corresponding Author:
Jordi Piera-Jiménez, MSc
Open Evidence Research Group
Universitat Oberta de Catalunya
Rambla de Poblenou 156
Barcelona, 08018
Spain
Phone: 34 651041515
Email: jpieraj@uoc.edu

Abstract

Background: Information and communication technology may provide domiciliary care programs with continuity of care. However, evidence about the effectiveness and cost-effectiveness of information and communication technology in the context of integrated care models is relatively scarce.

Objective: The objective of our study was to provide evidence on the clinical effectiveness and cost-effectiveness of the BeyondSilos project for patients enrolled in the Badalona city pilot site in Spain.

Methods: A quasi-experimental study was used to assess the cost-effectiveness of information and communication technology–enhanced integration of health and social care, including the third sector (intervention), compared to basic health and social care coordination (comparator). The study was conducted in Badalona between 2015 and 2016. Participants were followed for 8 months.

Results: The study included 198 patients: 98 in the intervention group and 100 in the comparator group. The mean Barthel index remained unchanged in the intervention group (mean change 0.14, 95% CI –4.51 to 4.78; *P*=.95) but decreased in the comparator group (mean change –3.23, 95% CI –5.34 to –1.11; *P*=.003). Instrumental Activities of Daily Living significantly decreased in both groups: mean changes of –0.23 (95% CI –0.44 to –0.02; *P*=.03) and –0.33 (95% CI –0.46 to –0.20; *P*<.001) in the intervention and comparator groups, respectively. No differences were found in the Geriatric Depression Scale (intervention: mean change 0.28, 95% CI –0.44 to 1.01, *P*=.44; comparator: mean change –0.29, 95% CI –0.59 to 0.01, *P*=.06). The intervention showed cost-effectiveness (incremental cost-effectiveness ratio €6505.52, approximately US $7582).

Conclusions: The information and communication technology–enhanced integrated domiciliary care program was cost-effective. The beneficial effects of this approach strongly rely upon the commitment of the professional staff involved.

Trial Registration: ClinicalTrials.gov NCT03111004; http://clinicaltrials.gov/ct2/show/ NCT03111004

(JMIR Med Inform 2020;8(10):e20938) doi:10.2196/20938

KEYWORDS

integrated care; telemedicine; telecare; digital health; cost-effectiveness; clinical effectiveness; chronic disease

Introduction

Background

Domiciliary care programs are increasingly used to deliver health care to patients—particularly older patients and those with chronic conditions—who are unable to go to a primary care center due to their medical condition or disability, thus improving their health and functional independence, while reducing hospitalizations [1-4]. Among domiciliary care programs, integrated care models prioritize continuity in the sense that the same care provider supports the patient both at home and the primary care center. However, the need for integration with social care is often undervalued [5]. The relevance of social care is not limited to the role of social workers, but also that of stakeholders in the third sector, which in some areas may strongly contribute to day-to-day welfare of these patients [6].

Regardless of the involvement of stakeholders from the social domain, integrated domiciliary care models face the challenge of being efficient enough to absorb the rapidly rising number of care recipients in this setting, likely prompted by social and demographic shifts [7]. In fact, the current overloaded schedule of primary care teams involved in integrated domiciliary care programs has been already identified as a significant drawback of this care model [8,9].

Among the interventions designed to increase the efficiency of health care systems, the use of information and communication technologies have shown promising results in various areas, including the management of older people with chronic diseases [10-12]. Besides integrating all patient information and facilitating the coordination of the various professionals involved, information and communication technology provides domiciliary care with telemonitoring solutions, which may bring patients and professionals closer [13]. However, the evidence regarding the cost-effectiveness of these solutions in the context of integrated care models is scarce and heterogeneous in terms of quality [7,14,15].

The BeyondSilos Project

BeyondSilos aimed to promote community-based, independent lives by providing domiciliary care with information and communication technology solutions capable of crossing through domain boundaries that typically separate social and health care providers [16]. One of the key areas of integration (frequently referred as to horizontal integration) was for the common access of all cross-sectorial care teams, including those of the third sector, to telehealth platforms in order to improve coordination and promote continuity of care.

To overcome the traditional boundaries separating social and health care, information and communication technology solutions of the BeyondSilos project went hand-in-hand with innovative organizational designs. This approach was based on the assumption by Urošević and Mitić, who pointed out that "Successful service integration in policy and practice requires both technology innovation and service process innovation being pursued and implemented at the same time [17]." Because information and communication technology–based services are typically delivered within sociotechnical system (ie, organizational frameworks where people interact with technology), their success often depends on the value of people applying technology. Hence, information and communication technology can effectively support well-designed care service delivery processes, but it cannot replace them because of the emotional aspects of physical meetings [18].

The first step in achieving a combined innovation approach was the development of common integrated care pathways that were to be supported by information and communication technology. For this purpose, the project adopted 2 generic service pathways of the SmartCare project which were adapted to fit local context through service process modeling techniques (Multimedia Appendix 1). The first pathway addressed needs for integrated home care during an acute episodes and immediately after hospital discharge. The second pathway was directed toward people needing integrated long-term care (eg, frail patients with multiple comorbidities).

We hypothesized that the provision of information and communication technology–enhanced integrated care services that encompass health and social care in the setting of domiciliary care would improve health outcomes and reduce health system costs. Herein, we report the clinical effectiveness and cost-effectiveness of the BeyondSilos intervention for patients enrolled in the long-term pathway in a Badalona city pilot site (Spain).

Methods

Study Design, Setting, and Participants

As part of the BeyondSilos project, an observational prospective cohort study was carried out to assess the implementation of an information and communication technology–enhanced integrated care model in the setting of domiciliary care in *Badalona Serveis Assistencials* (BSA), a public provider of health and social care services to the City Council of Badalona, the most populated suburban area to the north of Barcelona, Spain with a reference population of 433,175 inhabitants. BSA has recently been shifting toward integrated care models [19-30].

In Spain, the health and social care systems are centrally managed by the Ministry of Health, Consumerism, and Social Services, which provides the basic regulations and guidelines. The political control and jurisdiction over the organization and provision of health and social services are transferred to the 17 regional governments (autonomous communities). The health system is based on a Beveridge model, characterized by universal coverage, funded by the government through tax, and delivered by an extensive network of public and private health providers. The regions have the main responsibility for social services provision, together with municipalities [31,32]. Third-sector organizations (voluntary and nonprofit) play an essential role in responding to many and different social needs of the general population that are beyond the reach of the scarce public resources (eg, volunteer care and accompaniment of those at risk of social exclusion and isolation) [33].

Care recipients assessed for eligibility were involved in a domiciliary care program as described by Burgos-Díez et al (study condition) [19] and were recruited among care recipients managed from 6 primary care centers. Centers acting as intervention and comparator were paired 1-to-1 for similar socioeconomic status in their area of influence. To this end, candidate sites were stratified into 3 categories of socioeconomic status of the catchment area (2 primary care centers per category). The information and communication technology–enhanced integrated care model (intervention) was first introduced in 1 center in each category; the remaining centers were used as comparators. The first care recipient was enrolled March 3, 2015, and the last care recipient exited the project October 20, 2016.

Eligibility Criteria

The main inclusion criteria were age ≥65 years, special health needs due to the presence of chronic diseases (ie, heart failure, stroke, diabetes, or chronic obstructive pulmonary disease plus at least 1 additional chronic disease included in the Charlson Comorbidity Index [34]), and the need for social care based on Barthel Index of Activities of Daily Living [35] and Instrumental Activities of Daily Living. To be assessed for eligibility, patients were not require to have an active internet or mobile contract but had to have reliable 4G coverage at home (required by the telehealth solution provided). Participants with an active cancer or AIDS diagnosis, in a terminal state, those who had undergone an organ transplant, or who were on dialysis before enrolment were excluded from the study.

Ethics

The study protocol was approved by the Independent Ethics Committee of the *Hospital Germans Trias i Pujol*, and all participants provided informed consent before entering the study.

Intervention

Participants from both groups received health and social care, integrated through a corporate enterprise resource planner which was used as a facilitator for administrative coordination between BSA and the municipality (ie, management of admissions and discharges). Health and social care information were stored in 2 centralized repositories linked to each other through interoperability. Domiciliary care was coordinated using a homecare department software, which stored the Shared Care Plan, accessible for both health and social care professionals. Based on this Plan, professionals scheduled regular visits or phone contacts with care recipients.

In addition to the aforementioned common resources, participants in the intervention group were provided with a telehealth platform, the Health Insight Solutions Homecare Platform, which included the following components: security sensors (ie, fire and water detectors, behavioral movement sensors, and a cell phone with GPS tracking and fall detection), medical devices (ie, weight scale, blood pressure meter, glucometer, and oximeter), serious games, a personal diary, and a videoconferencing system (Multimedia Appendix 2). The telehealth platform was used by the participants and their close relatives to continuously track their health status following the care plan defined by their formal caregivers. Information collected within the telehealth platform was checked daily by the primary care team responsible for the patient. Exacerbation of health conditions (eg, weight increase over 20% in a 1-week period) and out-of-hours alarms (ie, fall detection, fire, or water leak) automatically triggered an alert (SMS text message) to the team on call. In the intervention group, third-sector care providers had access to basic clinical information (ie, main diagnostics and visits from other professional staff) throughout the Shared Care Plan and provided volunteer accompaniment support to patients at risk of social exclusion.

Recruitment

Potential study participants were identified in a 2-stage process. The first part of the process was conducted by the Information Systems Department of BSA and consisted of identifying possible candidates through a database search using the inclusion and exclusion criteria. The initial selection process identified 4800 possible candidates receiving both health and social care services. Applying more specific inclusion criteria, such as diagnosis-based specificities, reduced the list to 430 patients. In a second stage, research assistants in each participating center approached the individuals and asked them if they were willing to participate.

Assessments

The effectiveness of the intervention was evaluated using the Model for Assessment of Telemedicine [36]. Primary outcome measures were related to the health status of study participants and established based on the Barthel index scale, the Instrumental Activities of Daily Living scales [37], and the Geriatric Depression Scale [38]. All questionnaires were collected online by trained researchers using a purpose-designed survey built on an open-source tool (LimeSurvey; Limesurvey GmbH) [39].

Costs were modeled and collected from both a health care and societal perspective using the ASSIST Tool [40] and were estimated in 2016 euros. For the intervention group, 2 types of costs were considered: one-off costs (ie, incurred only at implementation) and recurring costs (ie, costs derived from the service practice).

The health care costs perspective included the assessment of resource utilization and considered all characteristics regarding hospitalization (eg, number of admissions and readmissions, length of hospital stay, and type of admission) and contacts with health and social care professionals (eg, type of professional, number of contacts, and type and setting of the contact). For personnel costs, the average income for 1 full - time employee with employer contributions to social security was used. The average hourly wages were €29.23 for a physician (approximately US $34.07), €20.79 for a nurse (approximately US $24.23), and €18.19 for a social care worker (approximately US $21.20).

The societal cost perspectives considered were the health care costs plus those outside the health care sector. In this case, the costs for the intervention group included the time spent by patients using the new service. Moreover, the intervention brought savings in travel time and costs for patients and their caregivers. These were computed as a cost for the control group. The monetary equivalent for the time spent by the patients and informal caregivers was calculated using the minimal interprofessional wages for the year 2016 and resulted in an hourly wage of €6.07 (approximately US $7.07).

All costs were homogenized per patient and per year. Bed days of each group were multiplied by the estimated cost per bed - day in Spain (€733.56, approximately US $854.91).

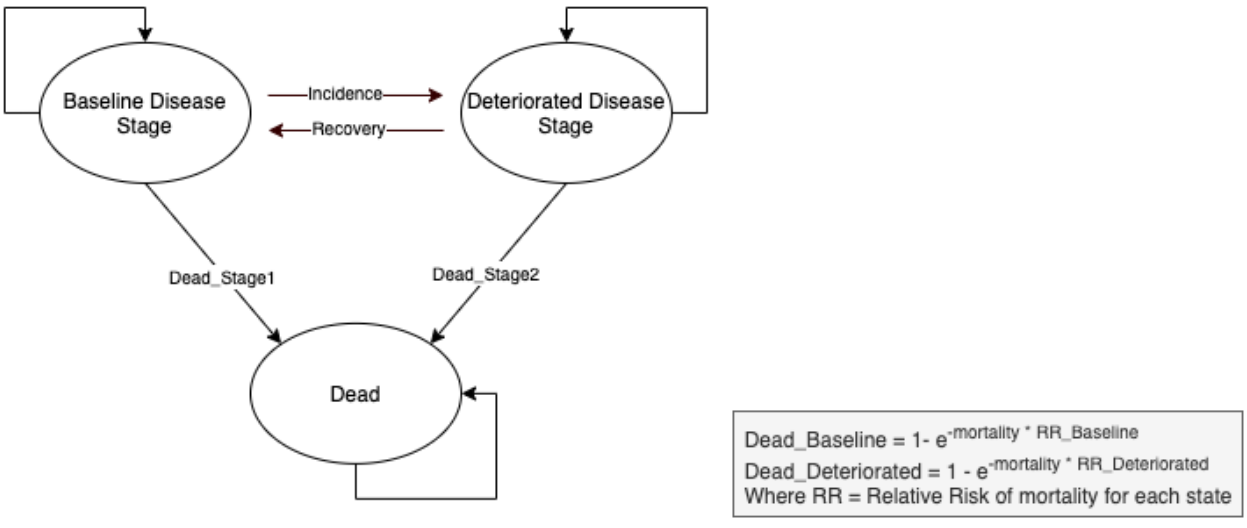
Analyses

Categorical variables were described as frequency and percentage of available data, whereas quantitative variables were described as mean and standard deviation or median and interquartile range; 95% confidence intervals were provided for mean differences. Between-group differences regarding the proportions of each category were compared using the chi-square test, whereas quantitative variables were compared using the *t* test, analysis of variance, or their nonparametric counterparts (Mann-Whitney *U* test and Kruskal-Wallis test, respectively). Normality was assessed using the Kolmogorov-Smirnov test [41]. In all comparisons, the significance threshold was set at a 2-sided α=.05. Descriptive and comparative analyses were performed using SPSS software (version 17.0; SPSS Inc).

Cost-effectiveness analysis was performed using Monitoring and Assessment Framework for the European Innovation Partnership on Active and Healthy Ageing (MAFEIP) [42]. MAFEIP is a free web-based tool promoted by the European Commission aimed at performing cost-utility analysis to estimate health outcomes and resource usage of a large sample of information and communication technology–enabled health and social care innovations, developed and implemented in the context of the European Innovation Partnership on Active and Healthy Ageing [42,43]. More precisely, the cost-effectiveness estimates are based on the principles of decision analytic modeling and a generic Markov model which provides the flexibility required to be tailored to the variety of solutions promoted by the European Innovation Partnership on Active and Healthy Ageing [44-46].

Quality-adjusted life years were computed using change in the Barthel index as a proxy of utility as described by Kaambwa et al [47], and based on a 3-states Markov model: *baseline disease stage* (the patient remains in the same state or improves), *deteriorated disease stage* (the patient worsens), and *dead* (Figure 1). The 3 states led to the corresponding transition probabilities: *recovery* (improving or remaining the same state), *incidence* (worsening), and *death*. Mortality rates were internally calculated by the MAFEIP tool using all-cause mortality rates (age- and sex-dependent) extracted from the Human Mortality Database. Discount factors for health outcomes and costs were both set to 3% following recommendations from local Health Technology Assessment authorities. In order to estimate the incremental costs and outcomes associated with the intervention, we ran the model over a 40-year time horizon, following the proposed standardization for economic analysis of health technologies in Spain, which recommends assessing the costs and benefits on a time horizon that covers the entire lifespan of the patients affected [48-50].

Figure 1. 3-state Markov model applied for the BeyondSilos cost-effectiveness analysis.



Results

Participant Characteristics

Of the 268 individuals considered for eligibility, 70 were excluded, resulting in a study sample of 198 patients: 98 (49.5%) were managed within the BeyondSilos project (intervention group) and 100 (50.5%) were managed according to usual care (comparator group) (Figure 2).

Figure 2. Flowchart of participant recruitment for the BeyondSilos project.

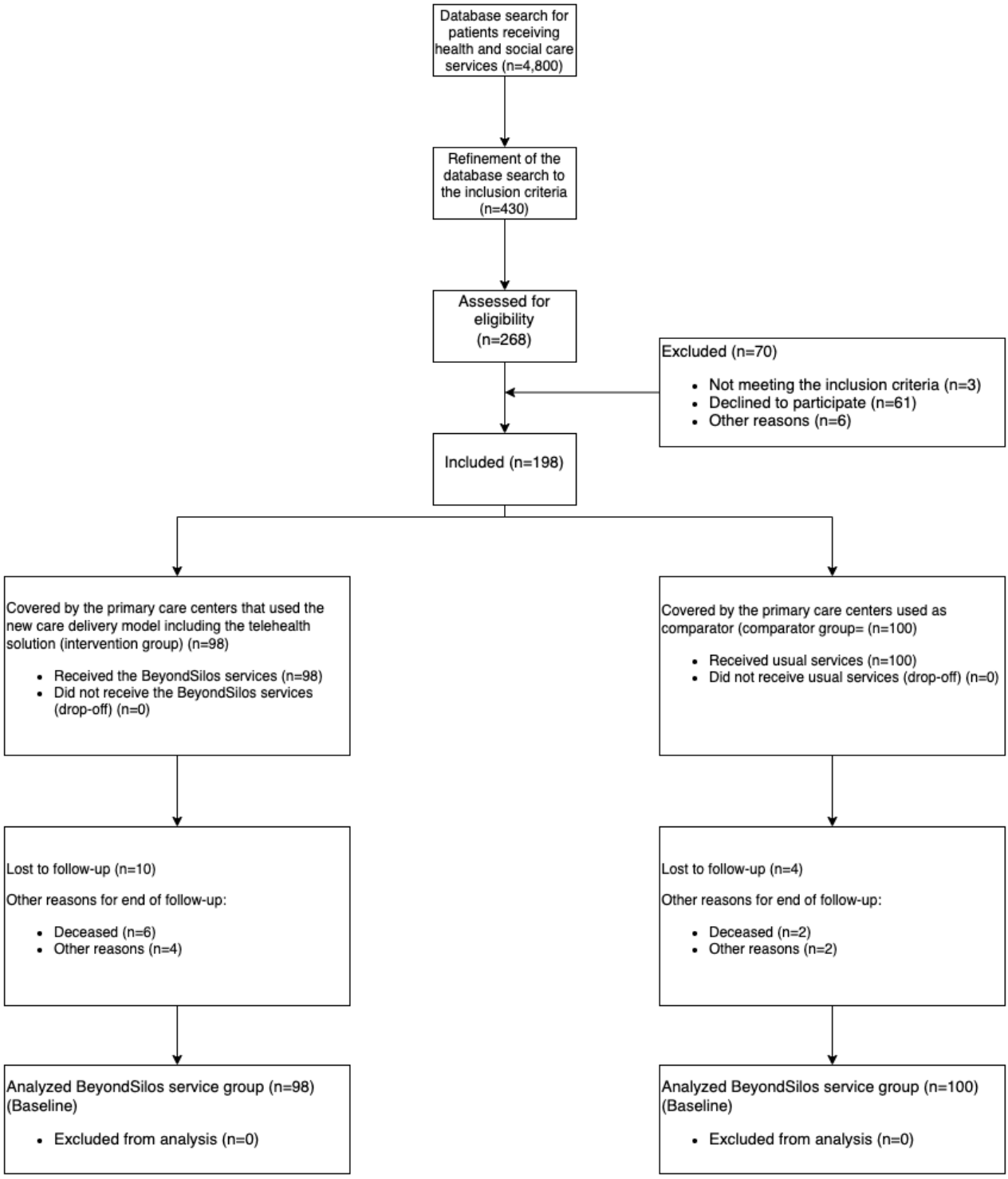


Table 1 summarizes the sociodemographic characteristics of study participants. Participants in the 2 study groups were balanced regarding education level and household income, but the intervention group tended to be overrepresented by older, female, and widowed individuals.

Table 1. Demographic baseline characteristics of the total sample (N=198).

Characteristics	Intervention group (n=98)	Comparator group (n=100)	*P* value
Gender, n (%)			.02
Male	26 (26)	43 (43)	
Female	72 (74)	57 (57)	
Age (years), median (IQR)	85.5 (7.3)	82.8 (8.3)	.01
Age group (years), n (%)			.003
<65	1 (1.0)	0 (0)	
65-75	7 (7.1)	24 (24.0)	
>75	90 (91.8)	76 (76.0)	
Marital status, n (%)			.053
Never married	2 (2.0)	3 (3.0)	
Currently married	30 (30.6)	47 (47.0)	
Separated	0	3 (3.0)	
Divorced	2 (2.0)	1 (1.0)	
Widowed	63 (64.3)	46 (46.0)	
Cohabitating	1 (1.0)	0 (0)	
Education level, n (%)			.23
Less than primary school	50 (53.2)	40 (40.8)	
Primary school	30 (31.9)	42 (42.9)	
Secondary school	5 (5.3)	10 (10.2)	
High school	7 (7.4)	4 (4.1)	
College/university	2 (2.1)	1 (1.0)	
Post graduate degree	0 (0)	1 (1.0)	
Household income (€[a] yearly), median (IQR)			.96
0-6999	7 (13.2)	11 (14.7)	
7000-13,999	32 (60.4)	45 (60.0)	
14,000-19,999	12 (22.6)	15 (20.0)	
20,000 or more	2 (3.8)	4 (5.3)	

[a]An approximate exchange rate of €1 to US $1.17 was applicable at the time of publication.

At baseline, study participants in both groups had a median of 3 comorbidities (IQR 2-4), with no significant differences regarding either the number of comorbidities (*P*=.96) or the prevalence of each comorbidity, except malignancies, which were 2.6-fold more frequent among those in the intervention group (Table 2). The mean Charlson Comorbidity index was 4.42 (SD 2.34) and 4.31 (SD 1.81) for the intervention and comparator groups, respectively (*P*=.79). Congestive heart failure was the most prevalent comorbidity in both study groups. The intervention group had significantly lower Barthel index scores (*P*=.001) and higher Geriatric Depression Scale scores (*P*=.002). This trend was not observed for the Instrumental Activities of Daily Living (*P*=.44).

Table 2. Clinical characteristics of study participants at baseline (N=198).

Characteristics	Intervention group (n=98)	Comparator group (n=100)	*P* value
Comorbidities, n (%)			
Myocardial infarction	17 (17.3)	23 (23.0)	.32
Congestive heart failure	61 (62.2)	71 (71.0)	.19
Peripheral vascular disease	1 (1.0)	3 (3.0)	.33
Cerebrovascular disease	43 (44.3)	25 (25.0)	.004
Dementia	3 (3.1)	5 (5.0)	.49
Chronic pulmonary disease	1 (1.0)	3 (3.0)	.33
Rheumatic disease	3 (3.1)	10 (10.0)	.051
Peptic ulcer disease	19 (19.6)	16 (16.0)	.51
Mild liver disease	22 (22.7)	34 (34.0)	.08
Diabetes without chronic complication	25 (26)	27 (27.0)	.88
Diabetes with chronic complication	31 (32.0)	19 (19.0)	.04
Hemiplegia or paraplegia	28 (28.9)	37 (37.0)	.22
Renal disease	1 (1.0)	1 (1.0)	.99
Malignancies[a]	23 (23.7)	9 (9.0)	.005
Moderate or severe liver disease	3 (3.1)	4 (4.0)	.73
Metastatic solid tumor	13 (13.4)	12 (12.0)	.77
Anthropometric and laboratory exams, mean (SD)			
Body mass index (kg/m^2)	28.8 (4.8)	27.3 (5.4)	.02
Blood glucose (mg/dL)	110.8 (34.6)	116.9 (44.5)	.44
HbA_{1c}[b] (%)	6.82 (1.70)	7.45 (1.81)	.11
eGFR ($mg/dL/1.73\ m^2$)	75.9 (38.1)	74.4 (43.2)	.40
Tobacco use, n (%)			.12
Never	75 (79.8)	69 (69.0)	
Former	19 (20.2)	29 (29.0)	
Current smoker	0 (0)	2 (2.0)	
E-cigarette	0 (0)	0 (0)	
Other	0 (0)	0 (0)	
Alcohol drinking (weekly drinks past 12 months), n (%)			.02
None	87 (88.8)	80 (80.0)	
<1	6 (6.1)	6 (6.0)	
1-7	3 (3.1)	14 (14)	
8-14	2 (2.0)	0 (0)	
15-21	0 (0)	0 (0)	
>21	0 (0)	0 (0)	
Assessment scores, mean (SD)			
Barthel index	44.66 (27.37)	71.58 (27.95)	.001
Instrumental Activities of Daily Living	1.45 (1.74)	2.94 (2.55)	.44
Geriatric Depression Scale	7.23 (3.47)	6.11 (3.51)	.002

[a]Any malignancy, including lymphoma and leukemia, except malignant neoplasm of skin.

[b]HbA_{1c}: glycohemoglobin.

[c]eGFR: estimated glomerular filtration rate.

Clinical Effectiveness

The Barthel index remained unchanged throughout the follow-up period in the intervention group (mean change from enrolment to end was 0.14, 95% CI –4.51 to 4.78; *P*=.95), but decreased in the comparator group (mean change –3.23, 95% CI –5.34 to –1.11; *P*=.003). The score of the Instrumental Activities of Daily Living significantly decreased in both groups: mean change of –0.23 (95% CI –0.44 to –0.02) in the intervention group (*P*=.03) and –0.33 (95% CI –0.46 to –0.20) in the comparator group (*P*<.001). The Geriatric Depression Scale score did not significantly change in either the intervention group (mean change 0.28, 95% CI –0.44 to 1.01; *P*=.44) or the comparator group (mean change –0.29, 95% CI –0.59 to 0.01; *P*=.06). None of the deaths were deemed to be related to the intervention or likely to been preventable with the intervention.

Resource Utilization

During the 8 months of follow-up, the study participants contacted the health care or social professionals 5209 times: 2556 times in the intervention group and 2653 times in the comparator group. The contact profile of the 2 groups differed significantly regarding the type of professional, the planned/unplanned contact, and the setting of contacts (Table 3). Overall, participants in the intervention group tended to have contact more with their general practitioner and the social worker, and less with the specialists. Regarding the type of visit, participants in the intervention group tended to have more planned visits, predominantly at home, compared to those of the comparator group.

Table 3. Resource utilization of study participants (N=198).

Resource use	Intervention group (n=98)	Comparator group (n=100)	*P* value
Hospitalization			
Hospitalized patients, n (%)	32 (32.7)	45 (45.0)	.08
Length of hospital stay per admission (days), mean (SD)	5.84 (8.81)	2.3 (2.8)	.02
Length of hospital stay per patient (days), mean (SD)	12.9 (15.0)	6.36 (9.0)	.02
Time to first admission (days), mean (SD)	56.3 (57.9)	70.8 (59.3)	.31
Admissions per patient (all patients), mean (SD)	0.85 (1.61)	1.12 (2.10)	.17
Readmissions within 30 days per patient, mean (SD)	1.73 (1.78)	2.11 (2.74)	.96
Type of admission, n (%)			.63
Planned	24 (28.9)	36 (32.1)	
Unplanned	59 (71.1)	76 (67.9)	
Annual length of hospital stay (unplanned admissions), mean (SD)	1.58 (5.15)	0.65 (1.41)	.74
Interaction with health and social professional			
Type of professional, n (%)			
General practitioners	895 (34.2)	670 (23.3)	<.001
Specialists	116 (4.4)	225 (7.8)	<.001
Nurses	1504 (57.5)	1901 (66.1)	<.001
Other health care provider	25 (1.0)	39 (1.4)	.17
Social workers	76 (2.9)	42 (1.5)	<.001
Volunteers	N/A[a]	N/A	N/A
Type of anticipation, n (%)			<.001
Planned	1677 (93.2)	1359 (87.6)	
Unplanned	123 (6.8)	193 (12.4)	
Setting of contacts, n (%)			
Physical meeting out of home	239 (9.4)	563 (21.2)	<.001
Home visit	1089 (42.6)	687 (25.9)	<.001
Telephone	759 (29.7)	535 (20.2)	<.001
Writing (email, SMS text message, etc)	463 (18.1)	857 (32.3)	<.001
Other	6 (0.2)	8 (0.3)	.82
Annual rates for contacts, mean (SD)			
Annual contacts rate	51.0 (36.1)	53.1 (40.3)	.85
Annual unplanned contacts rate	2.4 (3.5)	3.8 (5.3)	.07
Annual physical contacts rate	24.9 (23.5)	23.4 (18.1)	.66

[a]N/A: not applicable.

Cost-Effectiveness Analysis

Table 4 summarizes the costs with transition probabilities between the 3 health states of the Markov model used as inputs for the MAFEIP tool. Although the expenditures shared by the 2 care models were very similar, the intervention group was associated with an extra cost, resulting in an incremental cost of €4755 (approximately US $5542). The increase of costs was associated to the extra home visits and general practitioner contacts associated to the training and usage of the telehealth technology (for a detailed table of costs see Multimedia Appendix 3).

Table 4. Input used for the cost-effectiveness analysis based on the 3-state Markov model (N=198).

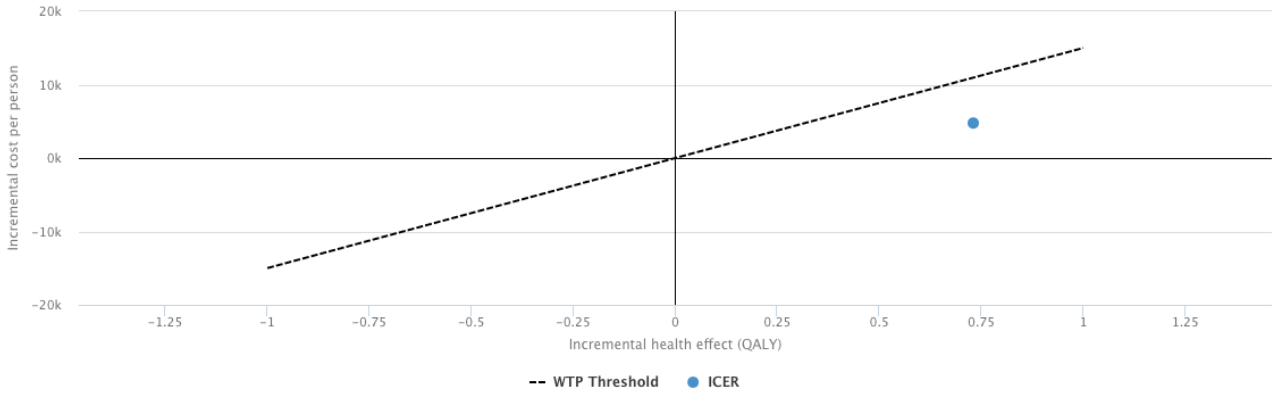
Input	Intervention group (n=98)	Comparator group (n=100)
Transition probabilities, %		
Incidence	34	36
Recovery	66	64
Relative risk (mortality)		
Baseline disease stage	1.005	1.005
Deteriorated disease stage	1.005	1.005
Utility after intervention		
Baseline disease stage	0.56	0.45
Deteriorated disease stage	0.3	0.33
Costs, € ($)[a]		
One-off cost per patient (intervention)	1268.89 (1484.60)	N/A[b]
Recurring cost per patient/year (intervention)	230.40 (269.57)	N/A
Health care cost—baseline disease stage	5664.89 (6627.92)	5198.62 (6082.39)
Health care cost—deteriorated disease stage	4502.89 (5268.38)	5221.69 (6109.38)
Societal cost—baseline disease stage	5953.15 (6965.19)	5259.14 (6153.19)
Societal cost—deteriorated disease stage	4791.15 (5605.65)	5282.21 (6180.19)

[a]An approximate exchange rate of €1 to US $1.17 was applicable at the time of publication.

[b]N/A: not applicable.

The effectiveness, computed based on transition probabilities between the 3 states of the Markov model, was also higher in the intervention group, yielding an incremental effect of 0.731. Overall, the incremental cost-effectiveness ratio was €6505.52 (approximately US $7582), making the intervention more effective than usual care for all willingness-to-pay thresholds above €6500 (approximately US $7575) per quality-adjusted life year (Figure 3).

Figure 3. Cost-effectiveness plane for a willingness-to-pay of €15,000 (approximately US $17,481)/quality-adjusted life year. ICER: incremental cost-effectiveness ratio; QALY: quality-adjusted life year; WTP: willingness-to-pay.



The sensitivity analysis showed that a change between 0% and 5% in the utility in the baseline health for the intervention group would place the incremental cost-effectiveness ratio still below the willingness-to-pay threshold of €15,000 (approximately US $17,481)/quality-adjusted life year (Figure 4).

Similarly, a change between 0% and 5% in the health care costs would not affect the result (Figure 5).

Figure 4. Sensitivity analysis showing effects between 0% and 5% change in utilities—willingness-to-pay of €15,000 (approximately US $17,481)/quality-adjusted life year. QALY: quality-adjusted life year; WTP: willingness-to-pay.

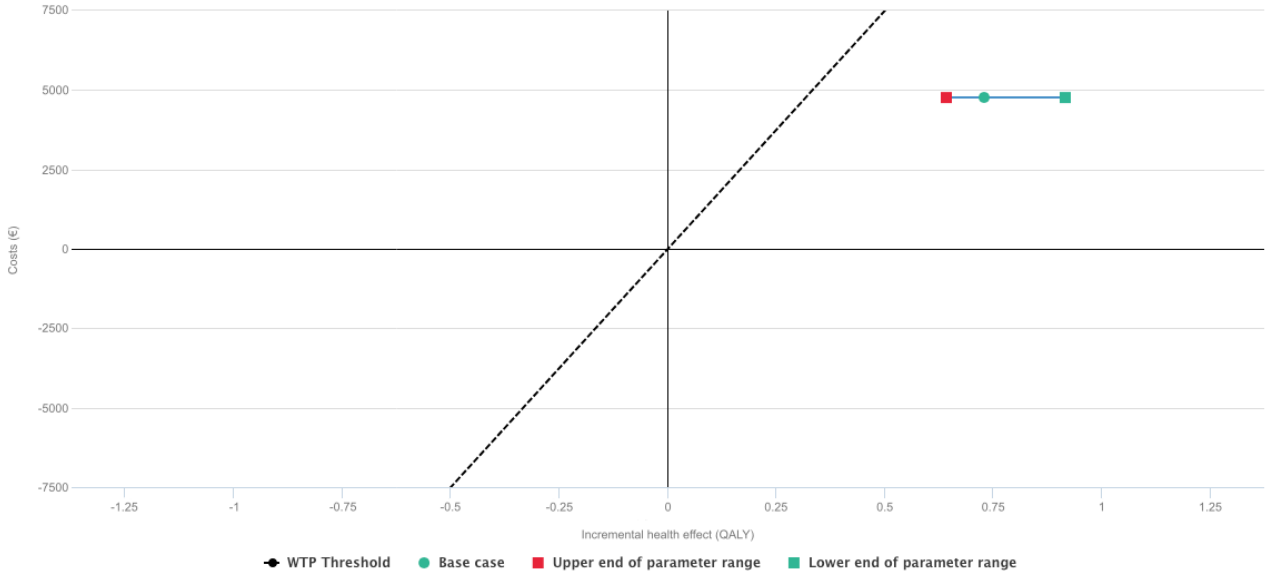
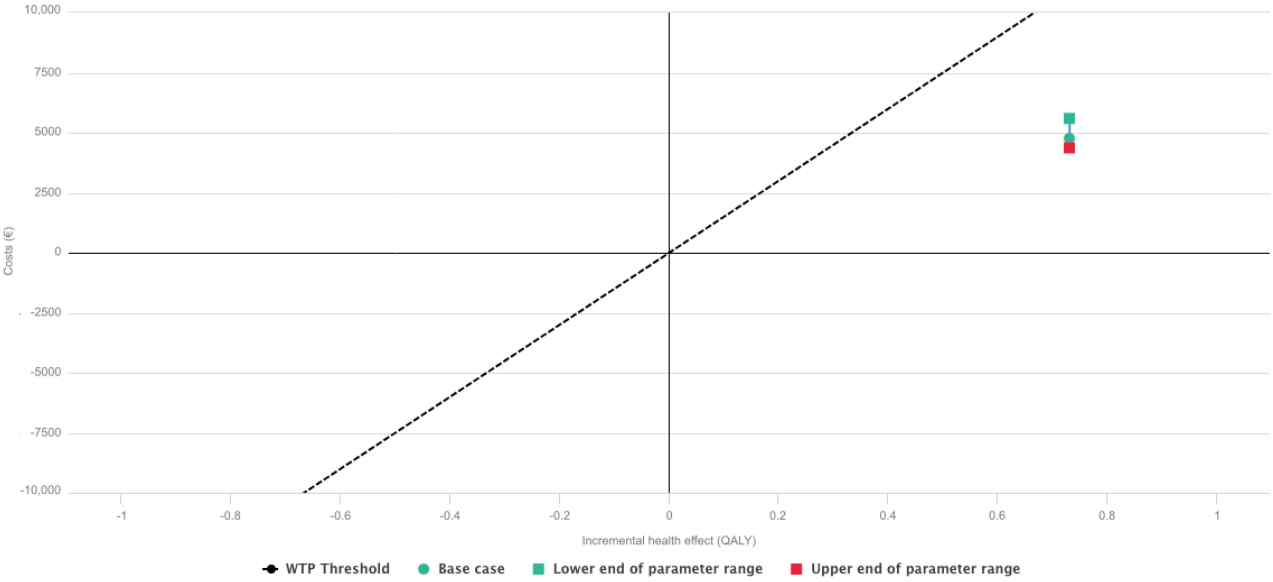


Figure 5. Sensitivity analysis showing effects between 0% and 5% change in costs—willingness-to-pay of €15,000 (approximately US $17,481)/quality-adjusted life year. QALY: quality-adjusted life year; WTP: willingness-to-pay.



Discussion

Summary of Main Results

In this observational prospective cohort study, we found that the addition of an information and communication technology solution (which also involved the third sector) to a basic integrated care model was more effective than integration only in terms of transition between health states established with the Barthel index and the Instrumental Activities of Daily Living. The superiority of the BeyondSilos intervention was confirmed by all willingness-to-pay thresholds above €6500 (approximately US $7575) per quality-adjusted life year, far below the €30,000 (approximately US $34,963) threshold traditionally considered in Spain [51].

Besides the specific context of the pilot site, these results must be interpreted by bearing in mind the challenges of assessing cost-effectiveness of a complex intervention (such as an integrated care model). First, the complexity of both the intervention and the usual care model (in this case, an integrated care framework) often blurs the different contribution of each element to costs [7,15]. This also applies to stakeholders in the third sector (volunteer care), which cannot be easily quantified. Furthermore, it is worth mentioning that quality-adjusted life years may not always be a useful indicator for decision making at the level of provider organizations, particularly when (1) the delivery of care is already constrained by decisions at national or regional level [50] and (2) additional factors such as patient and provider satisfaction need to be taken into account.

Contextualization of the Badalona Pilot Within the BeyondSilos Project

An important characteristic of projects aimed at implementing integrated care strategies is the need of tailor the overarching plan and methodology to the organizational framework of each area. Therefore, considering the expected differences between pilot sites in this regard, the original purpose was to provide a general integrated care framework so that pilot sites could tailor it to their health care environment. The most remarkable characteristic of Badalona pilot site was that, unlike other pilot sites enrolled in the BeyondSilos project, it was already delivering both health and social care services based on an integrated care approach. In this context, the BeyondSilos project added only 2 remarkable improvements compared with usual care: (1) a deeper commitment of the third-sector organizations and (2) the use of information and communication technology to enhance domiciliary care. The fact that the pilot site already operated under an integrated care approach had the advantage that the health care team was already used to integrated pathways, thus facilitating the incorporation of additional integrated care elements into the organizational model. However, this feature brought the trial to a challenging scenario in which the comparator (ie, comparator group) already included social care within the integrated care approach, thus reducing the benefits of the BeyondSilos model.

Strengths, Limitations, and Future Work

Our analysis was strengthened by the appropriate balance between the primary care centers that piloted the information and communication technology–based intervention and those acting as comparators (paired by socioeconomic status). Although this approach did not preclude baseline differences in some demographic and clinical characteristics, the study groups were balanced regarding sociodemographic characteristics that may influence attitudes toward information and communication technology, such as income and education level.

On the other hand, studies investigating the effectiveness of integrated care models have to deal with the difficulty of establishing an adequate comparator [7]. As a rule of thumb, usual care is the recommended comparator, but this had different meanings for the various pilot sites in the BeyondSilos project, with some comparing nonintegrated and integrated care models, and others—as in our pilot site—comparing 2 integrated care models with different intensities. The last approach has been increasingly used as more areas adopt integrated care approaches [52,53], although there is less room for improvement. Another challenge of the assessment of integrated care models includes patient profiles, often characterized by a multimorbid conditions, which may be rather heterogeneous [7,15]. In our study, the baseline demographic and clinical characteristics of patients in the 2 groups were similar, but patients in the intervention group tended to be female, older, widowed, more dependent, and with higher depression scores. These differences, likely because of the real-life setting, should be carefully considered when appraising the scope of our results. Specifically, the characteristics of the intervention group might be associated with a higher need of formal care and information and communication technology solutions than that in the control group, thus potentially shading the actual benefits of the intervention.

Keeping these limitations in mind, we found that the frequency of planned and home visit contacts was significantly higher in the intervention group ($P<.001$). Although this trend might be influenced by the higher complexity of patients in the intervention group, health care professionals explicitly explained that the usage of information and communication technology required more of their time and they were afraid that information and communication technology may replace their jobs. This attitude, together with the usual resistance of care recipients to losing contact with their formal caregivers [54,55], was likely to hinder the reduction of home visits that is expected with telemonitoring. Of note, the lack of differences in the estimated annual rates suggests that this phenomenon was not homogeneous throughout the follow-up period, being more pervasive during the first stages of the intervention. The temporal patterns of this attitude may reflect a certain resistance of professional staff to trust the new information and communication technology–supported integrated care model (ie, not fully taking advantage of the telemonitoring solution thus not abandoning the routine cadence of home care visits). Besides being a lesson for future implementation of information and communication technology solutions, this observation suggests that, in our study, uncontrolled factors such as the personal commitment of professionals to the project might influence the apparent cost-effectiveness of an information and communication technology solution, potentially overriding other factors such as patient characteristics. Future evaluations based on multicriteria decision analyses may provide interesting insights regarding the implementation of information and communication technology–enhanced integrated care programs [56].

Conclusion

Our study provided evidence regarding the clinical effectiveness and cost-effectiveness of an information and communication technology–enhanced integrated care model that enables telemonitoring and increases the intensity of integrated care by involving organizations of the third sector in the management of older patients in a domiciliary care setting. The cost-effectiveness analysis placed the intervention as more effective than usual care—and reasonably inexpensive. However, our findings confirm the difficulties of assessing the effectiveness of interventions and suggest that the beneficial effects of a new care model strongly depend on the commitment of health and social care professionals with the model.

Acknowledgments

The BeyondSilos study was cofunded by the European Commission's Competitiveness and Innovation Program (grant number: 621069). The funder had no role in data collection or analysis, the decision to publish, or the preparation of the manuscript. The

authors would like to thank Gerard Carot-Sans, PhD, for providing medical writing support and the BeyondSilos consortium members.

The development of the MAFEIP tool is funded by the European Commission's Horizon 2020 program under the projects WE4AHA (grant number: 769705) and DigitalHealthEurope (grant number: 826353).

Authors' Contributions

JP-J, SD, PS, IM, SM, LL, and PC contributed to the study design. JP-J was the principal investigator locally and managed the overall trial and data collection. JP-J, SD, PS, IM SM, and PC conducted the statistical analyses. JP-J drafted the manuscript and all authors critically revised and approved the final version of the manuscript. FLV and FF supervised the cost-effectiveness study. All authors agree to be accountable for all aspects of work ensuring integrity and accuracy.

Conflicts of Interest

The authors declare no conflicts of interest. The BeyondSilos team received 50% funding for Research and Innovation from the European Union for the project. The provider of the telehealth solution (Health Insight Solutions) is a privately-owned company which was not part of the project and was subcontracted under a public tendering process by BSA.

Multimedia Appendix 1
BeyondSilos integrated common care pathways for home care support.
[PDF File (Adobe PDF File), 460 KB - medinform_v8i10e20938_app1.pdf]

Multimedia Appendix 2
Telehealth solution used within the BeyondSilos project (Badalona pilot site).
[PDF File (Adobe PDF File), 472 KB - medinform_v8i10e20938_app2.pdf]

Multimedia Appendix 3
Tables of costs.
[PDF File (Adobe PDF File), 330 KB - medinform_v8i10e20938_app3.pdf]

References

1. Stuck AE, Egger M, Hammer A, Minder CE, Beck JC. Home visits to prevent nursing home admission and functional decline in elderly people: systematic review and meta-regression analysis. JAMA 2002 Feb 27;287(8):1022-1028. [doi: 10.1001/jama.287.8.1022] [Medline: 11866651]
2. Caplan GA, Williams AJ, Daly B, Abraham K. A randomized, controlled trial of comprehensive geriatric assessment and multidisciplinary intervention after discharge of elderly from the emergency department--the DEED II study. J Am Geriatr Soc 2004 Sep;52(9):1417-1423. [doi: 10.1111/j.1532-5415.2004.52401.x] [Medline: 15341540]
3. Dainty KN, Golden BR, Hannam R, Webster F, Browne G, Mittmann N, et al. A realist evaluation of value-based care delivery in home care: The influence of actors, autonomy and accountability. Soc Sci Med 2018 Jun;206:100-109. [doi: 10.1016/j.socscimed.2018.04.006] [Medline: 29727779]
4. Mogensen CB, Ankersen ES, Lindberg MJ, Hansen SL, Solgaard J, Therkildsen P, et al. Admission rates in a general practitioner-based versus a hospital specialist based, hospital-at-home model: ACCESS, an open-labelled randomised clinical trial of effectiveness. Scand J Trauma Resusc Emerg Med 2018 Apr 05;26(1):26 [FREE Full text] [doi: 10.1186/s13049-018-0492-3] [Medline: 29622029]
5. Leichsenring K. Developing integrated health and social care services for older persons in Europe. Int J Integr Care 2004 Sep 03;4(3):e10 [FREE Full text] [doi: 10.5334/ijic.107] [Medline: 16773149]
6. Dickinson H, Allen K, Alcock P, Macmillan R, Glasby J. The role of the third sector in delivering social care. NIHR School for Social Care Research. 2012. URL: http://eprints.lse.ac.uk/43538/ [accessed 2020-05-06]
7. Tsiachristas A, Stein KV, Evers S, Rutten-van Mölken M. Performing Economic Evaluation of Integrated Care: Highway to Hell or Stairway to Heaven? Int J Integr Care 2016 Oct 19;16(4):3 [FREE Full text] [doi: 10.5334/ijic.2472] [Medline: 28316543]
8. Low L, Yap M, Brodaty H. A systematic review of different models of home and community care services for older persons. BMC Health Serv Res 2011 May 09;11(1):93 [FREE Full text] [doi: 10.1186/1472-6963-11-93] [Medline: 21549010]
9. Booker C, Turbutt A, Fox R. Model of care for a changing healthcare system: are there foundational pillars for design? Aust. Health Review 2016;40(2):136-140. [doi: 10.1071/ah14173]
10. Koch S, Hägglund M. Health informatics and the delivery of care to older people. Maturitas 2009 Jul 20;63(3):195-199. [doi: 10.1016/j.maturitas.2009.03.023] [Medline: 19487092]

11. Loh KP, McHugh C, Mohile SG, Mustian K, Flannery M, Klepin H, et al. Using Information Technology in the Assessment and Monitoring of Geriatric Oncology Patients. Curr Oncol Rep 2018 Mar 06;20(3):25 [FREE Full text] [doi: 10.1007/s11912-018-0672-3] [Medline: 29511850]
12. Khosravi P, Ghapanchi AH. Investigating the effectiveness of technologies applied to assist seniors: A systematic literature review. Int J Med Inform 2016 Jan;85(1):17-26. [doi: 10.1016/j.ijmedinf.2015.05.014] [Medline: 26216463]
13. Bertoncello C, Colucci M, Baldovin T, Buja A, Baldo V. How does it work? Factors involved in telemedicine home-interventions effectiveness: A review of reviews. In: A review of reviews. Public Library of Science: Virgili G, editor. Vol. 13, PLoS ONE; 2018:e0207332.
14. Kadu M, Ehrenberg N, Stein V, Tsiachristas A. Methodological Quality of Economic Evaluations in Integrated Care: Evidence from a Systematic Review. Int J Integr Care 2019 Sep 09;19(3):17 [FREE Full text] [doi: 10.5334/ijic.4675] [Medline: 31565040]
15. Vondeling H. Economic evaluation of integrated care: an introduction. Int J Integr Care 2004 Mar 01;4(1):e20 [FREE Full text] [doi: 10.5334/ijic.95] [Medline: 16773144]
16. BeyondSilos project. Learning from integrated eCare practice and promoting deployment in European regions. CORDIS EU Research Results. 2017 Apr 25. URL: https://cordis.europa.eu/project/id/621069 [accessed 2020-05-06]
17. Urošević V, Mitić M. From generic pathways to ICT-supported horizontally integrated care: the SmartCare approach and convergence with future Internet assembly. Stud Health Technol Inform 2014;197:71-75. [Medline: 24743080]
18. Lindberg B, Nilsson C, Zotterman D, Söderberg S, Skär L. Using Information and Communication Technology in Home Care for Communication between Patients, Family Members, and Healthcare Professionals: A Systematic Review. In: Int J Telemed Appl. Using information and communication technology in home care for communication between patients, family members, and health care professionals: a systematic review. Int J Telemed Appl. 2013; 2013:461829.
19. Burgos-Díez C, Sequera-Requero RM, Tarazona-Santabalbina FJ, Contel-Segura JC, Monzó-Planella M, Santaeugènia-González SJ. Study protocol of a quasi-experimental trial to compare two models of home care for older people in the primary setting. BMC Geriatr 2020 Mar 12;20(1):101 [FREE Full text] [doi: 10.1186/s12877-020-1497-0] [Medline: 32164542]
20. Santaeugènia SJ, Mas MA, Tarazona-Santabalbina FJ, García-Lázaro M, Alventosa AM, Gutiérrez-Benito A, et al. Clinical effectiveness of an intermediate care inpatient model based on integrated care pathways. Geriatr Gerontol Int 2020 Apr 18;20(4):366-372. [doi: 10.1111/ggi.13877] [Medline: 32072727]
21. Mas M, Closa C, Gámez S, Inzitari M, Ribera A, Santaeugènia SJ, et al. Home as a Place for Care of the Oldest Stroke Patients: A Pilot from the Catalan Stroke Program. J Am Geriatr Soc 2019 Sep 24;67(9):1979-1981. [doi: 10.1111/jgs.15944] [Medline: 31018014]
22. Baltaxe E, Cano I, Herranz C, Barberan-Garcia A, Hernandez C, Alonso A, et al. Evaluation of integrated care services in Catalonia: population-based and service-based real-life deployment protocols. BMC Health Serv Res 2019 Jun 11;19(1):370 [FREE Full text] [doi: 10.1186/s12913-019-4174-2] [Medline: 31185997]
23. Closa C, Mas M, Santaeugènia SJ, Inzitari M, Ribera A, Gallofré M. Hospital-at-home Integrated Care Program for Older Patients With Orthopedic Processes: An Efficient Alternative to Usual Hospital-Based Care. J Am Med Dir Assoc 2017 Sep 01;18(9):780-784. [doi: 10.1016/j.jamda.2017.04.006] [Medline: 28578883]
24. Mas M, Closa C, Santaeugènia SJ, Inzitari M, Ribera A, Gallofré M. Hospital-at-home integrated care programme for older patients with orthopaedic conditions: Early community reintegration maximising physical function. Maturitas 2016 Jun;88:65-69. [doi: 10.1016/j.maturitas.2016.03.005] [Medline: 27105701]
25. Dueñas-Espín I, Vela E, Pauws S, Bescos C, Cano I, Cleries M, et al. Proposals for enhanced health risk assessment and stratification in an integrated care scenario. BMJ Open 2016 Apr 15;6(4):e010301. [doi: 10.1136/bmjopen-2015-010301] [Medline: 27084274]
26. RossiMori A, Piera-Jiménez J, Albano V, Mercurio G. A systematic analysis of the multi-annual journey of Badalona towards integrated care. Int J Integr Care 2019 Aug 08;19(4):344. [doi: 10.5334/ijic.s3344]
27. Rossi Mori A, Albano V, Piera-Jiménez J. Badalona Story: integrating the integration initiatives. Int J Integr Care 2017 Oct 17;17(5):315. [doi: 10.5334/ijic.3632]
28. Moharra M, Vela E, Dueñas-Espín I, Pauws S, Bescos C, Cano I, et al. Health risk assessment and stratification in an integrated care scenario. Int J Integr Care 2016 Dec 16;16(6):322. [doi: 10.5334/ijic.2870]
29. Mas M, Inzitari M, Sabaté S, Santaeugènia S, Miralles R. Hospital-at-home Integrated Care Programme for the management of disabling health crises in older patients: comparison with bed-based Intermediate Care. Age Ageing Jun 24 2017;46(6):925-931. [doi: 10.1093/ageing/afx099]
30. Mas M, Santaeugènia SJ, Tarazona-Santabalbina FJ, Gámez S, Inzitari M. Effectiveness of a Hospital-at-Home Integrated Care Program as Alternative Resource for Medical Crises Care in Older Adults With Complex Chronic Conditions. J Am Med Dir Assoc 2018 Oct;19(10):860-863. [doi: 10.1016/j.jamda.2018.06.013] [Medline: 30268290]
31. Bernal-Delgado E, Garcia-Armesto S, Oliva J, Sanchez Martinez FI, Repullo JR, Pena-Longobardo LM, et al. Spain: Health System Review. Health Syst Transit 2018 May;20(2):1-179 [FREE Full text] [Medline: 30277216]
32. González-Portillo A, Arroyo G. An Approach to Social Service Systems in Europe: The Spanish Case. An Anal Contemp Soc Welf Issues 2016;47. [doi: 10.5772/65121]

33. Izquieta Etulain JL, Callejo González JJ, Prieto Lobato JM. Third Sector and Public Administrations. Relationships at the Regional and Local Level. Rev Int Sociol 2008 Feb 07;LXVI(49). [doi: 10.3989/ris.2008.i49.85]
34. Charlson ME, Pompei P, Ales KL, MacKenzie C. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. Journal of Chronic Diseases 1987 Jan;40(5):373-383. [doi: 10.1016/0021-9681(87)90171-8]
35. Royal College of General Practitioners. Barthel Activities of Daily Living (ADL) Index. Occas Pap R Coll Gen Pract 1993 Apr(59):24 [FREE Full text] [Medline: 19790834]
36. Kidholm K, Ekeland AG, Jensen LK, Rasmussen J, Pedersen CD, Bowes A, et al. A model for assessment of telemedicine applications: MAST. Int J Technol Assess Health Care 2012 Jan 23;28(1):44-51. [doi: 10.1017/s0266462311000638]
37. Guo HJ, Sapra A. Instrumental Activity of Daily Living (IADL). StatPearls.: StatPearls Publishing; 2020 Jan. URL: https://www.ncbi.nlm.nih.gov/books/NBK553126/ [accessed 2020-09-28]
38. Yesavage J, Sheikh J. Geriatric Depression Scale. Clinical Gerontologist 2008 Oct 25;5(1-2):165-173. [doi: 10.1300/J018v05n01_09]
39. LimeSurvey: An Open Source survey tool. Limesurvey GmbH. URL: http://www.limesurvey.org [accessed 2020-05-06]
40. Hammerschmidt R, Meyer I. Socio-economic impact assessment and business models for integrated eCare. Achieving Effective Integrated E-Care Beyond the Silos 2014:136-163. [doi: 10.4018/978-1-4666-6138-7.ch007]
41. Hair J. Multivariate Data Analysis. New Jersey: Prentice Hall; 2010:785.
42. MAFEIP (Monitoring and Assessment Framework for the EIP on Active and Healthy Ageing). European Commission. URL: https://www.mafeip.eu/ [accessed 2020-05-06]
43. European Innovation Partnership on Active and Healthy Ageing. European Commission. URL: https://ec.europa.eu/eip/ageing/home_en [accessed 2020-05-06]
44. Briggs A, Sculpher M, Claxton K. Decision Modelling for Health Economic Evaluation. Oxford: University of Oxford Press; 2006.
45. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. Methods for the Economic Evaluation of Health Care Programmes. Oxford: Oxford University Press; 2015.
46. Piera-Jiménez J, Winters M, Broers E, Valero-Bover D, Habibovic M, Widdershoven JWMG, et al. Changing the Health Behavior of Patients With Cardiovascular Disease Through an Electronic Health Intervention in Three Different Countries: Cost-Effectiveness Study in the Do Cardiac Health: Advanced New Generation Ecosystem (Do CHANGE) 2 Randomized Controlled Trial. J Med Internet Res 2020 Jul 28;22(7):e17351 [FREE Full text] [doi: 10.2196/17351] [Medline: 32720908]
47. Kaambwa B, Billingham L, Bryan S. Mapping utility scores from the Barthel index. Eur J Health Econ 2013 Apr 2;14(2):231-241. [doi: 10.1007/s10198-011-0364-5] [Medline: 22045272]
48. Neumann PJ, Sanders G, Russell L, Siegel J, Ganiats T. Cost-Effectiveness in Health and Medicine second edition. Oxford: Oxford University Press; 2016.
49. Rovira J, Antoñanzas F. Economic analysis of health technologies and programmes. a Spanish proposal for methodological standardisation. Pharmacoeconomics 1995 Sep;8(3):245-252. [doi: 10.2165/00019053-199508030-00007] [Medline: 10155620]
50. Sculpher M, Fenwick E, Claxton K. Assessing quality in decision analytic cost-effectiveness models. a suggested framework and example of application. Pharmacoeconomics 2000 May;17(5):461-477. [doi: 10.2165/00019053-200017050-00005] [Medline: 10977388]
51. Vallejo-Torres L, García-Lorenzo B, Serrano-Aguilar P. Estimating a cost-effectiveness threshold for the Spanish NHS. Health Econ 2018 Apr 28;27(4):746-761. [doi: 10.1002/hec.3633] [Medline: 29282798]
52. Tsiachristas A, Burgers L, Rutten-van Mölken MPMH. Cost-Effectiveness of Disease Management Programs for Cardiovascular Risk and COPD in The Netherlands. Value Health 2015 Dec;18(8):977-986 [FREE Full text] [doi: 10.1016/j.jval.2015.07.007] [Medline: 26686781]
53. Tsiachristas A, Cramm J, Nieboer AP, Rutten-van Mölken MP. Changes in costs and effects after the implementation of disease management programs in the Netherlands: variability and determinants. Cost Eff Resour Alloc 2014;12(1):17 [FREE Full text] [doi: 10.1186/1478-7547-12-17] [Medline: 25089122]
54. McDonald J, McKinlay E, Keeling S, Levack W. How family carers engage with technical health procedures in the home: a grounded theory study. BMJ Open 2015 Jul 06;5(7):e007761. [doi: 10.1136/bmjopen-2015-007761] [Medline: 26150143]
55. Jennett B. Bringing the Hospital Home: Ethical and Social Implications of High-Tech Home Care. BMJ 1996 Mar 02;312(7030):587-587. [doi: 10.1136/bmj.312.7030.587a]
56. Tsiachristas A. Payment and economic evaluation of integrated care. Int J Integr Care 2015 Apr 22;15(2):e013 [FREE Full text] [doi: 10.5334/ijic.2009] [Medline: 26034470]

Abbreviations

BSA: Badalona Serveis Assistencials
MAFEIP: Monitoring and Assessment Framework for the European Innovation Partnership on Active and Healthy Ageing

Edited by C Lovis; submitted 02.06.20; peer-reviewed by J Roca, H Ide; comments to author 17.07.20; revised version received 12.08.20; accepted 06.09.20; published 06.10.20.

Please cite as:
Piera-Jiménez J, Daugbjerg S, Stafylas P, Meyer I, Müller S, Lewis L, da Col P, Folkvord F, Lupiáñez-Villanueva F
BeyondSilos, a Telehealth-Enhanced Integrated Care Model in the Domiciliary Setting for Older Patients: Observational Prospective Cohort Study for Effectiveness and Cost-Effectiveness Assessments
JMIR Med Inform 2020;8(10):e20938
URL: https://medinform.jmir.org/2020/10/e20938
doi:10.2196/20938
PMID:33021490

Original Paper

Feasibility of Asynchronous and Automated Telemedicine in Otolaryngology: Prospective Cross-Sectional Study

Dongchul Cha[1], MD; Seung Ho Shin[1], MD; Jungghi Kim[1], MD; Tae Seong Eo[1], MD; Gina Na[1], MD; Seonghoon Bae[1], MD; Jinsei Jung[1], MD; Sung Huhn Kim[1], MD; In Seok Moon[1], MD; Jaeyoung Choi[1*], MD; Yu Rang Park[2*], PhD

[1]Department of Otorhinolaryngology, Yonsei University College of Medicine, Seoul, Republic of Korea

[2]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Yu Rang Park, PhD
Department of Biomedical Systems Informatics
Yonsei University College of Medicine
50-1 Yonsei-ro
Seodaemun-gu
Seoul, 03722
Republic of Korea
Phone: 82 2 2228 2363
Email: yurangpark@yuhs.ac

Abstract

Background: COVID-19 often causes respiratory symptoms, making otolaryngology offices one of the most susceptible places for community transmission of the virus. Thus, telemedicine may benefit both patients and physicians.

Objective: This study aims to explore the feasibility of telemedicine for the diagnosis of all otologic disease types.

Methods: A total of 177 patients were prospectively enrolled, and the patient's clinical manifestations with otoendoscopic images were written in the electrical medical records. Asynchronous diagnoses were made for each patient to assess Top-1 and Top-2 accuracy, and we selected 20 cases to conduct a survey among four different otolaryngologists to assess the accuracy, interrater agreement, and diagnostic speed. We also constructed an experimental automated diagnosis system and assessed Top-1 accuracy and diagnostic speed.

Results: Asynchronous diagnosis showed Top-1 and Top-2 accuracies of 77.40% and 86.44%, respectively. In the selected 20 cases, the Top-2 accuracy of the four otolaryngologists was on average 91.25% (SD 7.50%), with an almost perfect agreement between them (Cohen kappa=0.91). The automated diagnostic model system showed 69.50% Top-1 accuracy. Otolaryngologists could diagnose an average of 1.55 (SD 0.48) patients per minute, while the machine learning model was capable of diagnosing on average 667.90 (SD 8.3) patients per minute.

Conclusions: Asynchronous telemedicine in otology is feasible owing to the reasonable Top-2 accuracy when assessed by experienced otolaryngologists. Moreover, enhanced diagnostic speed while sustaining the accuracy shows the possibility of optimizing medical resources to provide expertise in areas short of physicians.

(JMIR Med Inform 2020;8(10):e23680) doi:10.2196/23680

KEYWORDS

telemedicine; otolaryngology; otology; automated diagnosis; asynchronous; COVID-19; diagnosis; feasibility; cross-sectional

Introduction

COVID-19, which was declared a pandemic by the World Health Organization, has shifted societies toward noncontact. Since the disease is highly transmissible between humans and often has respiratory symptoms [1], otolaryngologists are among the most susceptible physicians for infection. Hospital visits raise the risk of hospital-acquired COVID-19 infections, which calls for telemedicine. Currently, telemedicine is widely deployed in the United States and is on the rise [2]. Telemedicine can be synchronous or asynchronous [3]. For example, in otolaryngology, a Veterans Affairs model uses a

community-based outpatient clinic to connect with an otolaryngologist [4]. It is similar to walk-in office clinics in that a clinic visit happens in real time. However, from the physician's point of view, connecting and interviewing the patient in real time through videoconferencing is likely to cause a temporal overhang compared to meeting the patient in person and, therefore, is less efficient. In an asynchronous model, the patient's information and physical findings are presented to the physician as medical records, and therefore, further interview with the patient is not possible. It is more prevalent in consultations between health care providers, and a study by Liddy et al [5] showed it to improve access to specialists. In otology, a study with asynchronous video-otoscopy taken by a tele-health facilitator reported a diagnostic capability equivalent to direct otoscopy by physicians [6].

Diagnosing otologic diseases is mostly noninvasive; initial diagnosis usually involves detailed history taking and physical examination by otoscopy. The problem of telemedicine in the field of otology lies in the presentation of physical inspection. Videoconferencing methods enable physicians to see lesions on the skin; however, looking inside the external ear canal and at the eardrums requires a dedicated imaging device. Conventional otoscopy is not expensive, but sharing raw images is not possible. To share an otoscopic image, one has to rely on expensive otoendoscopic imaging systems. In a recent study, smartphone-enabled otoscopy was shown to be just as effective as conventional otoscopy [7], showing promising possibilities of telemedicine. Nowadays, some consumer-grade tools are more affordable, costing under $40 [8]. With more accessibility to consumer-grade otoendoscopes, the possibility of patients directly contacting physicians for medical advice is increasing, especially during the COVID-19 pandemic era. We, therefore, evaluated the possibility of using telemedicine in otology.

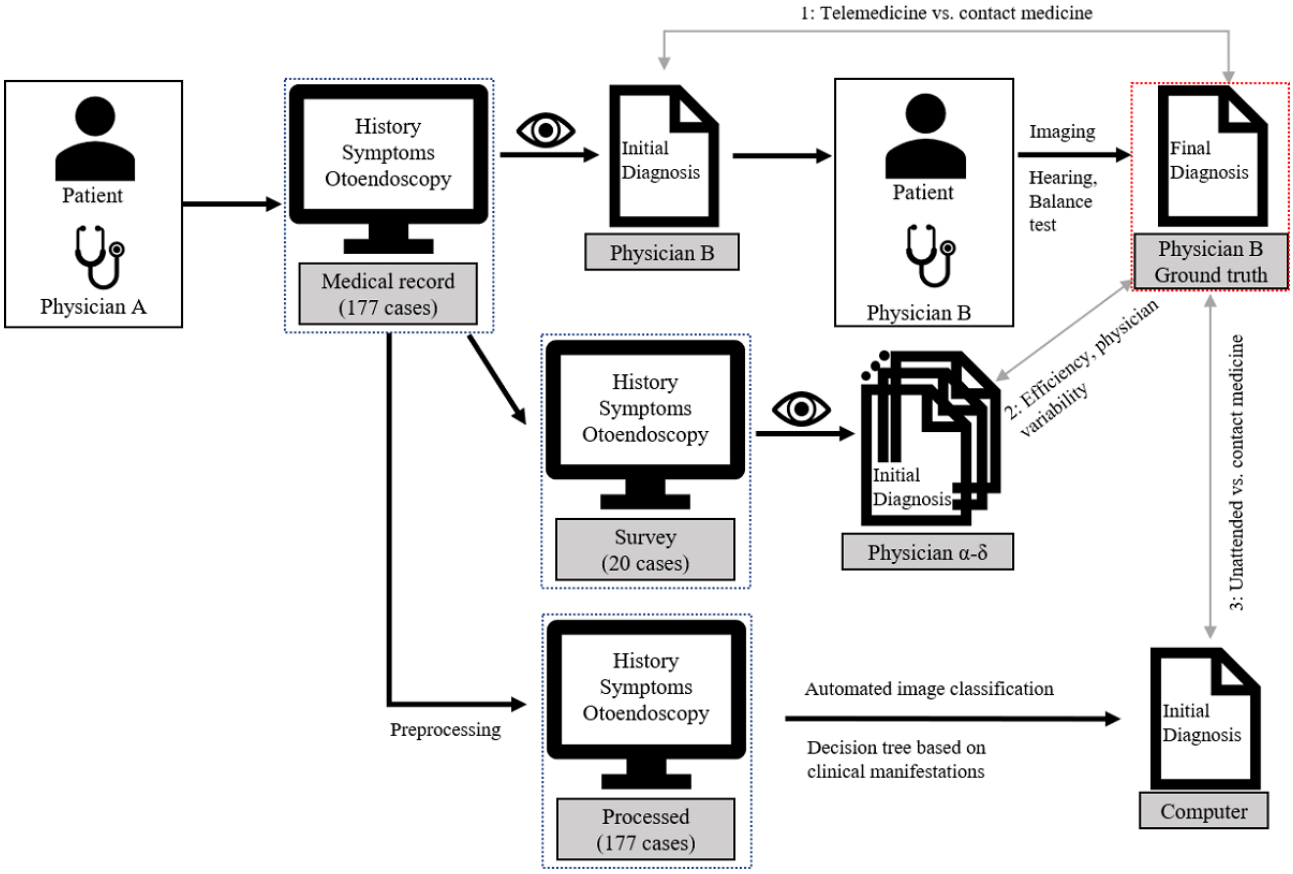
Previous studies focused mainly on findings of otoendoscopy [6,9,10], which did not include other ear diseases such as dizziness, facial palsy, and tumors. In this study, we target all otologic diseases to explore the possibility of converting the operations of the entire otology clinical office from offline to online. Accordingly, we prospectively performed a telemedicine scenario for every new visitor to the otology department clinic. First, we compared the accuracy and likelihood of remote diagnosis to that of office walk-in visits. Second, we conducted a diagnostic survey among otolaryngologists to explore interrater reliability and measure the speed of diagnosis as measures of diagnostic consistency and efficiency. Finally, we created a decision tree to perform an automated diagnostic system, using our previously created otoendoscopic image classification system [11], and compared its accuracy and speed to those of otolaryngologists.

Methods

Patient Selection

All first-time visitors to the Severance Hospital's Otolaryngology Outpatient Clinic from June 8, 2020, to June 19, 2020, who were 18 years or older and having otologic symptoms, were prospectively recruited for the study. A total of 201 patients were eligible for the study. However, 17 patients refused to enroll in the study, and we excluded 2 patients who had Alzheimer disease and one with a known genetic disease (Usher syndrome). There were 4 patients who were workers inside the hospital. They were excluded because of their vulnerability to the rejection of the study. Therefore, 177 patients were included in the final analysis. The Severance Hospital's review board approved this study (Institutional Review Board no 4-2020-0428), and written consent was obtained from all participants. All methods were performed following the Declaration of Helsinki (1964). Figure 1 shows the study design.

Figure 1. Summary of three telemedicine scenarios in this study. Otoendoscopic images were acquired with a consumer-grade device. The eye sign indicates that the records were viewed by corresponding physicians. Physician A is the otolaryngologist or otolaryngology resident, Physician B is the attending physician (otolaryngologist), and Physician α-δ is the otolaryngologist.



Asynchronous Telemedicine Versus Contact Medicine

After obtaining written consent, an otolaryngologist or otolaryngology resident recorded the detailed patient history, noted the clinical manifestations, and acquired an otoendoscopic image of the eardrum and the external auditory canal using a consumer-grade otoendoscopy device tethered to an Android smartphone. In cases of facial palsy or external ear diseases, the default camera app on the Android smartphone was used to acquire facial or lesional photos. The written medical record mainly included chief complaints, duration, present illness, and medico-surgical history. Additional structured profiles, including onset, duration, characteristics, and aggravating and relieving factors, were filled out if the patient had dizziness. All this information and the otoendoscopic image were placed into the electronic medical record. The attending physician then made an initial diagnosis with the information in the electronic medical record before interviewing the patient, thus mimicking an asynchronous telemedicine setting. After the initial impression was set, the patient walked into the office for the attending physician to perform a more detailed history taking; take images using a professional-grade otoendoscope; and perform a physical examination, audiologic tests, and imaging studies if necessary. Based on these, the attending physician made a final diagnosis. Attending physicians were allowed to make no more than initial and final impressions. Of note, our clinic is a tertiary referral center. Still, referral letters and additional study results from previous clinics were not seen by the attending physician when making the initial impression. We should also mention that the patients' reservations were assigned mainly according to the attending physician's subspecialty (acoustic tumors, dizziness, or hearing rehabilitation); therefore, to some extent, the attending physician had an advantage of anticipating the patient's diagnosis.

Efficiency and Interphysician Variability in Telemedicine Diagnosis

An online survey was conducted with four otolaryngologists to evaluate the diagnostic variability between physicians. The survey was conducted in an open type question to reflect real-life clinical settings. Each surveyee was allowed to make up to two diagnoses based on the written patient information, otoendoscopic images acquired with the consumer-grade otoendoscopy device, and photos of facial expressions or lesions acquired with the built-in camera app of the smartphone, if applicable. There were 20 patients that were randomly selected from the 177 patients, and all surveyees evaluated them in the same order. When answering the questionnaire, we timed the total test time to assess the speed of diagnosis. Diagnostic accuracy was calculated as Top-1 and Top-2 accuracy using the Cohen kappa method. Diagnostic speed was measured as the number of diagnoses per minute.

Unattended (Automated) Versus Contact Medicine and Asynchronous Telemedicine Diagnoses

An automated diagnosis decision tree–based model was fed with the patient's symptoms and otoendoscopic images. Otoendoscopic images were classified into normal, tympanic membrane perforation, attic retraction, myringitis, otitis media with effusion, and tumors using an automated otoendoscopic image classification model [11]. Since the automated classification model could not handle facial palsy and preauricular sinuses, these were manually marked as a correct diagnosis, as all surveyees had correctly identified these diseases. The decision tree was based on single-label classification, so only Top-1 was used for comparison of accuracy and likelihood of diagnosis (Cohen kappa). The speed of diagnosis was measured as the total execution time when running under a system environment of Intel Core i7-8750H (Intel Corporation), 16 GB of RAM, and GeForce RTX 2070 (Nvidia Corporation). We converted runtime to a scale of diagnoses per minute.

Classification of the Diseases and Evaluation Metrics

The agreement between telemedicine (initial impression) and contact medicine (final impression) was assessed by Top-1 and Top-2 accuracies. The diagnosis was categorized into 21 categories, based on the International Classification of Diseases, 10th revision diagnostic hierarchy [12]. Accordingly, the next necessary clinical steps were indicated. For example, in cases of hearing impairment, we additionally categorized sudden onset, since the treatment strategy (steroids administration) differs from typical deafness (hearing rehabilitation). Likewise, we divided patients with suspected peripheral vestibulopathy as having benign paroxysmal positional vertigo, vestibulopathy, or Meniere disease. The agreement between the initial and final diagnoses was measured using the Cohen kappa method [13]. The kappa (κ) scores were interpreted as follows: $\kappa<0$, poor; 0.01-0.20, slight; 0.21-0.40, fair; 0.41-0.60, moderate; 0.61-0.80, substantial; and 0.81-1, near perfect agreement [14].

Statistical Analysis

The Cohen kappa method was used to calculate the diagnostic accuracy and the likelihood of agreement in the diagnosis between survey participants, using the Scikit-learn python package [15]. Continuous variables are presented as mean and standard deviation.

Results

Patient Distribution

There were 177 patients in total, ranging in age from 19 to 95 years (69 male and 108 female; mean 55.57, SD 17.05 years). The distribution of diseases in these patients is summarized in Table 1. Hearing impairment was the most common cause of visits, followed by dizziness and chronic otitis media.

Table 1. Diagnoses during contact medicine.

Diagnosis	Count (n=177), n
Hearing loss	41
Chronic otitis media	25
Benign paroxysmal positional vertigo	23
Sudden sensorineural hearing loss	19
Vestibulopathy	16
Tinnitus	14
External ear disease (preauricular fistula, otohematoma)	7
Meniere disease	5
Normal finding	5
Bell palsy	4
Schwannoma (vestibular, facial)	4
Eustachian tube dysfunction	3
Cerumen impaction	3
Middle ear tumors	2
Otitis media with effusion	2
Acute otitis media	1
Postoperative complication (cochlear implant device exposure)	1
Traumatic eardrum perforation	1
Superior canal dehiscence syndrome or perilymph fistula	1

Contact medicine included additional diagnostic modalities (audiologic test, imaging, vestibular function tests) if necessary and was considered as the ground-truth label.

Asynchronous Telemedicine Versus Contact Medicine

We evaluated the accuracy of telemedicine versus contact medicine. The mean Top-1 and Top-2 accuracies were 83.05% and 88.14%, respectively. Of the 177 patients, 23 had a second likely diagnosis, accounting for the difference between mean Top-1 and Top-2 accuracies. We individually calculated each diagnostic class sensitivity in the two most likely diagnoses in the telemedicine setting (Table 2). Diseases with strong clinical correlations, such as sudden sensorineural hearing loss, had 100% sensitivity. In addition, apparent findings such as external ear disease and facial palsy had 100% sensitivity. Diseases related to dizziness had relatively low sensitivity due to the absence of the required physical examination results, such as nystagmus tests, at the time of electronic medical record creation. Vestibular or facial nerve schwannoma is diagnosed by magnetic resonance imaging, not by physical examination; therefore, low sensitivity was inevitable. All predictions and ground-truth labels are presented as a confusion matrix in Multimedia Appendix 1.

Table 2. Telemedicine sensitivities in diagnosing different otological diseases (Top-2).a

Diagnoses	Sensitivity (%)
Hearing loss	100.00
Sudden sensorineural hearing loss	100.00
External ear disease	100.00
Facial palsy	100.00
Chronic otitis media	95.83
Tinnitus	80.00
Meniere disease	80.00
Benign paroxysmal positional vertigo	79.17
Schwannoma	75.00
Vestibular neuritis	73.33
Normal finding	60.00

[a]Incidences of less than 4 were excluded.

Efficiency and Interphysician Variability in Telemedicine Diagnosis

Four otolaryngologists reviewed 20 randomly selected cases in the same order (Table 3). The average Top-1 and Top-2 accuracies were 76.25% (SD 10.31%) and 91.25% (SD 7.50%), respectively. The mean Top-2 accuracy in this assessment was higher than the asynchronous telemedicine's mean Top-2 accuracy for the entire cohort (86.44% vs 91.25%). In the questionnaire, the surveyees were more likely to add a second probable diagnosis, compared to the original telemedicine scenario. Out of the 177 patients, there were 23 (13.0%) cases with a second-likely diagnosis assessed in the original telemedicine scenario. In the questionnaire assessment, there were 8 out of 20 (mean 40.0%, SD 21.21%) second-line impressions on average. Therefore, despite the Top-1 accuracy being lower, the Top-2 accuracy was higher in the survey scenario. Interrater variability was also assessed (Table 4). Substantial agreement was present among the four surveyees (κ=0.71) when only the first-line diagnosis was considered. An almost perfect agreement was observed (κ=0.91) when considering whether the survey participants agreed on one of the first-line or second-line diagnoses. Answering 20 questionnaires took an average of 14 minutes and 2 seconds (SD 20 seconds), making an average of 1.55 (SD 0.48) diagnoses per minute.

Table 3. Accuracies and diagnostic speed in different approaches.

Approach	Top-1 (%)	Top-2 (%)	Diagnoses/min
Asynchronous telemedicine	77.40	86.44	N/A[a]
Survey of 20 cases, average	76.25	91.25	1.55
Otologist 1	75.00	85.00	1.63
Otologist 2	90.00	100.00	1.51
Otologist 3	65.00	85.00	2.11
Otologist 4	75.00	95.00	0.95
Automated system	69.50	N/A	667.90

[a]N/A: not applicable.

Table 4. Interrater reliability among four otolaryngologists based on Cohen kappa in Top-1 and Top-2.

Diagnoses	Cohen kappa
Top-1	
Overall	0.71
External ear disease	1.00
Postoperative complication	1.00
Cerumen impaction	1.00
Superior semicircular canal dehiscence syndrome	1.00
Benign paroxysmal positional vertigo	0.82
Tinnitus	0.82
Otitis media with effusion	0.82
Eustachian tube dysfunction	0.82
Facial palsy	0.82
Chronic otitis media	0.69
Acute otitis media	0.37
Hearing loss	0.19
Meniere disease	0.18
Normal findings	0.18
Top-2	
Overall	0.91
Middle ear tumor	1.00
Benign paroxysmal positional vertigo	1.00
Vestibular neuritis	1.00
Tinnitus	1.00
Meniere disease	1.00
External ear disease	1.00
Postoperative complication	1.00
Cerumen impaction	1.00
Superior semicircular canal dehiscence syndrome	1.00
Facial palsy	0.94
Chronic otitis media	0.92
Hearing loss	0.92
Eustachian tube dysfunction	0.88
Otitis media with effusion	0.82
Schwannoma	0.74
Acute otitis media	0.50
Normal findings	0.18

Looking at the agreement of individual classes, some classes had a higher agreement in Top-2 than Top-1. This is due to high interconnection between different conditions. In these cases, the diagnosis often relies on the physician's experience and personal tendencies. For example, tinnitus is often present with hearing impairment; therefore, it is up to the physician to diagnose the patient with tinnitus or hearing impairment. This flexibility reduced the first-line diagnostic agreement but had a higher agreement when first-line and second-line diagnoses were considered together.

Unattended (Automated) Versus Contact Medicine and Asynchronous Telemedicine Diagnoses

With an experimental classifier that is based on automated otoendoscopic image classification and a decision tree, Top-1 accuracy was 69.5%. The Top-1 results were compared to the

initial impression of the telemedicine scenario to examine how similar the computer systems and the physicians' predictions were. This comparison yielded a κ of 0.70, indicating substantial agreement. We measured the total execution time (loading the patient's information file, classifying the image, and writing the prediction to a file) seven times. It took an average of 15.9 (SD 0.2) seconds to diagnose all 177 patients, which is equivalent to an average of 667.9 (SD 8.3) diagnoses per minute. All predictions and ground-truth labels are presented as a confusion matrix in Multimedia Appendix 1.

Discussion

Principal Results

When diagnosed by four otolaryngologists, the interrater agreement was substantial (κ=0.71) and almost perfect (κ=0.91) in Top-1 and Top-2 diagnoses, respectively. Diagnostic accuracy was stable across different survey participants, especially in Top-2. In real-world clinics, physicians often make more than one differential diagnosis at the initial visit, so it is rational to regard the Top-2 accuracy as a reliable metric. Since by its nature, asynchronous telemedicine does not permit additional interviews with the patients, information may be limited compared to traditional synchronous telemedicine or walk-in clinic visits; under such circumstances, the clinical experience might become a more critical key factor for an accurate diagnosis.

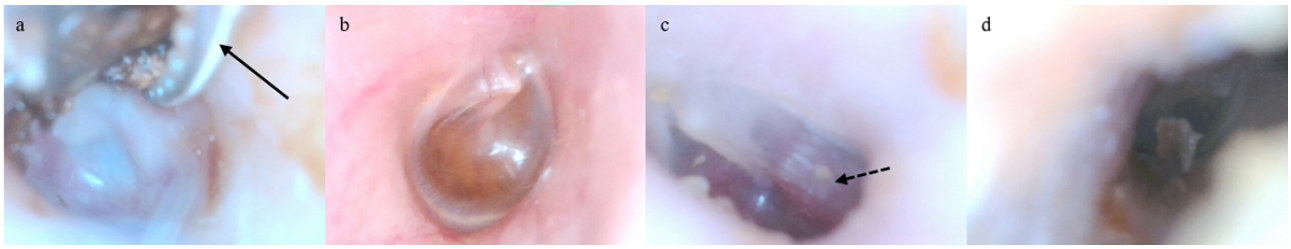
The automated diagnosis system's Top-1 accuracy was 69.50%. Although this is acceptable when compared to other groups' Top-1 accuracies, in real-life clinics, second or third likely differential diagnoses are essential for appropriate treatments. Therefore, we do not think the automated system is ready for clinical use. It needs further refinements.

We additionally tested the time for making a diagnosis as a measure of diagnostic efficiency. Often, it takes more than 10 minutes for a physician to make an initial impression of a new patient. In a study in a primary-care office, the median visit time was 15.7 minutes [16]. In synchronous telemedicine models, there is no advantage in terms of saving time for diagnosis; instead, it may cause temporal overhang in the connection between patients or centers. In this study, physicians were able to make an average of 1.55 diagnoses per minute, with Top-2 accuracy comparable to an attending physician. In the automated diagnosis model, computers diagnosed diseases with lightning speed. With modifications, we may use the system as a computer-aided diagnosis before asking for a second opinion by clinicians.

Comparison With Prior Work and Limitations

In this study, we expand the aim and scope of asynchronous telemedicine in otolaryngology by assessing all otologic diseases, rather than confining it to otoscopic findings alone as in previous studies [6,9,10]. Since we included all patients visiting our tertiary referral center, the disease spectrum was broad, including rare complications such as cochlear implant electrode exposure following surgery. Most of the middle ear diseases could be diagnosed using consumer-grade smartphone otoendoscopy and clinical manifestations, with a high degree of accuracy. Some lesions were easily identified, whereas in some cases, the consumer-grade device showed limitations associated with its resolution, contrast, or size being too large for narrow ear canals (Figure 2). With time, as technology advances, such problems might be solved. Facial palsies and external ear tumors or pits have apparent symptoms and findings, and could be diagnosed with facial photos with almost 100% accuracy. Hearing impairment and tinnitus are symptoms as well as diagnoses; if a patient claims to have it, it is likely there. However, the diagnostic accuracy of dizziness was relatively low because additional physical examinations could not be performed. In this study, since eye cameras are not currently affordable to most consumers, diagnosing dizziness was solely based on clinical representations and history, thus resulting in low diagnostic accuracy. When an eye camera or eye-tracking device becomes widely available, the feasibility of diagnosing dizziness of peripheral origin may be re-evaluated, and the overall accuracy of asynchronous telemedicine is likely to improve.

Figure 2. Capabilities and limitations of a consumer-grade otoendoscope images. (a) A cochlear implant electrode is exposed in the ear canal (arrow), suggesting a postoperative complication. (b) Consumer-grade otoendoscopes offer sufficient image quality to identify otitis media with effusion. (c) It is hard to diagnose middle ear tumors (dashed arrow) due to haziness of the picture. (d) The ear canal is too narrow for a consumer-grade otoendoscope to pass.



This study has some limitations. First, we could not randomly allocate the patients to telemedicine or contact medicine groups since the current legal regulations in South Korea prohibit telemedicine. However, we strictly followed the flow previously mentioned when diagnosing new patients. Second, the same attending physician with the initial impression made the final diagnosis, which may lead to bias in either the initial impression or the final diagnosis. We made sure that attending physicians did not modify the initial impression once they made an impression, and for the final diagnosis, we trusted the attending physicians to make professional clinical decisions, regardless of study enrolment or initial impressions. Last, the automated classification system was trained on images acquired by expensive otoendoscopic imaging towers (OTOLUX 0-degree telescope tethered to Olympus OTV-SP1 video imaging system), not consumer-grade images. Moreover, images of postsurgical

status were not included in the original classification model; we simply used the classifying system using these lower quality otoendoscopic images. Therefore, we suspect that image quality was one of the reasons for low diagnostic accuracy in the automated diagnosis model.

Conclusions

To our knowledge, this is the first prospective study to assess the feasibility and effectiveness of asynchronous telemedicine in otolaryngology. The findings of the study could be used in many ways, especially in the era of the COVID-19 pandemic. In offices, patients may be prediagnosed with appropriate interviews and otoscopic findings, and make all the arrangements for additional tests (hearing, balance, and imaging), if necessary, to minimize waiting time and hospital visits. In areas abundant with otolaryngologists, these may work together to draw a more accurate diagnosis by voting. Where otolaryngologists are scarce, local tele-health facilitators may interview the patients and take otoendoscopic images. These can be sent to a central server for diagnosis. Ideally, patients may also provide symptoms along with otoendoscopic images directly to a server via structured questionnaires or chatbots. Once some amount of patient's records are stacked, otolaryngologists may remotely assess these patients in batches and provide further strategies such as observation, prescription, or recommend an office visit. Since diagnosis time is significantly reduced, it is not likely to impose a heavy burden on the existing medical resources. We claim that this approach might help alleviate the global burden of ear disease by medical resources optimization.

Acknowledgments

This study received support from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (grant number: HI19C1015).

Authors' Contributions

DC, YRP, and JYC designed the study. SHS, JK, and TSE initiated patient interviews and obtained written consent, followed by otoendoscopic pictures, and curated the records into the electronic medical records. GN, SHB, JJ, SHK, ISM, and JYC performed clinical interventions in contact medicine settings. SHS, SHB, JJ, and SHK performed online surveys. GN performed statistical analysis of interrater agreements. DC wrote the machine learning code with YRP and performed the automated medicine scenario. DC and SHS wrote the first draft of the manuscript, and all other authors reviewed, modified, and approved the final manuscript. DC, YRP, and JYC are guarantors of the study. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Confusion matrix (Top-1) of all classes in asynchronous telemedicine (Left) and automatic (Right). Overall accuracy: 83.05% (Left), 69.49% (Right). Contact medicine was considered a ground-truth label. AOM: acute otitis media or myringitis; BPPV: benign paroxysmal positional vertigo; Cer: cerumen impaction; COE: chronic otitis externa; COM: chronic otitis media; ETD: eustachian tube dysfunction; Ext: external ear disease (includes preauricular fistula and otohematoma); FP: facial palsy (Bell palsy); HL: hearing loss (includes conductive, sensorineural, and mixed); MD: Meniere disease; MET: middle ear tumor; Nor: normal finding; OME: otitis media with effusion; POC: postoperative complication; SCDS: superior semicircular canal dehiscence syndrome; Sch: schwannoma (vestibular and facial); S-SNHL: sudden sensorineural hearing loss; Tin: tinnitus; Tr: traumatic eardrum perforation; VN: vestibular neuritis, vestibulopathy.

[PNG File , 389 KB - medinform_v8i10e23680_app1.png]

References

1. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. N Engl J Med 2020 Mar 26;382(13):1199-1207. [doi: 10.1056/nejmoa2001316]
2. Dorsey ER, Topol EJ. Telemedicine 2020 and the next decade. Lancet 2020 Mar;395(10227):859. [doi: 10.1016/s0140-6736(20)30424-4]
3. Allely EB. Synchronous and asynchronous telemedicine. J Med Syst 1995 Jun;19(3):207-212. [doi: 10.1007/bf02257174]
4. McCool RR, Davies L. Where does telemedicine fit into otolaryngology? An assessment of telemedicine eligibility among otolaryngology diagnoses. Otolaryngol Head Neck Surg 2018 Apr;158(4):641-644. [doi: 10.1177/0194599818757724] [Medline: 29436270]
5. Liddy C, Moroz I, Mihan A, Nawar N, Keely E. A systematic review of asynchronous, provider-to-provider, electronic consultation services to improve access to specialty care available worldwide. Telemed J E Health 2019 Mar;25(3):184-198. [doi: 10.1089/tmj.2018.0005] [Medline: 29927711]

6. Biagio L, Swanepoel DW, Adeyemo A, Hall JW, Vinck B. Asynchronous video-otoscopy with a telehealth facilitator. Telemed J E Health 2013 Apr;19(4):252-258. [doi: 10.1089/tmj.2012.0161] [Medline: 23384332]
7. Moshtaghi O, Sahyouni R, Haidar YM, Huang M, Moshtaghi A, Ghavami Y, et al. Smartphone-enabled otoscopy in neurotology/otology. Otolaryngol Head Neck Surg 2017 Mar;156(3):554-558. [doi: 10.1177/0194599816687740] [Medline: 28118550]
8. Meng X, Dai Z, Hang C, Wang Y. Smartphone-enabled wireless otoscope-assisted online telemedicine during the COVID-19 outbreak. Am J Otolaryngol 2020;41(3):102476 [FREE Full text] [doi: 10.1016/j.amjoto.2020.102476] [Medline: 32305252]
9. Lundberg T, Biagio de Jager L, Swanepoel DW, Laurent C. Diagnostic accuracy of a general practitioner with video-otoscopy collected by a health care facilitator compared to traditional otoscopy. Int J Pediatr Otorhinolaryngol 2017 Aug;99:49-53. [doi: 10.1016/j.ijporl.2017.04.045] [Medline: 28688565]
10. Cottrell E, George A, Coulson C, Chambers R. Telescopic otology referrals: evaluation of feasibility and acceptability. Laryngoscope Investig Otolaryngol 2020 Apr;5(2):221-227 [FREE Full text] [doi: 10.1002/lio2.367] [Medline: 32337353]
11. Cha D, Pae C, Seong S, Choi J, Park HJ. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. EBioMedicine 2019 Jul;45:606-614 [FREE Full text] [doi: 10.1016/j.ebiom.2019.06.050] [Medline: 31272902]
12. International Statistical Classification of Diseases and Related Health Problems. Geneva, Switzerland: World Health Organization; 2004.
13. Fleiss JL, Cohen J. The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychological Meas 2016 Jul 02;33(3):613-619. [doi: 10.1177/001316447303300309]
14. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005 May;37(5):360-363 [FREE Full text] [Medline: 15883903]
15. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Machine Learning Res 2011 Oct;12:2825-2830.
16. Tai-Seale M, McGuire TG, Zhang W. Time allocation in primary care office visits. Health Serv Res 2007 Oct;42(5):1871-1894 [FREE Full text] [doi: 10.1111/j.1475-6773.2006.00689.x] [Medline: 17850524]

Edited by G Eysenbach; submitted 19.08.20; peer-reviewed by J Knitza; comments to author 10.09.20; revised version received 12.09.20; accepted 22.09.20; published 19.10.20.

Please cite as:
Cha D, Shin SH, Kim J, Eo TS, Na G, Bae S, Jung J, Kim SH, Moon IS, Choi J, Park YR
Feasibility of Asynchronous and Automated Telemedicine in Otolaryngology: Prospective Cross-Sectional Study
JMIR Med Inform 2020;8(10):e23680
URL: http://medinform.jmir.org/2020/10/e23680/
doi:10.2196/23680
PMID:33027033

Original Paper

A Novel Approach to Assessing Differentiation Degree and Lymph Node Metastasis of Extrahepatic Cholangiocarcinoma: Prediction Using a Radiomics-Based Particle Swarm Optimization and Support Vector Machine Model

Xiaopeng Yao[1,2], PhD; Xinqiao Huang[3], MD; Chunmei Yang[3], MD; Anbin Hu[1,2], PhD; Guangjin Zhou[4], BA; Mei Ju[5], MA; Jianbo Lei[1,6*], PhD; Jian Shu[3*], PhD

[1]School of Medical Information and Engineering, Southwest Medical University, Luzhou, China
[2]Central Nervous System Drug Key Laboratory of Sichuan Province, Southwest Medical University, Luzhou, China
[3]Department of Radiology, The Affiliated Hospital of Southwest Medical University, Luzhou, China
[4]Department of Radiology, Peking University Third Hospital, Beijing, China
[5]School of Nursing, Southwest Medical University, Luzhou, Sichuan Province, China
[6]Center for Medical Informatics/Institute of Medical Technology, Peking University, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Jian Shu, PhD
Department of Radiology
The Affiliated Hospital of Southwest Medical University
25 Taiping Street
Luzhou, 646000
China
Phone: 86 18980253083
Email: shujiannc@163.com

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2021/1/e25337

Abstract

Background: Radiomics can improve the accuracy of traditional image diagnosis to evaluate extrahepatic cholangiocarcinoma (ECC); however, this is limited by variations across radiologists, subjective evaluation, and restricted data. A radiomics-based particle swarm optimization and support vector machine (PSO-SVM) model may provide a more accurate auxiliary diagnosis for assessing differentiation degree (DD) and lymph node metastasis (LNM) of ECC.

Objective: The objective of our study is to develop a PSO-SVM radiomics model for predicting DD and LNM of ECC.

Methods: For this retrospective study, the magnetic resonance imaging (MRI) data of 110 patients with ECC who were diagnosed from January 2011 to October 2019 were used to construct a radiomics prediction model. Radiomics features were extracted from T1-precontrast weighted imaging (T1WI), T2-weighted imaging (T2WI), and diffusion-weighted imaging (DWI) using MaZda software (version 4.6; Institute of Electronics, Technical University of Lodz). We performed dimension reduction to obtain 30 optimal features of each sequence, respectively. A PSO-SVM radiomics model was developed to predict DD and LNM of ECC by incorporating radiomics features and apparent diffusion coefficient (ADC) values. We randomly divided the 110 cases into a training group (88/110, 80%) and a testing group (22/110, 20%). The performance of the model was evaluated by analyzing the area under the receiver operating characteristic curve (AUC).

Results: A radiomics model based on PSO-SVM was developed by using 110 patients with ECC. This model produced average AUCs of 0.8905 and 0.8461, respectively, for DD in the training and testing groups of patients with ECC. The average AUCs of the LNM in the training and testing groups of patients with ECC were 0.9036 and 0.8889, respectively. For the 110 patients, this model has high predictive performance. The average accuracy values of the training group and testing group for DD of ECC were

82.6% and 80.9%, respectively; the average accuracy values of the training group and testing group for LNM of ECC were 83.6% and 81.2%, respectively.

Conclusions: The MRI-based PSO-SVM radiomics model might be useful for auxiliary clinical diagnosis and decision-making, which has a good potential for clinical application for DD and LNM of ECC.

(JMIR Med Inform 2020;8(10):e23578) doi:10.2196/23578

KEYWORDS

PSO-SVM algorithm; magnetic resonance imaging; lymph node metastases; differentiation degree; extrahepatic cholangiocarcinoma; radiomics feature; algorithm; MRI; radiomics; lymph; cancer; oncology

Introduction

Cholangiocarcinoma is a highly aggressive neoplasm with a poor prognosis. Cholangiocarcinomas are commonly classified as either extrahepatic cholangiocarcinoma (ECC) or intrahepatic cholangiocarcinoma (ICC), on the basis of their anatomic position in regard to the second-order bile ducts. Generally, ECCs account for approximately 80-90% of diagnosed cases of cholangiocarcinoma [1]. Most (60%-70%) of ECCs are perihilar or “Klatskin” tumors, including the hepatic duct bifurcation; the rest of ECCs incorporate in the distal common bile duct [1].

Radical surgical resection is still the uniquely definitive and effective therapy for the long-term survival of patients with ECC. Patients with ECC show a low survival rate, attributed to hidden early clinical symptoms and a lack of effective nonsurgical therapeutic agents, which lead to local lymph vascular invasion and lymph node metastases (LNMs) [2]. In general, surgical resection with a cure expectation is associated with an 18%-54% 5-year survival rate for ECC [3-5]. Among clinicopathological features, tumor differentiation, positive lymph node, and lymphatic invasion were considered independent predictors of the overall survival rate of ECC [6-8]. Therefore, the accurate preoperative assessment of tumor pathological differentiation degree and lymph node status (especially lymph node status) could provide considerable help for the planning of treatment as soon as possible.

Ultrasonography, computerized tomography (CT), 18-fluorodeoxyglucose positron emission tomography/computerized tomography (18F–FDG-PET/CT), magnetic resonance imaging (MRI) and magnetic resonance cholangiopancreatography (MRCP), direct cholangiography, and endoscopy are traditional imaging methods for observing and diagnosing ECC [3,9]. MRI is regarded as a noninvasive and precise imaging modality for patients with ECC. MRI can provide information about lymph node metastases and survival results [10]. However, we should recognize some of the inherent defects of MRI. Traditional techniques mainly depend on radiologists’ subjective visual and qualitative observations. Therefore, we still have no quantitative way of predicting pathological differentiation degree (DD) and LNM of ECC, including MRI [11]. More importantly, it’s quite difficult to analyze the tremendous digital characteristics of the cells, physiology, and genetic variation of patients in the images, which cannot be distinguished by human eyes [12]. In current clinical studies, preoperative morphological features of lymph nodes, such as size, number, ratio, morphology, signal intensity, and lymph node changes, can be used to evaluate the preoperative lymph node status of ECC [13-15]. However, the accurate prediction method for assessing DD and LNM of ECC is incomprehensive.

By extracting traditional MRI, a large number of radiologic features can be obtained. Radiomics can be intuitively regarded as an approach that can quantify the conversion of visual image information into deep features [16,17]. This radiomics model is based on a machine-learning approach that can help doctors make the most accurate diagnosis by mining and analyzing radiological features. So far, radiomics have been successfully used to assist in decision making on the diagnosis and risk stratification of several types of cancer, such as hepatocellular carcinoma [18], glioma [19], rectal cancer [20], lung cancer [21], breast cancers [22], and thymic epithelial tumors [23]. Nonetheless, the diagnostic significance of radiomics in patients with ECC has not be evaluated.

In this paper, a radiomics model based on particle swarm optimization and a support vector machine (PSO-SVM) was developed for predicting DD and LNM of patients with ECC.

Methods

Patient Selection

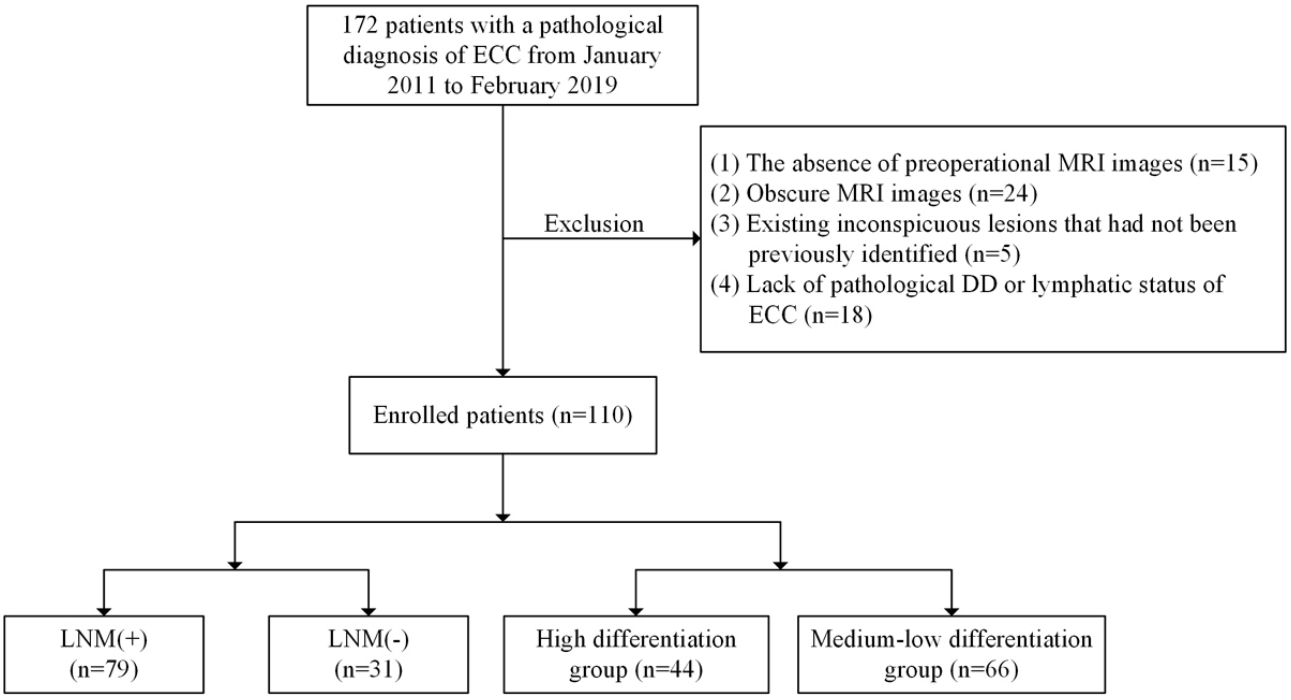
We retrospectively collected a total of 110 consecutive patients’ data (which included 60 men and 50 women) with ECC who underwent radical surgical resection between January 2011 and October 2019 at our hospital (The Affiliated Hospital of Southwest Medical University). Every inpatient underwent an abdominal MRI examination within 2 weeks before surgical resection, chemotherapy, or radiotherapy. With approval from the local Institutional Review Board and Ethics Committee, all features for patients with ECC were retrospectively investigated. Retrieved data included clinical symptoms, laboratory examination, surgery notes, MRI features, and pathological outcomes (including pathological DD and lymph node status). All identifying information in the records was deleted to protect patients’ privacy.

The inclusion criteria were as follows: (1) All patients had pathologically confirmed ECC; (b) the regional LNMs dissection was performed during the operation; (3) abdominal MRI scans were obtained within 2 weeks before surgical resection, chemotherapy, or radiotherapy; and (4) the clinical and follow-up data were available. The final diagnosis of ECC was based on a combination of pathological examination results and MRI examination. Exclusion criteria were as follows: (1) the

absence of preoperational MRI images; (2) obscure MRI images; (3) the presence of unidentified, inconspicuous lesions; (d) a lack of pathological DD or lymphatic status of ECC.

Of the initial 172 patients with a pathological diagnosis of ECC from January 2011 to February 2019, we excluded 62 patients because of insufficient medical examination information, such as the absence of preoperational MRI images (n=15), obscure MRI images (n=24), the existence of unidentified, inconspicuous lesions (n=5), and a lack of pathological DD or lymphatic status of ECC (n=18). Consequently, 110 patients were used for DD and LNM of ECC. A flow diagram summarizing the study selection and inclusion is reported in Figure 1. The DD of ECC was divided into a high-risk differentiation group (n=44) and a low-medium risk differentiation group (n=66). The LNM of ECC was divided into a positive lymph node metastases group [LNM (+); n=79] and a negative lymph node metastases group [LNM (-); n=31].

Figure 1. Flow diagram of patient cohort selection (n=110). DD: differentiation degree; ECC: extrahepatic cholangiocarcinoma; LNM: lymph node metastases; MRI: magnetic resonance imaging.



Histopathologic Analysis of the Study Population

All study patients underwent surgical resection, lesions were made into paraffin-embedded specimens, and the patients were histologically diagnosed with ECC. The samples were colored with a hematoxylin-eosin stain for regular histopathologic assessment. All specimens were identified by a seasoned histopathologist, who had over five years of work experience and was trained not to disclose individual participants' relevant information.

According to the American Joint Committee on Cancer (AJCC) and the College of American Pathologists, the ECC can be divided into 3 pathological grades: high-differentiation (G1), medium-differentiation (G2), and low-differentiation (G3) [24]. For G1, more than 95% of the tumor is composed of glands, and the perniciousness of the degree of the tumor is relatively low; for G2, 50–95% of the tumor is composed of glands, and the degree of the tumor is moderately malignant; for G3, less than 50% of the tumor is composed of glands, and the perniciousness of the degree of the tumor is relatively large. This pathological differentiation has a certain significance for the clinical treatment and prognosis of ECC. Generally, G1 has a better prognosis and less metastasis than G2 and G3. G3 has a worse prognosis and more metastasis than G2.

MRI Acquisition Protocol

A Philips Achieva 3.0T superconducting MRI scanner with a quasar dual gradient system and a 16-channel phased-array torso coil was used to create all magnetic resonance images. Patients were asked to fast for 4-8 hours before the examination, with no restriction on drinking water. They also practiced breathing and holding their breath in the supine position. The imaging protocol mainly described the data acquisition and MRI sequences analysis. The MRI sequences were the following: an axial T1-weighted high-resolution isotropic volume excitation sequence (T1WI), an axial fat-suppressed turbo spin-echo (TSE) T2-weighted spectral attenuated inversion recovery (T2WI), a coronal TSE T2WI sequence, an axial dual-echo T1WI breath-hold gradient-echo sequence for the acquisition of in-phase and out-of-phase images, axial diffusion-weighted imaging (DWI), and T1-weighted dynamic contrast-enhanced MR images (including arterial, portal venous, transitional, and delayed phase). In this study, we mainly selected T1WI, T2WI, DWI, and ADC as the image data. The parameters of MRI sequences (T1WI, T2WI, DWI, ADC) are shown in Table 1.

Table 1. The acquisition parameters of the abdominal magnetic resonance imaging (MRI) protocol.

Acquisition parameters	Imaging protocol			
	T1WI[a]	T2WI[b]	DWI[c]	ADC[d]
Repetition time (milliseconds)	300	2610	2103	N/A[e]
Echo time (milliseconds)	14	70	70	N/A
Flip angle (degrees)	10	90	90	N/A
Field of view (mm×mm)	365×305	280×305	375×305	N/A
Matrix size (mm×mm)	204×154	176×20 1	128×256	N/A
Slice thickness (mm)/gap(mm)	7/1	7/1	7/1	N/A
Slices (mm)	24	24	72	24
Averaged number of signals	1	2	4	N/A
b values (s/mm^2)	N/A	N/A	0 and 800	800

[a]T1WI: T1-weighted imaging high spatial resolution isotropic volume exam.

[b]T2WI: fat-suppressed turbo spin-echo T2-weighted imaging spectral attenuated inversion recovery.

[c]DWI: diffusion-weighted imaging.

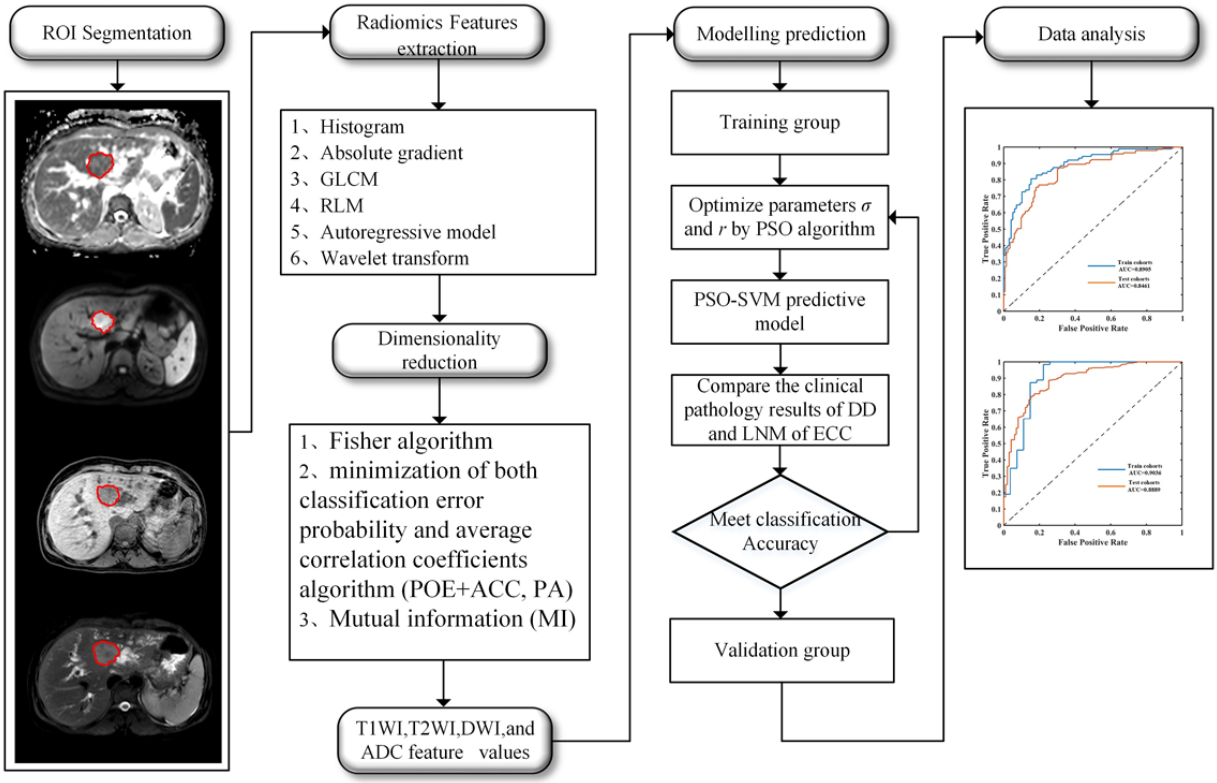
[d]ADC: apparent diffusion coefficient.

[e]N/A: not available.

Workflow

The workflow of this paper is shown in Figure 2. It includes five main parts: (1) imaging and region of interest (ROI) segmentation, (2) radiomics features extraction, (3) dimension reduction, (4) PSO-SVM model construct, and (5) data analysis. These 5 parts will be detailed in the following section.

Figure 2. Research workflow of the paper. ADC: apparent diffusion coefficient; DD: differentiation degree; DWI: diffusion-weighted imaging; ECC: extrahepatic cholangiocarcinoma; GLCM: grey-level co-occurrence matrix; LMN: lymph node metastases; PSO-SVM: particle swarm optimization and support vector machine; RLM: grey-level run-length matrix; ROI: receiver operating characteristic curve; T1WI: T1-weighted imaging high spatial resolution isotropic volume exam; T2WI: fat-suppressed turbo spin-echo T2-weighted imaging spectral attenuated inversion recovery.



ROI Segmentation

All patients were followed up, and whether the lesion had recurred or metastasized was determined by radiological and pathological diagnosis. The relevant MRI images of patients were collected in the PACS-DICOM (picture archiving and communication system–Digital Imaging and Communications in Medicine) system, where the sequences of ECC were clearly selected. Given the 1515×1114-pixel image of cholangiocarcinoma, the average area of the lesions was 125.522 mm^2. We did not exclude any images, and the radiology feature extraction used the entire ROI image.

Radiomics Feature Extraction

The MRI radiomics features of ECC were extracted using MaZda software (version 4.6; Institute of Electronics, Technical University of Lodz). The MRI analysis started with the definition of the ROIs. Under the guidance of an experienced radiologist, the ROI of the lesion was outlined to avoid adjacent vessels and bile ducts, and to locate the inside of the parenchyma of the tissue as much as possible. To outline lesions in MRI images, it is necessary to maintain about 1-2 mm from the edge of the tumor and to minimize the average volume of the surrounding structures when extracting image features. In the feature extraction process, the image intensity within the range of μ (SD 3) was normalized to minimize the influence of contrast and brightness variation. We finally extracted 300 radiomics features from the ROI of each sequence based on the following algorithms: first-order histogram, grey-level co-occurrence matrix (GLCM), grey-level run-length matrix (RLM), autoregressive model, and wavelet transform.

Data Dimensionality Reduction

All ROI features are high-dimensional data, and it may be difficult to select the required features if data dimensionality reduction (DDR) is not performed before the feature data is inputted into the classifier.

The purpose of DDR was to reduce the number of attributes under consideration so as to obtain the optimal features from the original features. Therefore, before we performed image classification and recognition, the significant features were selected to reduce the bias in features modeling. Based on MaZda software, we provided 3 methods for performing DDR and obtaining the optimal features: (1) the Fisher algorithm (F), (2) minimization of both classification error probability and the average correlation coefficients algorithm (POE+ACC, PA), and (3) mutual information (MI). These methods were used to deal with each feature separately and to remove almost indistinguishable features. Finally, 30 optimal features were selected from 300 radiomics features of each sequence (T1WI, T2WI, DWI, and ADC, respectively).

PSO-SVM Model Construction

After implementing DDR of the ROI features, the optimal features were adopted to build the prediction model. In the modeling process, all feature data had been normalized in the interval (0,1) to eliminate the dimensional difference of radiomics features. The min-max normalization algorithm was used to normalize the radiomics features value cohort. In order to calculate uniformly, the main purpose was to convert the different magnitudes data into the same magnitude order. The min-max normalization algorithm can be described from the following equation:

$$X=(x-x_{min})/(x_{max}-x_{min}) \quad (1)$$

X is the normalized value of the optimal features, x is the value of the optimal features, x_{max} is the maximum value of the optimal features, and x_{min} is the minimum value of the optimal features.

Because cholangiocarcinoma is a rare disease and the number of cases is relatively small, we faced a typical prediction modeling problem of small sample sizes. The basic principle of the PSO-SVM algorithm is to construct a hyperplane and distinguish high-dimensional mappings of feature data classification. The space of the feature data was taken as an input variable, and then the penalty parameters (c and g) of the support vector machine (SVM) were optimized by using the PSO algorithm. Then, the SVM algorithm was used to construct the prediction model for DD and LNM of ECC. To improve the performance of the prediction model, cross-validation and iterative training was used to verify data in this study.

Data Analysis

Development, Performance, and Validation of a Radiomics Model

In this paper, a radiomics model based on the PSO-SVM algorithm was established to predict DD and LNM of ECC by combining the optimal features of the tumor ROI and clinical outcomes. All patients were divided into high-risk and low-medium risk differentiated groups according to the pathological examination results. The min-max algorithm was used to normalize 120 features, including 90 radiomics features from 3 sequences (T1WI, T2WI, and DWI) and 30 ADC values of the tumors, which can eliminate the negative effects caused by different sample dimensions. The distribution of DD and LNM cases of ECC was imbalanced. Statistically, there were mainly 2 methods to solve the problem: one was the under-sampling algorithm, and the other was the synthetic minority oversampling algorithm (SMOTE) [25]. The under-sampling algorithm could mainly achieve the sample balance by reducing the data set. This method was suitable for statistical problems with sufficient samples. Because there were fewer cases of ECC in this study, the under-sampling algorithm is not suitable for statistical problems with fewer samples. On the contrary, the oversampling algorithm was artificial to synthesize minority samples and add new samples to achieve sample balance. For the DD of ECC, the number of low-medium risk differentiated groups (n=68) was significantly larger than that of high-risk differentiated groups (n=42) for the DD of ECC, and the cases were extremely class-imbalanced. The number of low-medium-risk and high-risk differentiation groups were adjusted to be the same (n=1428) by using the SMOTE algorithm, respectively. For the LNM of ECC, the number of metastasis cases (n=33) was significantly less than nonmetastasis cases (n=77). Similarly, the numbers of metastasis and nonmetastatic groups were adjusted to be the same (n=231) by using the SMOTE algorithm, respectively. In this way, the number of ECC cases was balanced.

During the modeling process, we randomly selected 88 cases as the training group and the remaining 22 as the test group for DD and LNM of ECC. The PSO algorithm was used to obtain the optimal penalty parameters c of 7.3607 and g of 0.2132 so as to improve the classification accuracy and the robustness of this prediction model.

We determined the receiver operating characteristic curve (ROC) and the area under the curve (AUC) to evaluate the predictive performance of the PSO-SVM radiomics model. Furthermore, the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy of the proposed model were calculated. Then, this model was evaluated by all of the above indicators for the validation cohort.

Statistics, Comparison, and Analysis

All continuous data (age and lesion area) were respectively given as means and medians (with interquartile ranges). ROC analysis was adopted to test the PSO-SVM model. We used the MATLAB statistics package (version 9.1; MathWorks) to conduct statistical analysis. We compared the result from the same case with independent *t* tests and Wilcoxon rank sum tests, whereas the categorical variables, including gender and tumor location, were compared using a chi-square test. The evaluation indicators of the proposed model were also designed by MATLAB, which included AUC, classification accuracy, PPV, NPV, sensitivity, and specificity. A 2-tailed *P* value of less than .05 was considered statistically significant.

Results

Clinical Features of the Studied Patients

A total of 110 patients were selected from The Affiliated Hospital of Southwest Medical University. The mean age of patients was 57.0 (SD 10.0, range 28-83) years and the group included 60 (54.5%) men and 50 (45.5%) women. The clinical and baseline characteristics are summarized in Table 2. According to the pathological results of ECC, all patient cases were divided into high-risk differentiation groups (n=42) and low-medium risk differentiation groups (n=68). Simultaneously, there were no significant heterogeneity differences between the 2 groups of data features for DD of ECC.

According to the pathological examination report, of the 110 patients, a total of 33 cases (30%) were diagnosed with lymph node metastasis, and the other 77 cases (70%) were diagnosed as being without lymph node metastasis. By analyzing the 5 characteristics in Table 2, there were no significant heterogeneity differences between the 2 groups of data features for non-LNM and LNM of ECC.

Table 2. Clinical and pathological characteristics of patients with extrahepatic cholangiocarcinoma (ECC; n=110).

Characteristics	Differentiation degree of ECC			LNM[a] of ECC		
	High-risk group	Low-medium risk group	*P* value	Non-LNM	LNM	*P* value
Age in years, mean (SD)	56.4 (10.3)	57.5 (9.8)	.957	58.0 (9.6)	54.4 (10.6)	.272
Gender, n (%)			.434			.969
Male	22(50)	38(57.6)		43(54.4)	17(54.8)	
Female	22(50)	28(42.4)		36(45.6)	14(45.2)	
Lesion location, n (%)			.876			.174
Porta	20(45.5)	29(43.9)		32(40.5)	17(54.8)	
Distal bile duct	24(54.5)	37(56.1)		47(59.5)	14(45.2)	
Lesion area[b] (mm^2), mean (SD)	115.144 (SD 78.425)	131.8649 (SD 73.069)	.495	133.199 (SD 86.93)	103.515 (SD 70.998)	.816

[a]Lymph node metastases.

[b]Lesion size was defined as the maximum diameter on transverse images.

Reliability of Radiomics Feature Selection

In order to construct a high-performance prediction model of PSO-SVM, we needed to obtain reliable ROI features. First, we randomly selected feature data of 30 patients from the 3 MRI sequences of T1WI, T2WI, and DWI, which had outlined ROI segmentation and extracted radiomics features. To evaluate the repeatability between intra-observer and inter-observer, we provided 2 radiologists (JS and XH), each of whom have over 5 years of experience in abdominal oncologic imaging diagnosis. They performed ROI segmentation and feature extraction of the MRI images in a blinded fashion.

To ensure the objectivity of radiomics features, the 2 radiologists were aware of the diagnosis of ECC but were blinded to the clinical and pathologic details. The first radiologist repeatedly followed the same procedure to outline and determine the ROI twice within a week, and then we compared the 2 groups of radiomics features to evaluate intra-observer reliability. The second radiologist also independently outlined the ROI area and extracted radiomics features according to the same operating procedure. Then we evaluated inter-observer reliability by comparing the extracted results of the ROI area between the first radiologist and the second radiologist. The intraclass correlation coefficient (ICC) was used to evaluate the repeatability of radiomics features extracted by intra-observer and inter-observer. ICC can be obtained by using SPSS software according to the following equation:

Cov (X,Y) is covariance; σ_X is X standard deviation; σ_Y is Y standard deviation.

The radiomics features with ICC values of both the intra-observer and inter-observer greater than 0.75 (indicating satisfactory repeatability) were selected for subsequent modeling research. According to the above requirement, since all 300 radiation features extracted from each sequence have satisfactory consistency, no abnormal feature data were found and eliminated. The average value of the ICC within the inter-observer reached 0.97 (range 0.812-1, *P*<.001), and the average ICC among the intra-observers reached 0.98 (range 0.826-1, *P*<.001), as shown in Table 3. According to the above calculation results, because the radiology features extracted in each sequence (T1WI, T2WI, DWI, ADC) have satisfactory consistency, no abnormal feature data was found and eliminated. Therefore, no abnormal characteristic data was found and eliminated.

Table 3. The intraclass correlation coefficient (ICC) between the intra-observer and inter-observer.

Data	Intra-observer	Inter-observer
Patients, n	30	30
MRI sequence	T1WI, T2WI, DWI	T1WI, T2WI, DWI
ICC		
Mean	0.9849	0.9749
Maximum	0.9999	1
Minimum	0.8256	0.8641
SD	0.0278	0.0333

PSO-SVM Model Construction

We selected 90 optimal features from 3 sequences (T1WI, T2WI, and DWI) and 30 ADC values by reducing dimensionality as the sample set. All of the data was normalized to be used for modeling. We randomly selected the optimal features of 88 patients as the training cohorts and the remaining optimal features of 22 patients as the test cohorts. The training cohorts were used to optimize the penalty parameters (c and g) of the SVM by using the PSO algorithm. To further improve the performance of the SVM classifier, the test cohorts were used to verify the performance and accuracy of the SVM classifier. Consequently, we built a radiomics prediction model based on PSO-SVM using the MRI images for predicting DD and LNM of ECC.

Overall Validation of the PSO-SVM Radiomics Model

In order to verify the robustness and deliverability of the PSO-SVM radiomics prediction model, we mainly evaluated the classification accuracy through the ROC curve. The ROC curve is a basic tool used for diagnostic test evaluation, which could reflect the performance of the PSO-SVM radiomics prediction model; it should ensure that the classification rates of the high-risk and low–medium-risk differentiated cases are as high as possible. However, the prediction model would make sure that a lot of the true positive cases are detected, even at the cost of some false positives during the screening phase.

Based on the PSO-SVM radiomics model, the performance of this model for predicting DD and LNM of ECC is shown in Figure 3, and the detailed data is listed in Table 4. The average accuracy of the training group and the testing group for DD of ECC were 82.6% and 80.9%, respectively; the average sensitivity was 80.5% and 78.1%, respectively; the average specificity was 83.1% and 81.5%, respectively; the positive predictive value was 77.2% and 75.6%, respectively; and the negative predictive value was 84.6% and 81.8%, respectively. The average accuracy of the training group and the testing group for LNM of ECC was 83.6% and 81.2%, respectively; the average sensitivity was 85.8% and 83.2%, respectively; the average specificity was 82.1% and 79.6%, respectively; the positive predictive value was 79.1% and 76.9%, respectively; and the negative predictive value was 89.5% and 86.5%, respectively.

Figure 3. Receiver operating characteristic curves (ROC) of the performance evaluation for (a) differentiation degree prediction of extrahepatic cholangiocarcinoma in the training and testing cohorts and (b) lymphatic node metastasis of extrahepatic cholangiocarcinoma in the training and testing cohorts. AUC: area under the curve.

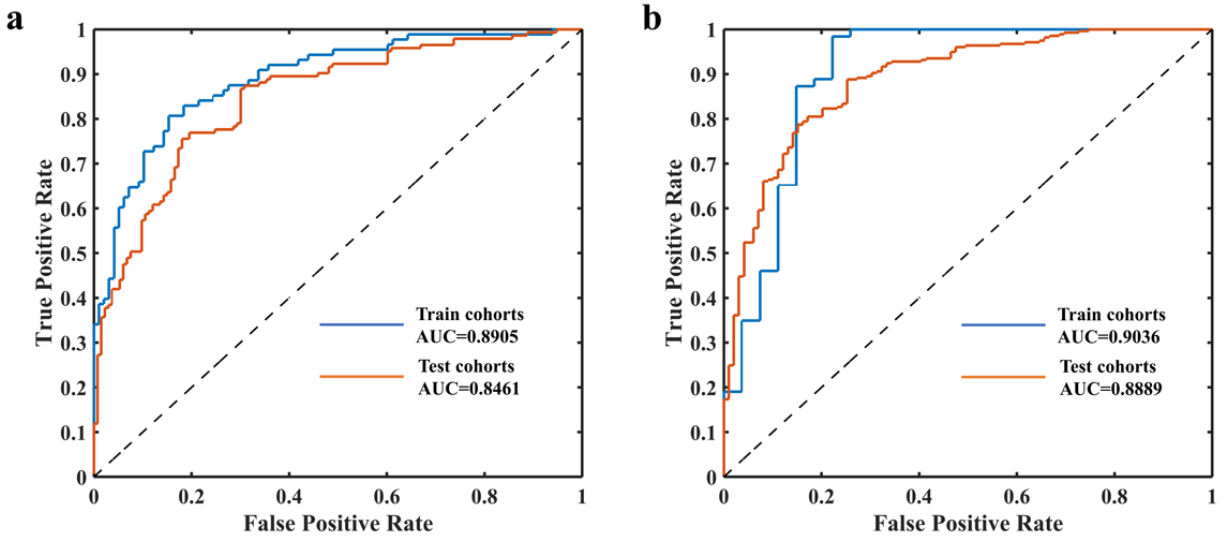


Table 4. The performance of the radiomics prediction model for predicting differentiation degree (DD) and lymph node metastases (LNM) of extrahepatic cholangiocarcinoma (ECC) by using a particle swarm optimization and support vector machine (PSO-SVM) model.

Evaluation indicators (%)	DD of ECC		LNM of ECC	
	Training group	Testing group	Training group	Testing group
Average AUC[a]	89.1[b]	84.6	90.4[b]	88.9
Average accuracy	82.6	80.9	83.6	81.2
Average sensitivity[c]	80.5	78.1	85.8	83.2
Average specificity[d]	83.1	81.5	82.1	79.6
Average PPV[e]	77.2	75.6	79.1	76.9
Average NPV[f]	84.6	81.8	89.5	86.8

[a]AUC: area under the curve.

[b]$P<.001$.

[c]Sensitivity is computed at average radiologist specificity.

[d]Specificity is computed at average radiologist sensitivity.

[e]PPV: positive predictive value; positive predictive value is computed at average radiologist sensitivity.

[f]NPV: negative predictive value.

Discussion

Principal Findings

We developed and validated a PSO-SVM prediction model for DD and LNM of ECC by using a radiomics approach. We performed this study to evaluate ECC and improved the efficiency of clinical diagnosis by using machine learning algorithms and a radiological approach. Our preliminary findings indicate that the radiological model incorporating the patients' MRI image sequence (T1WI, T2WI, DWI) and ADC values has superior diagnostic performance. The prediction performance of this model is shown in Figure 3. In the training and test groups, the average AUC of patients for high, medium, and low DD of ECC were 0.8905 and 0.8461 (the maximum AUC was 0.97), respectively. The average AUC of patients for LNM of ECC were 0.9036 and 0.8889 (with a maximum AUC of 1.00), respectively. Compared with the literature [20,26], our research results have higher prediction accuracy. The entire prediction model has the characteristics of multi-modality and high robustness, which comprehensively considered the radiomics feature of multiple sequences (T1WI, T2WI, DWI,

ADC). Therefore, the proposed PSO-SVM prediction model can help clinicians choose an optimal treatment strategy, improve the prognosis of patients with ECC, and reduce complications, making it a potential postoperative evaluation tool in clinical practice.

It is generally recognized that imaging is the most important method for preoperative evaluation of ECC. However, traditional imaging methods have many defects in accurately evaluating the DD and LNM of ECC. The continuous development of ultrasonography, CT, 18-FDG PET/CT, and MRI technology in medical research have provided a great leap forward with respect to the LNM status of ECC [27-30]. Ercolani et al [29-31] reported that the sensitivity, specificity, and accuracy of CT examination of ECC were 35.2%, 91.8%, and 46.1%, respectively. Lewis et al [32,33] showed that CT and MRI can evaluate the degree of pathological differentiation of ECC. However, the traditional techniques, which mainly rely on the subjective observation of radiologists, have many limitations. Transabdominal ultrasonography may only detect the dilatation of bile ducts in the majority of patients with intraductal tumors. CT can be used for X-ray imaging, but X-ray itself may be harmful to the health of patients. PET/CT is expensive and may be affected by false-positive results of benign lesions, such as biliary tract infection or sclerosing cholangitis [34,35]. Most importantly, it is difficult to analyze the tremendous digital characteristics of the biological features of patients in images using traditional techniques.

In contrast, radiomics can conquer these shortcomings. Researchers of radiomics can develop predictive models for clinical outcomes, such as survival, distant metastasis, and molecular feature classification [34-37], by mining potential associations between the quantitative features and pathophysiological characteristics of images [36-39]. According to our literature review, there is a sparse number of studies on DD and LNM that use a machine learning algorithm combined with radiomics to predict ECC, and the prediction accuracy is low. In this study, we innovatively proposed a PSO-SVM model based on radiomics to predict the DD and LNM of ECC. In the training and testing groups, the average prediction accuracy values of DD and LNM of patients with ECC were 82.6% and 83.6%, respectively, and the average AUC values were 0.8680 and 0.89690, respectively. The prediction results of this model were superior to those obtained from traditional image evaluation, such as ultrasonography, CT, 18-FDG PET/CT, and MRI technology. The results of our research indicate that the PSO-SVM model based on radiomics has potential clinical value as an auxiliary diagnostic method for the preoperative quantitative prediction of DD and LNM of ECC.

Furthermore, in order to use the extracted feature information to describe the shape and internal heterogeneity of the lesion area, the radiological features were integrated with the cellular and molecular features of the lesion to improve the accuracy of diagnosis prediction. So far, only a few studies have reported the relationship between the radiological features and the biological features of cholangiocarcinoma lesions. Researchers discovered that certain texture parameters correlate significantly with microvascular invasion, perineural invasion, differentiation, Ki-67, vascular endothelial growth factor, and cytokeratin 7 based on ultrasonography medical images [40]. They proposed radiomics signatures that have moderate efficiency in predicting the biological behaviors of cholangiocarcinoma noninvasively [40]. Gu-Wei Ji et al [41] regarded a radiomics model based on arterial phase CT scans as a valuable diagnostic tool to forecast LNM of ICC. Zhao et al [42] discovered that the combined model, containing enhancement MRI patterns, vascular endothelial growth factor (VEGFR), and radiomics features, showed a preferable early recurrence predictive performance compared to the radiomics model or clinic radiologic-pathological model alone, with AUC, sensitivity, and specificity values of 0.949, 0.875, and 0.774, respectively. Liang et al [43] showed that the noninvasive radiomics nomogram developed using the radiomics signature and clinical stage could be used to predict early recurrence of ICC after partial hepatectomy. Compared with ultrasound and CT examination, MRI has become the imaging modality of choice for bile duct disease examination, especially for diagnosis and staging of cholangiocarcinoma. The contrast of high soft tissue helps to better discover and identify the infiltrating lesions. Magnetic resonance cholangiopancreatography (MRCP) is the most noninvasive method for evaluating bile ducts, allowing for assessments of tumor spread and the level of obstruction [44]. Dynamic contrast-enhanced MRI can not only provide crucial information about tumors, but it can also flag the appearance of distant metastasis and vascular invasion. The MRI examination can provide precise information on the biliary system, lesion range, and local tumor invasion.

As there were many differences between ECC and other liver lesions, such as origin, morbidity, growth pattern, imaging features, and tumor prognosis, the single evaluation method of ECC using radiological characteristics is prone to diagnostic blind spots. Since the ADC value could describe the diffusion capacity of water molecules in the lesion cells, the tissue structure and functional location of the lesion at the cellular and molecular level could be evaluated by combining the ADC value and radiological characteristics. Therefore, another innovation of this study is that we innovatively integrated 90 radiomics features from 3 MRI sequences (T1WI, T2WI, and DWI) and 30 ADC values to improve the prediction accuracy of the PSO-SVM model. At the same time, during the entire training process, the algorithm was repeatedly optimized with 200 iterations to ensure the reliability of the model. Therefore, our model can provide clinicians with auxiliary decision-making for ECC and provide a more personalized treatment plan for patients.

Limitations

The proposed research has certain limitations and deficiencies. First, since ECC is a rare disease, all patients were obtained from a single medical institution (The Affiliated Hospital of Southwest Medical University) for our study, and the sample number of the cases was relatively small. In order to further improve the accuracy and robustness of the prediction model, the next research work is mainly dedicated to collecting more patient data from other medical institutions. Secondly, the design of the study was retrospective in this paper; thus, there were missing data regarding clinical factors and disease progression. Finally, this model has certain predictive barriers in this study,

which cannot make multi-modal prediction results for patients with time-variance. As the radiomics diagnosis is a systematic project, the models should take into account as many factors as possible, and the radiomics features should be correlated with other clinical results, such as biochemical examination, pathology, radiology, and genomic features, and provide quantitative clinical analysis results. With the development of various hospital information technologies and personal wearable devices, it has become more feasible to use real-time collected health data for comprehensive health management [45,46] or hospital data to support intelligent auxiliary diagnosis and decision-making. Therefore, the multi-modal and big data prediction model for ECC will become the focus of the next research study.

Conclusions

In this paper, we developed a PSO-SVM radiomics model that incorporates the qualitative and quantitative radiomics features and pathological characteristics for predicting DD and LNM of ECC. The techniques used include image sketching, ROI region segmentation, feature extraction, dimension reduction, preprocessing, and classification. This model has the advantages of a simple principle, low computational cost, good robustness, and less manual intervention. The prediction result of the PSO-SVM radiomics model might be useful in the assistance of clinical diagnosis and decision-making, and the guidance of patients toward more individualized and accurate treatment.

Acknowledgments

This work was supported by Sichuan Science and Technology Program (Grant No. 2020YJ0151), the Health Committee of Sichuan province (Grant No. 19PJ151), the Applied Basic Research Program of Southwest Medical University (Grant No.2017-ZRZD-019), and the National Natural Science Foundation of China (NSFC) (Grants NO. 81771937 and No. 81871455).

We thank the staff of the Department of Radiology at the Affiliated Hospital of Southwest Medical University and the Center for Medical Informatics of Peking University for technical assistance.

This study was also partly supported by PKU-Baidu Fund project of Intelligent auxiliary diagnosis using medical images and Research project of constructing health big data platform and service system for medical and nursing combined elderly care institutions.

Authors' Contributions

JS and JL were guarantors of the integrity of the entire study. All the authors were involved in the formulation of the study concepts, study design, data acquisition, and data analysis and interpretation. All the authors were involved in manuscript drafting, manuscript revision, or approval of the final version of the manuscript. GZ, XH, CY, and XY participated in literature research. JS reviewed clinical studies. All authors were involved in experimental studies. XY performed statistical analysis and manuscript editing.

Conflicts of Interest

None declared.

References

1. Clements O, Eliahoo J, Kim JU, Taylor-Robinson SD, Khan SA. Risk factors for intrahepatic and extrahepatic cholangiocarcinoma: A systematic review and meta-analysis. Journal of Hepatology 2020 Jan;72(1):95-103. [doi: 10.1016/j.jhep.2019.09.007]
2. Lin H, Yang L, Tian F, Nie S, Zhou H, Liu J, et al. Up-regulated LncRNA-ATB regulates the growth and metastasis of cholangiocarcinoma via miR-200c signals . OTT 2019 Sep;Volume 12:7561-7571. [doi: 10.2147/ott.s217676]
3. Esnaola NF, Meyer JE, Karachristos A, Maranki JL, Camp ER, Denlinger CS. Evaluation and management of intrahepatic and extrahepatic cholangiocarcinoma. Cancer 2016 Jan 22;122(9):1349-1369. [doi: 10.1002/cncr.29692]
4. Andrianello S, Paiella S, Allegrini V, Ramera M, Pulvirenti A, Malleo G, et al. Pancreaticoduodenectomy for distal cholangiocarcinoma: surgical results, prognostic factors, and long-term follow-up. Langenbecks Arch Surg 2015 Jul 2;400(5):623-628. [doi: 10.1007/s00423-015-1320-0]
5. Kiriyama M, Ebata T, Aoba T, Kaneoka Y, Arai T, Shimizu Y, et al. Prognostic impact of lymph node metastasis in distal cholangiocarcinoma. Br J Surg 2015 Jan 22;102(4):399-406. [doi: 10.1002/bjs.9752]
6. Suzuki S, Shimoda M, Shimazaki J, Maruyama T, Oshiro Y, Nishida K, et al. Number of positive lymph nodes and lymphatic invasion are significant prognostic factors after pancreaticoduodenectomy for distal cholangiocarcinoma . CEG 2019 Jun;Volume 12:255-262. [doi: 10.2147/ceg.s207333]
7. Hu H, Jin Y, Shrestha A, Ma W, Wang J, Liu F, et al. Predictive factors of early recurrence after R0 resection of hilar cholangiocarcinoma: A single institution experience in China. Cancer Med 2019 Mar 13;8(4):1567-1575. [doi: 10.1002/cam4.2052]
8. Kato Y, Takahashi S, Gotohda N, Konishi M. Prognostic Impact of the Initial Postoperative CA19-9 Level in Patients with Extrahepatic Bile Duct Cancer. J Gastrointest Surg 2016 Jun 1;20(8):1435-1443. [doi: 10.1007/s11605-016-3180-5]

9. Doherty B, Nambudiri VE, Palmer WC. Update on the Diagnosis and Treatment of Cholangiocarcinoma. Curr Gastroenterol Rep 2017 Jan 21;19(1). [doi: 10.1007/s11894-017-0542-4]
10. Jhaveri KS, Hosseini-Nik H. MRI of cholangiocarcinoma. J. Magn. Reson. Imaging 2014 Dec 01;42(5):1165-1179. [doi: 10.1002/jmri.24810]
11. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018 Dec;18(8):500-510 [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
12. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
13. Nanashima A, Sakamoto I, Hayashi T, Tobinaga S, Araki M, Kunizaki M, et al. Preoperative Diagnosis of Lymph Node Metastasis in Biliary and Pancreatic Carcinomas: Evaluation of the Combination of Multi-detector CT and Serum CA19-9 Level. Dig Dis Sci 2010 Mar 18;55(12):3617-3626. [doi: 10.1007/s10620-010-1180-y]
14. Noji T, Kondo S, Hirano S, Tanaka E, Suzuki O, Shichinohe T. Computed tomography evaluation of regional lymph node metastases in patients with biliary cancer. Br J Surg 2007 Sep 13;95(1):92-96. [doi: 10.1002/bjs.5920]
15. Kiriyama M, Ebata T, Aoba T, Kaneoka Y, Arai T, Shimizu Y, et al. Prognostic impact of lymph node metastasis in distal cholangiocarcinoma. Br J Surg 2015 Jan 22;102(4):399-406. [doi: 10.1002/bjs.9752]
16. Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. Eur J Nucl Med Mol Imaging 2019 Jun 18;46(13):2656-2672. [doi: 10.1007/s00259-019-04372-x]
17. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016 Feb;278(2):563-577. [doi: 10.1148/radiol.2015151169]
18. Jiang H, Liu X, Chen J, Wei Y, Lee JM, Cao L, et al. Man or machine? Prospective comparison of the version 2018 EASL, LI-RADS criteria and a radiomics model to diagnose hepatocellular carcinoma. Cancer Imaging 2019 Dec 5;19(1). [doi: 10.1186/s40644-019-0266-9]
19. Li L, Mu W, Wang Y, Liu Z, Liu Z, Wang Y, et al. A Non-invasive Radiomic Method Using 18F-FDG PET Predicts Isocitrate Dehydrogenase Genotype and Prognosis in Patients With Glioma. Front. Oncol 2019 Nov 14;9. [doi: 10.3389/fonc.2019.01183]
20. Bibault J, Giraud P, Housset M, Durdux C, Taieb J, Berger A, et al. Author Correction: Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. Sci Rep 2018 Nov 12;8(1). [doi: 10.1038/s41598-018-35359-7]
21. Cong M, Feng H, Ren J, Xu Q, Cong L, Hou Z, et al. Development of a predictive radiomics model for lymph node metastases in pre-surgical CT-based stage IA non-small cell lung cancer. Lung Cancer 2020 Jan;139:73-79. [doi: 10.1016/j.lungcan.2019.11.003]
22. Li H, Zhu Y, Burnside ES, Huang E, Drukker K, Hoadley KA, et al. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. npj Breast Cancer 2016 May 11;2(1). [doi: 10.1038/npjbcancer.2016.12]
23. Xiao G, Rong W, Hu Y, Shi Z, Yang Y, Ren J, et al. MRI Radiomics Analysis for Predicting the Pathologic Classification and TNM Staging of Thymic Epithelial Tumors: A Pilot Study. American Journal of Roentgenology 2020 Feb;214(2):328-340. [doi: 10.2214/ajr.19.21696]
24. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th Edition of the AJCC Cancer Staging Manual and the Future of TNM. Ann Surg Oncol 2010 Feb 24;17(6):1471-1474. [doi: 10.1245/s10434-010-0985-4]
25. Elreedy D, Atiya AF. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. Information Sciences 2019 Dec;505:32-64. [doi: 10.1016/j.ins.2019.07.070]
26. Lewis S, Besa C, Wagner M, Jhaveri K, Kihira S, Zhu H, et al. Prediction of the histopathologic findings of intrahepatic cholangiocarcinoma: qualitative and quantitative assessment of diffusion-weighted imaging. Eur Radiol 2017 Dec 12;28(5):2047-2057. [doi: 10.1007/s00330-017-5156-6]
27. Meng Z, Lin X, Zhu J, Han S, Chen Y. A nomogram to predict lymph node metastasis before resection in intrahepatic cholangiocarcinoma. Journal of Surgical Research 2018 Jun;226:56-63. [doi: 10.1016/j.jss.2018.01.024]
28. Ercolani G, Grazi GL, Ravaioli M, Grigioni WF, Cescon M, Gardini A, et al. The Role of Lymphadenectomy for Liver Tumors. Annals of Surgery 2004;239(2):202-209. [doi: 10.1097/01.sla.0000109154.00020.e0]
29. Li X, Zhang Y, Zhang Y. 18F-FDG PET/CT may be a suitable method for preoperative diagnosis and evaluation of Chinese older patients with hilar cholangiocarcinoma. BMC Geriatr 2018 Jul 6;18(1). [doi: 10.1186/s12877-018-0846-8]
30. D'Antuono F, De Luca S, Mainenti PP, Mollica C, Camera L, Galizia G, et al. Comparison Between Multidetector CT and High–Field 3T MR Imaging in Diagnostic and Tumour Extension Evaluation of Patients with Cholangiocarcinoma. J Gastrointest Canc 2019 Jul 29;51(2):534-544. [doi: 10.1007/s12029-019-00276-z]
31. Hu H. Prognostic factors and long-term outcomes of hilar cholangiocarcinoma: A single-institution experience in China. WJG 2016;22(8):2601. [doi: 10.3748/wjg.v22.i8.2601]
32. Pawlik TM, Pulitano C, Alexandrescu S, Gamblin TC, Ferrone C, Sotiropoulos G, et al. Intrahepatic cholangiocarcinoma: An international, multi-institutional analysis of prognostic factors and lymph node assessment. JCO 2011 Feb 01;29(4_suppl):162-162. [doi: 10.1200/jco.2011.29.4_suppl.162]

33. Amini N, Ejaz A, Spolverato G, Maithel SK, Kim Y, Pawlik TM. Management of Lymph Nodes During Resection of Hepatocellular Carcinoma and Intrahepatic Cholangiocarcinoma: A Systematic Review. J Gastrointest Surg 2014 Oct 10;18(12):2136-2148. [doi: 10.1007/s11605-014-2667-1]
34. Anderson C. Fluorodeoxyglucose PET imaging in the evaluation of gallbladder carcinoma and cholangiocarcinoma. Journal of Gastrointestinal Surgery 2004 Jan 01;8(1):90-97. [doi: 10.1016/j.gassur.2003.10.003]
35. Wakabayashi H, Akamoto S, Yachida S, Okano K, Izuishi K, Nishiyama Y, et al. Significance of fluorodeoxyglucose PET imaging in the diagnosis of malignancies in patients with biliary stricture. European Journal of Surgical Oncology (EJSO) 2005 Dec;31(10):1175-1179. [doi: 10.1016/j.ejso.2005.05.012]
36. Caudell JJ, Torres-Roca JF, Gillies RJ, Enderling H, Kim S, Rishi A, et al. The future of personalised radiotherapy for head and neck cancer. The Lancet Oncology 2017 May;18(5):e266-e273. [doi: 10.1016/s1470-2045(17)30252-8]
37. Lambin P, Leijenaar RT, Deist TM, Peerlings J, de Jong EE, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017 Oct 4;14(12):749-762. [doi: 10.1038/nrclinonc.2017.141]
38. Limkin E, Sun R, Dercle L, Zacharaki E, Robert C, Reuzé S, et al. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. Annals of Oncology 2017 Jun;28(6):1191-1206. [doi: 10.1093/annonc/mdx034]
39. Verma V, Simone C, Krishnan S, Lin S, Yang J, Hahn S. The rise of radiomics and implications for oncologic management. J Natl Cancer Inst 2017;109(7). [doi: 10.1093/jnci/djx055]
40. Peng Y, Zhou C, Lin P, Wen D, Wang X, Zhong X, et al. Preoperative Ultrasound Radiomics Signatures for Noninvasive Evaluation of Biological Characteristics of Intrahepatic Cholangiocarcinoma. Academic Radiology 2020 Jun;27(6):785-797. [doi: 10.1016/j.acra.2019.07.029]
41. Ji G, Zhu F, Zhang Y, Liu X, Wu F, Wang K, et al. A radiomics approach to predict lymph node metastasis and clinical outcome of intrahepatic cholangiocarcinoma. Eur Radiol 2019 Mar 26;29(7):3725-3735. [doi: 10.1007/s00330-019-06142-7]
42. Zhao L, Ma X, Liang M, Li D, Ma P, Wang S, et al. Prediction for early recurrence of intrahepatic mass-forming cholangiocarcinoma: quantitative magnetic resonance imaging combined with prognostic immunohistochemical markers. Cancer Imaging 2019 Jul 15;19(1). [doi: 10.1186/s40644-019-0234-4]
43. Liang W, Xu L, Yang P, Zhang L, Wan D, Huang Q, et al. Novel Nomogram for Preoperative Prediction of Early Recurrence in Intrahepatic Cholangiocarcinoma. Front. Oncol 2018 Sep 4;8. [doi: 10.3389/fonc.2018.00360]
44. Romagnuolo J, Bardou M, Rahme E, Joseph L, Reinhold C, Barkun AN. Magnetic Resonance Cholangiopancreatography. Ann Intern Med 2003 Oct 07;139(7):547. [doi: 10.7326/0003-4819-139-7-200310070-00006]
45. Xie J, Wen D, Liang L, Jia Y, Gao L, Lei J. Evaluating the Validity of Current Mainstream Wearable Devices in Fitness Tracking Under Various Physical Activities: Comparative Study. JMIR Mhealth Uhealth 2018 Apr 12;6(4):e94 [FREE Full text] [doi: 10.2196/mhealth.9754] [Medline: 29650506]
46. Wen D, Zhang X, Liu X, Lei J. Evaluating the Consistency of Current Mainstream Wearable Devices in Health Monitoring: A Comparison Under Free-Living Conditions. J Med Internet Res 2017 Mar 07;19(3):e68 [FREE Full text] [doi: 10.2196/jmir.6874] [Medline: 28270382]

Abbreviations

ADC: apparent diffusion coefficient
AUC: area under the curve
DD: differentiation degree
DDR: data dimensionality reduction
DICOM: Digital Imaging and Communications in Medicine
DWI: diffusion-weighted imaging
ECC: extrahepatic cholangiocarcinoma
ICC: intraclass correlation coefficient
LNM: lymph node metastasis
MRI: magnetic resonance imaging
NPV: negative predictive value
PACS: picture archiving and communication system
PPV: positive predictive value
PSO-SVM: particle swarm optimization and support vector machine
ROC: receiver operating characteristic
ROI: region of interest
T1WI: T1-weighted imaging
T2WI: T2-weighted imaging
18F–FDG-PET/CT: 18-fluorodeoxyglucose positron emission tomography/computerized tomography

Edited by C Lovis, G Eysenbach; submitted 17.08.20; peer-reviewed by F Wang, D Liu; comments to author 06.09.20; accepted 18.09.20; published 05.10.20.

Please cite as:
Yao X, Huang X, Yang C, Hu A, Zhou G, Ju M, Lei J, Shu J
A Novel Approach to Assessing Differentiation Degree and Lymph Node Metastasis of Extrahepatic Cholangiocarcinoma: Prediction Using a Radiomics-Based Particle Swarm Optimization and Support Vector Machine Model
JMIR Med Inform 2020;8(10):e23578
URL: https://medinform.jmir.org/2020/10/e23578
doi:10.2196/23578
PMID:33016889

Original Paper

Assessment of Myosteatosis on Computed Tomography by Automatic Generation of a Muscle Quality Map Using a Web-Based Toolkit: Feasibility Study

Dong Wook Kim[1], MD; Kyung Won Kim[1], MD, PhD; Yousun Ko[2], PhD; Taeyong Park[3], PhD; Seungwoo Khang[3], MS; Heeryeol Jeong[3], MS; Kyoyeong Koo[3], MS; Jeongjin Lee[3], PhD; Hong-Kyu Kim[4], MD, PhD; Jiyeon Ha[1], MD; Yu Sub Sung[5,6], PhD; Youngbin Shin[2], MS

[1]Department of Radiology and Research Institute of Radiology, Asan Medical Center, Seoul, Republic of Korea

[2]Biomedical Research Center, Asan Medical Center, Seoul, Republic of Korea

[3]School of Computer Science and Engineering, Soongsil University, Seoul, Republic of Korea

[4]Health Screening and Promotion Center, Asan Medical Center, Seoul, Republic of Korea

[5]Clinical Research Center, Asan Medical Center, Seoul, Republic of Korea

[6]Department of Convergence Medicine, University of Ulsan College of Medicine, Seoul, Republic of Korea

Corresponding Author:
Kyung Won Kim, MD, PhD
Department of Radiology and Research Institute of Radiology
Asan Medical Center
88 Olympic-ro, 43-gil, Songpa-gu
Seoul, 05505
Republic of Korea
Phone: 82 2 3010 4377
Fax: 82 2 476 4719
Email: medimash@gmail.com

Abstract

Background: Muscle quality is associated with fatty degeneration or infiltration of the muscle, which may be associated with decreased muscle function and increased disability.

Objective: The aim of this study is to evaluate the feasibility of automated quantitative measurements of the skeletal muscle on computed tomography (CT) images to assess normal-attenuation muscle and myosteatosis.

Methods: We developed a web-based toolkit to generate a muscle quality map by categorizing muscle components. First, automatic segmentation of the total abdominal muscle area (TAMA), visceral fat area, and subcutaneous fat area was performed using a predeveloped deep learning model on a single axial CT image at the L3 vertebral level. Second, the Hounsfield unit of each pixel in the TAMA was measured and categorized into 3 components: normal-attenuation muscle area (NAMA), low-attenuation muscle area (LAMA), and inter/intramuscular adipose tissue (IMAT) area. The myosteatosis area was derived by adding the LAMA and IMAT area. We tested the feasibility of the toolkit using randomly selected healthy participants, comprising 6 different age groups (20 to 79 years). With stratification by sex, these indices were compared between age groups using 1-way analysis of variance (ANOVA). Correlations between the myosteatosis area or muscle densities and fat areas were analyzed using Pearson correlation coefficient r.

Results: A total of 240 healthy participants (135 men and 105 women) with 40 participants per age group were included in the study. In the 1-way ANOVA, the NAMA, LAMA, and IMAT were significantly different between the age groups in both male and female participants ($P \leq .004$), whereas the TAMA showed a significant difference only in male participants (male, $P < .001$; female, $P = .88$). The myosteatosis area had a strong negative correlation with muscle densities ($r = -0.833$ to -0.894), a moderate positive correlation with visceral fat areas ($r = 0.607$ to 0.669), and a weak positive correlation with the subcutaneous fat areas ($r = 0.305$ to 0.441).

Conclusions: The automated web-based toolkit is feasible and enables quantitative CT assessment of myosteatosis, which can be a potential quantitative biomarker for evaluating structural and functional changes brought on by aging in the skeletal muscle.

(JMIR Med Inform 2020;8(10):e23049) doi:10.2196/23049

KEYWORDS

body composition; muscle; skeletal; sarcopenia; computed tomography; x-ray; scan; web-based tool; feasibility; automated; CT

Introduction

Measurement of muscle is one of the fastest growing research areas in medicine, as quantity and quality of muscle in individuals are reportedly associated with morbidity and mortality in various diseases [1,2]. Muscle mass is frequently assessed using medical imaging such as computed tomography (CT) or magnetic resonance imaging (MRI) [3]. However, an issue of discrepancy between the muscle mass measured on imaging and muscle functions, such as strength and mobility, has been raised. Therefore, the assessment of muscle quality by imaging is gaining a focus in the research and clinical diagnoses of relevant diseases. Specifically, the CT density or attenuation can be quantified using a standardized unit (ie, Hounsfield unit [HU]), enabling a standardized muscle quality evaluation on CT.

Recent studies demonstrated that muscle quality is associated with fatty degeneration or fatty infiltration of the muscle (ie, myosteatosis), which may be associated with decreased muscle function and increased disability [4]. Based on the imaging characteristics in CT, fat can be stored as follows: (1) in the intermuscular adipose tissue, which is observed as gross fat between muscle groups; (2) in the intramuscular adipose tissue as extramyocellular lipids, which are denoted as gross fat tissues between muscle fibers in the same muscle group; and (3) in the intramyocellular lipid droplets, which are not visually demonstrated as fat but are indicated by decreased muscle density on CT [5]. It is known that the intermuscular and intramuscular adipose tissues (IMAT) can be depicted as areas with gross fat density of –190 HU to –30 HU, whereas the intramyocellular lipid can be reflected as low-attenuation muscle area (LAMA) with –29 HU to +29 HU within the muscle [6,7]. The healthy muscle without myosteatosis is observed as normal-attenuation muscle area (NAMA) with +30 HU to +150 HU [6].

The muscle density at the L3 vertebral level has been used for muscle quality evaluation on abdominal CT in majority of prior studies because it easily measures muscle density [1,8]. Recent advancements in technology evaluate myosteatosis based on the distribution and amount of IMAT and LAMA. In this paper, we evaluate the feasibility of fully automated quantitative measurement of the skeletal muscle on CT using a web-based toolkit for assessing normal-attenuation muscle and myosteatosis.

Methods

This study, which uses retrospective data, was approved by the institutional review board of Asan Medical Center, and a waiver for written informed consent was obtained.

Generation of a Muscle Quality Map

Among the abdominal CT images, a single axial image at the inferior endplate level of the L3 vertebra at the portal venous phase was selected for image assessment [9]. The selected CT image was saved as a digital imaging and communications in medicine (DICOM) file and was uploaded to our web-based toolkit in a drag-and-drop manner. First, we performed automatic segmentation of the abdominal compartments using a predeveloped deep learning model based on a fully convolutional network [10]. This model was reported to segment abdominal body compartments into the total abdominal muscle area (TAMA), subcutaneous fat area, and visceral fat area with a dice similarity coefficient of 0.97 and mean cross-sectional area error of 2.26% [10].

Second, we performed postprocessing to divide the TAMA into 3 muscle components using the pixel-wise measurement of CT density with the following range of HU for each component [6,7]: NAMA with +30 HU to +150 HU, LAMA with –29 HU to +29 HU, and IMAT with –190 HU to –30 HU. Subsequently, a muscle quality map was generated by combining NAMA, LAMA, and IMAT, which were displayed in different colors on the web-based toolkit (Figure 1A), enabling the quantitative assessment of muscle properties based on histogram analysis. For example, the participants with similar TAMA values can be differentiated into participants with healthy muscles (ie, mostly composed of NAMA) and those with fatty degenerated muscles (ie, comprising several LAMA and IMAT). The technique has been packaged and can be accessed at an iAID Viewer repository [11]. Figure 1 shows the process and examples of muscle quality map generation using the web-based iAID toolkit. A CT image is uploaded to the toolkit in a drag-and-drop manner. Muscle quality map and its measurement values are displayed upon clicking on the “L3 Area Measure” button (Figure 1A). Muscle quality maps and a histogram for a 28-year-old man with higher quality muscle are shown in Figure 1B; and muscle quality maps and a histogram for a 72-year-old man with lower quality muscle are shown in Figure 1C.

Figure 1. Muscle quality map generation using an automated web-based toolkit. IMAT: inter/intramuscular adipose tissue area; LAMA: low-attenuation muscle area; NAMA: normal-attenuation muscle area; SMA: skeletal muscle area; TAMA: total abdominal muscle area.

The quality of the generated muscle quality map was evaluated by a board-certified abdominal radiologist by comparing the original CT image and the muscle quality map image. If the segmentation quality was not acceptable, manual adjustment was performed using a stand-alone software.

Feasibility Study in Healthy Participants

From the electronic database of Asan Medical Center, we retrospectively identified 3928 abdominal CT scans from healthy participants who had no disease or history of previous treatments and had undergone CT for medical checkup or workup for liver donation during the period from January 2008 to June 2016. We used the block randomization technique to randomly select 240 healthy participants (135 men and 105 women) belonging to 6 different age groups, from 20 to 79 years, with 10-year intervals (40 participants per age group).

CT was performed using a 16-channel (LightSpeed 16, GE Healthcare; Somatom Sensation 16, Siemens Medical Solution), a 64-channel (Discovery CT 750 HD, GE Healthcare; Somatom Definition AS, Siemens Medical Solution), or a 128-channel (Somatom Definition Flash, Siemens Medical Solution) CT scanner with the following parameters: 120 kVp, 200–220 mAs (maximum tube current with automated dose modulation), 1.5-5 mm section thickness and intervals, and a pitch of 0.6-1. Contrast agents were administered at a rate of 3-4 mL/s, and CT images (regardless of the CT protocols) were obtained, including the portal venous phase (65-72 seconds after contrast agent injection) in the craniocaudal direction.

Statistical Analysis

To eliminate effects by different sexes, participants within each subgroup were stratified by sex. The mean values of muscle components (TAMA, NAMA, LAMA, and IMAT) were compared among the age groups using 1-way analysis of variance. A $P<.05$ was considered statistically significant. Correlation between the mean density of TAMA or skeletal muscle area (SMA; ie, NAMA + LAMA) and muscle components related to the adipose tissue (ie, LAMA, IMAT, or myosteatosis area represented by LAMA + IMAT) were calculated using Pearson correlation (absolute magnitude of Pearson correlation coefficient r=0-0.1, negligible correlation; r=0.1-0.4, weak correlation; r=0.4-0.7, moderate correlation; r=0.7-0.9, strong correlation; and r=0.9-1, very strong correlation) [12]. In addition, to investigate the correlation between the adipose tissue within and outside the muscle compartment, the correlation between the aforementioned muscle components related to the adipose tissue—LAMA, IMAT, and myosteatosis area—and the visceral and subcutaneous fat area was also analyzed using Pearson correlation. All statistical analyses were performed using the SPSS, version 21 (IBM Corp).

Results

The characteristics of the 240 participants are presented in Table 1. In male participants, height and weight were significantly different across age groups (P<.001). In female participants, height was significantly different across age groups (P<.001). Body mass index was not significantly different both in male (P=.11) and female participants (P=.13).

Table 1. Characteristics of the study population.

Characteristics	Age group, years						*P* value
	20-29	30-39	40-49	50-59	60-69	70-79	
Participants, n	40	40	40	40	40	40	N/A[a]
Age (years), mean (SD)	25.1 (2.83)	34.5 (2.39)	45.2 (2.93)	54.2 (2.63)	62.5 (2.62)	72.8 (2.35)	N/A[a]
Male:female ratio	29:11	17:23	22:18	15:25	25:15	27:13	.007
Height (m), mean (SD)	1.71 (0.08)	1.65 (0.10)	1.67 (0.09)	1.59 (0.07)	1.62 (0.08)	1.63 (0.09)	<.001
Male	1.74 (0.06)	1.74 (0.06)	1.73 (0.06)	1.67 (0.04)	1.67 (0.04)	1.68 (0.05)	<.001
Female	1.64 (0.07)	1.59 (0.06)	1.59 (0.05)	1.55 (0.04)	1.54 (0.04)	1.53 (0.05)	<.001
Weight (kg), mean (SD)	69.8 (14.9)	66.8 (14.5)	67.5 (11.7)	62.3 (7.8)	63.3 (9.5)	62.8 (10.0)	.02
Male	74.2 (14.7)	78.9 (12.7)	73.4 (10.9)	69.8 (4.0)	67.2 (9.0)	65.7 (8.9)	.001
Female	58.3 (7.2)	57.8 (7.7)	60.3 (8.1)	57.8 (5.8)	56.8 (6.4)	56.8 (9.7)	.78
BMI (kg/m^2), mean (SD)	23.6 (4.0)	24.3 (3.4)	24.2 (3.2)	24.5 (2.2)	24.0 (2.6)	23.5 (3.0)	.65
Male	24.3 (4.3)	26.1 (3.6)	24.7 (3.6)	25.1 (2.0)	24.1 (2.7)	23.2 (2.5)	.11
Female	21.7 (2.2)	23.0 (2.5)	23.7 (2.6)	24.2 (2.2)	24.0 (2.6)	24.2 (3.9)	.13
Cause of CT[b] examination, n (%)							<.001
Medical checkup	27 (67.5)	15 (37.5)	15 (37.5)	20 (50.0)	31 (77.5)	39 (97.5)	
Workup for liver donation	13 (32.5)	25 (67.5)	25 (62.5)	20 (50.0)	9 (22.5)	1 (2.5)	

[a]N/A: not applicable.

[b]CT: computed tomography.

Figure 2 illustrates the mean areas of muscle components among different age groups of male and female participants.

The detailed information is presented in Table 2. TAMA showed a significant difference only in male participants (male, P<.001; female, P=.88). The age group of 30s had the highest TAMA (mean 187.9 cm^2, SD 29.6 cm^2) with a gradual decrease in area with aging in male participants. The TAMA in female participants did not differ significantly across the age groups. In contrast, NAMA, LAMA, and IMAT were significantly different among the age groups both in male and female participants (P<.001). In male participants, NAMA showed a gradual decrease with age (peak age: 20 years; mean 153.4 cm^2, SD 20.0 cm^2). In female participants, there was a peak in the 30s age group (mean 92.6 cm^2, SD 15.1 cm^2), followed by a gradual decrease in area with aging. The myosteatosis area showed a gradual increase with aging in female participants.

Figure 2. Area of muscle components according to the age groups in male and female participants. IMAT: inter/intramuscular adipose tissue area; LAMA: low-attenuation muscle area; NAMA: normal-attenuation muscle area; TAMA: total abdominal muscle area.

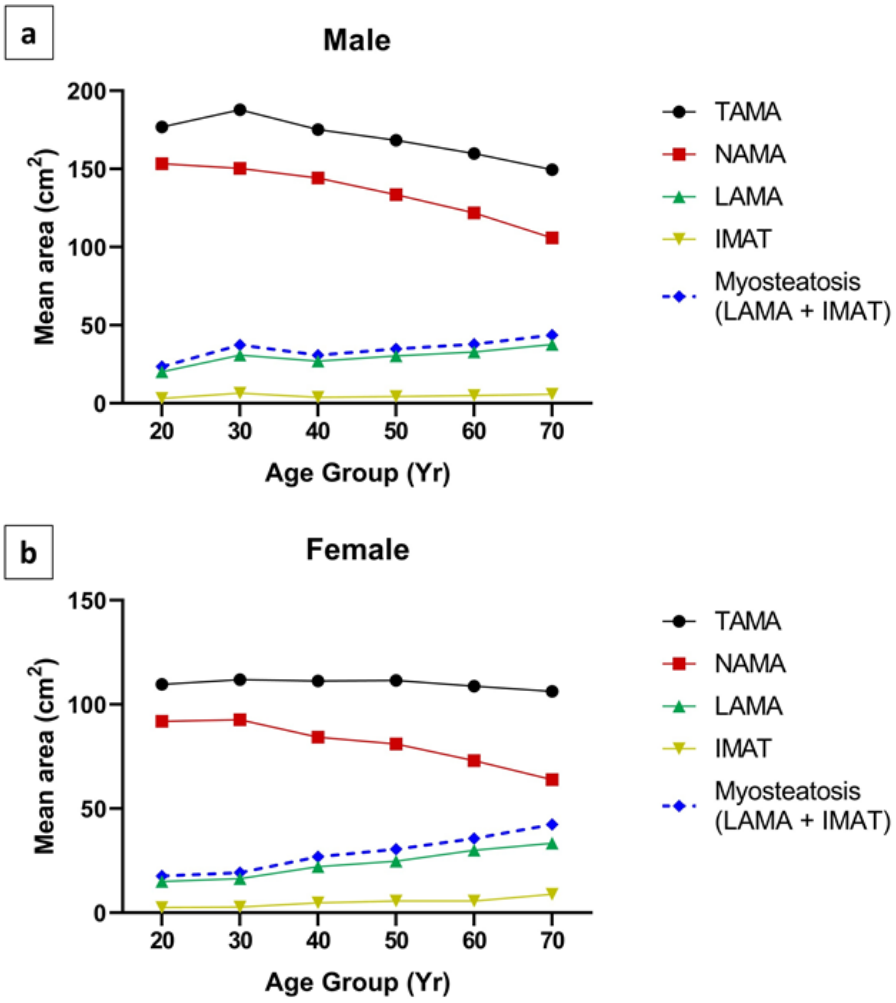


Table 2. Mean area and indices of muscle components in male and female participants.

Muscle components in participants	Age group, years						*P* value
	20-29	30-39	40-49	50-59	60-69	70-79	
Male participants, n	29	17	22	15	25	27	
TAMA[a], mean (SD)	177.0 (22.9)	187.9 (29.6)	175.3 (24.5)	168.4 (18.0)	160.0 (22.5)	149.6 (21.0)	<.001
NAMA[b], mean (SD)	153.4 (20.0)	150.4 (26.1)	144.3 (19.5)	133.5 (19.1)	122.0 (16.7)	105.9 (17.5)	<.001
LAMA[c], mean (SD)	20.2 (8.9)	30.9 (11.7)	27.0 (10.5)	30.5 (6.7)	32.9 (10.3)	37.8 (14.4)	<.001
IMAT[d], mean (SD)	3.3 (2.5)	6.6 (5.1)	3.9 (2.4)	4.4 (1.7)	5.1 (2.4)	5.9 (3.6)	.004
Myosteatosis area (LAMA+IMAT), mean (SD)	23.5 (11.0)	37.5 (16.3)	31.0 (12.3)	34.9 (7.9)	38.0 (11.6)	43.7 (17.6)	<.001
Female participants, n	11	23	18	25	15	13	
TAMA, mean (SD)	109.6 (10.9)	112.0 (15.3)	111.3 (11.1)	111.5 (15.4)	108.8 (13.8)	106.3 (17.2)	.88
NAMA, mean (SD)	91.9 (9.5)	92.6 (15.1)	84.3 (11.2)	81.0 (13.1)	73.1 (14.2)	63.9 (17.3)	<.001
LAMA, mean (SD)	15.1 (5.7)	16.5 (5.7)	22.1 (8.4)	24.9 (8.1)	30.0 (8.8)	33.5 (11.2)	<.001
IMAT, mean (SD)	2.6 (1.4)	2.8 (1.6)	4.8 (3.1)	5.7 (2.5)	5.7 (3.3)	8.9 (5.1)	<.001
Myosteatosis area (LAMA+IMAT), mean (SD)	17.7 (6.9)	19.3 (7.1)	27.0 (11.1)	30.5 (10.0)	35.7 (11.4)	42.4 (15.6)	<.001

[a]TAMA: total abdominal muscle area.

[b]NAMA: normal-attenuation muscle area.

[c]LAMA: low-attenuation muscle area.

[d]IMAT: inter/intramuscular adipose tissue area.

All muscle components related to the adipose tissue (ie, LAMA, IMAT, and myosteatosis area) had a moderate-to-strong negative correlation with the mean density of TAMA (*r*=–0.629 to –0.884) and SMA (*r*=–0.647 to –0.898) (Table 3). They had a moderate positive correlation (*r*=0.474 to 0.686) with visceral fat area and weak-to-moderate positive correlation (*r*= 0.274 to 0.459) with subcutaneous fat area in both male and female participants (Table 3). In particular, LAMA showed higher correlation with visceral and subcutaneous fat compartments than that shown by IMAT. All the correlations in Table 3 are statistically significant (*P*<.05).

Table 3. Correlation between muscle components containing fat and muscle density and fat compartments.

Muscle components	Mean density of TAMA[a]		Mean density of SMA[b] (NAMA[c]+LAMA[d])		Visceral fat area		Subcutaneous fat area	
	Male	Female	Male	Female	Male	Female	Male	Female
LAMA	–0.845	–0.884	–0.847	–0.898	0.686	0.617	0.274	0.459
IMAT[e]	–0.629	–0.728	–0.647	–0.763	0.474	0.495	0.365	0.329
Myosteatosis area (LAMA+IMAT)	–0.833	–0.874	–0.838	–0.894	0.669	0.607	0.305	0.441

[a]TAMA: total abdominal muscle area.

[b]SMA: skeletal muscle area.

[c]NAMA: normal-attenuation muscle area.

[d]LAMA: low-attenuation muscle area.

[e]IMAT: inter/intramuscular adipose tissue area.

Discussion

Principal Results

We developed a web-based toolkit to generate a pixel-based automatic categorization of muscle components (ie, muscle quality map) within the generated segmented muscle compartment using predeveloped deep learning models. These muscle quality maps can illustrate spatial distribution of fat infiltration and provide insights into the muscle quality in individuals. In our study, NAMA gradually decreased with age; and LAMA and IMAT gradually increased with age. These results indicate that fat infiltration or fatty degeneration increases with aging. Therefore, NAMA might be an effective imaging

biomarker for evaluation of muscle in individuals. Additionally, LAMA and IMAT might be used as biomarkers of myosteatosis and related diseases, such as metabolic syndrome [1,5].

Comparison With Previous Work

Sarcopenia, which is defined as the loss of muscle mass or function, is associated with increased morbidity and mortality in various diseases [1,2]. Recently, the revised consensus of the European Working Group on Sarcopenia in Older People emphasized muscle strength and quality as a key characteristic of sarcopenia in the definition and diagnostic criteria of sarcopenia [8]. Accordingly, both qualitative and quantitative evaluations of muscle are highly recommended in sarcopenia diagnosis. Among the diagnostic tests for muscle quantity measurements, muscle quality evaluation can be performed only using cross-sectional imaging such as CT and MRI. The dual-energy x-ray absorptiometry and bioelectrical impedance analysis cannot differentiate healthy muscles from fatty degenerated muscles. Currently, the CT density or attenuation is well-calibrated and standardized (ie, HU of zero for water) across the CT acquisition protocols and imaging machines, whereas the signal intensity of MRI might differ between protocols and machines. Therefore, the muscle quality map on CT might be the optimal option to evaluate myosteatosis.

There are several software programs (both open source and licensed) that analyze muscle and fat composition in CT [13]. However, a software providing muscle quality maps at the L3 vertebral level is not yet available. We developed our own software for performing a fully automated segmentation of body compartments and generating a muscle quality map at the L3 vertebral level [11]. This software is publicly available for academic purposes. Considering the increasing recognition of myosteatosis and sarcopenia as prognostic biomarkers for various diseases in older patients, measuring muscle quality and muscle mass easily is highly recommended for researchers and clinicians.

Limitations

Our study has some limitations. First, we tested the muscle quality map in a relatively small number of participants (N=240). We randomly assigned 40 participants into each age group; thus, our study is an experimental study rather than an epidemiologic study and may not reflect the real-world trends of muscle quality with aging. There was an imbalance between male and female participants, which may preclude the real-world data. In addition, we could not provide any criteria or cutoff value to diagnose myosteatosis. Therefore, an epidemiological study using large populations might be required. Second, the segmented areas of each muscle component might be affected by the different CT protocols and scanners. However, this is beyond the scope of the current study and needs to be evaluated in a separate study. Third, the general applicability of our toolkit should be validated in an external dataset, particularly because the predeveloped segmentation model was trained at our institution and could thus lead to overfitting.

Conclusion

In conclusion, the automated web-based toolkit is feasible and enables a quantitative CT assessment of normal-attenuation muscle and myosteatosis, which can be effective quantitative biomarkers for evaluating structural and functional changes brought on by aging in the skeletal muscle.

Acknowledgments

This study was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C1216).

Conflicts of Interest

KWK, TP, SK, HJ, KK, JL, YSS, and YS are inventors on patent issued by the Korean Intellectual Property Office (KR patent application No. 10-2018-0035284). All the other authors declare no conflicts of interest.

References

1. Lee K, Shin Y, Huh J, Sung YS, Lee IS, Yoon KH, et al. Recent Issues on Body Composition Imaging for Sarcopenia Evaluation. Korean J Radiol 2019 Feb;20(2):205-217 [FREE Full text] [doi: 10.3348/kjr.2018.0479] [Medline: 30672160]
2. Kim Y, Seo D, Kang J, Huh JW, Kim KW, Kim WY. Impact of Body Composition Status on 90-Day Mortality in Cancer Patients with Septic Shock: Sex Differences in the Skeletal Muscle Index. J Clin Med 2019 Oct 02;8(10) [FREE Full text] [doi: 10.3390/jcm8101583] [Medline: 31581650]
3. Park J, Gil JR, Shin Y, Won SE, Huh J, You M, et al. Reliable and robust method for abdominal muscle mass quantification using CT/MRI: An explorative study in healthy subjects. PLoS One 2019;14(9):e0222042 [FREE Full text] [doi: 10.1371/journal.pone.0222042] [Medline: 31536542]
4. Miljkovic I, Zmuda JM. Epidemiology of myosteatosis. Curr Opin Clin Nutr Metab Care 2010 May;13(3):260-264 [FREE Full text] [doi: 10.1097/MCO.0b013e328337d826] [Medline: 20179586]
5. Engelke K, Museyko O, Wang L, Laredo J. Quantitative analysis of skeletal muscle by computed tomography imaging-State of the art. J Orthop Translat 2018 Oct;15:91-103 [FREE Full text] [doi: 10.1016/j.jot.2018.10.004] [Medline: 30533385]
6. Aubrey J, Esfandiari N, Baracos VE, Buteau FA, Frenette J, Putman CT, et al. Measurement of skeletal muscle radiation attenuation and basis of its biological variation. Acta Physiol (Oxf) 2014 Mar;210(3):489-497 [FREE Full text] [doi: 10.1111/apha.12224] [Medline: 24393306]

7. Poltronieri TS, de Paula NS, Chaves GV. Assessing skeletal muscle radiodensity by computed tomography: An integrative review of the applied methodologies. Clin Physiol Funct Imaging 2020 Jul;40(4):207-223. [doi: 10.1111/cpf.12629] [Medline: 32196914]
8. Cruz-Jentoft AJ, Bahat G, Bauer J, Boirie Y, Bruyère O, Cederholm T, Writing Group for the European Working Group on Sarcopenia in Older People 2 (EWGSOP2),the Extended Group for EWGSOP2. Sarcopenia: revised European consensus on definition and diagnosis. Age Ageing 2019 Jan 01;48(1):16-31 [FREE Full text] [doi: 10.1093/ageing/afy169] [Medline: 30312372]
9. Amini B, Boyle SP, Boutin RD, Lenchik L. Approaches to Assessment of Muscle Mass and Myosteatosis on Computed Tomography: A Systematic Review. J Gerontol A Biol Sci Med Sci 2019 Sep 15;74(10):1671-1678 [FREE Full text] [doi: 10.1093/gerona/glz034] [Medline: 30726878]
10. Park HJ, Shin Y, Park J, Kim H, Lee IS, Seo DW, et al. Development and Validation of a Deep Learning System for Segmentation of Abdominal Muscle and Fat on Computed Tomography. Korean J Radiol 2020 Jan;21(1):88-100 [FREE Full text] [doi: 10.3348/kjr.2019.0470] [Medline: 31920032]
11. iAID Sarcopenia. URL: https://iaidimage.com/app/aid-u/sarcopenia-l3 [accessed 2020-09-21]
12. Schober P, Boer C, Schwarte LA. Correlation Coefficients: Appropriate Use and Interpretation. Anesth Analg 2018 May;126(5):1763-1768. [doi: 10.1213/ANE.0000000000002864] [Medline: 29481436]
13. Mullie L, Afilalo J. CoreSlicer: a web toolkit for analytic morphomics. BMC Med Imaging 2019 Feb 11;19(1):15 [FREE Full text] [doi: 10.1186/s12880-019-0316-6] [Medline: 30744586]

Abbreviations

ANOVA: analysis of variance
CT: computed tomography
DICOM: digital imaging and communications in medicine
HU: Hounsfield Unit
IMAT: inter/intramuscular adipose tissue
LAMA: low-attenuation muscle area
MRI: magnetic resonance imaging
NAMA: normal-attenuation muscle area
SMA: skeletal muscle area
TAMA: total abdominal muscle area

Edited by G Eysenbach; submitted 30.07.20; peer-reviewed by J Huh, TH Kim; comments to author 22.08.20; revised version received 07.09.20; accepted 15.09.20; published 19.10.20.

Please cite as:
Kim DW, Kim KW, Ko Y, Park T, Khang S, Jeong H, Koo K, Lee J, Kim HK, Ha J, Sung YS, Shin Y
Assessment of Myosteatosis on Computed Tomography by Automatic Generation of a Muscle Quality Map Using a Web-Based Toolkit: Feasibility Study
JMIR Med Inform 2020;8(10):e23049
URL: http://medinform.jmir.org/2020/10/e23049/
doi:10.2196/23049
PMID:33074159

Original Paper

Blood Uric Acid Prediction With Machine Learning: Model Development and Performance Comparison

Masuda Begum Sampa[1], PhD; Md Nazmul Hossain[2], PhD; Md Rakibul Hoque[3], PhD; Rafiqul Islam[4], PhD; Fumihiko Yokota[5], PhD; Mariko Nishikitani[4], MPH, PhD; Ashir Ahmed[1], PhD

[1]Department of Advanced Information Technology, Kyushu University, Fukuoka, Japan

[2]Department of Marketing, Faculty of Business Studies, University of Dhaka, Dhaka, Bangladesh

[3]School of Business, Emporia State University, Kansas, KS, United States

[4]Medical Information Center, Kyushu University Hospital, Fukuoka, Japan

[5]Institute of Decision Science for a Sustainable Society, Kyushu University, Fukuoka, Japan

Corresponding Author:
Masuda Begum Sampa, PhD
Department of Advanced Information Technology
Kyushu University
744 Motooka, Nishi-ku
Fukuoka
Japan
Phone: 81 8079893966
Email: sampa@kyudai.jp

Abstract

Background: Uric acid is associated with noncommunicable diseases such as cardiovascular diseases, chronic kidney disease, coronary artery disease, stroke, diabetes, metabolic syndrome, vascular dementia, and hypertension. Therefore, uric acid is considered to be a risk factor for the development of noncommunicable diseases. Most studies on uric acid have been performed in developed countries, and the application of machine-learning approaches in uric acid prediction in developing countries is rare. Different machine-learning algorithms will work differently on different types of data in various diseases; therefore, a different investigation is needed for different types of data to identify the most accurate algorithms. Specifically, no study has yet focused on the urban corporate population in Bangladesh, despite the high risk of developing noncommunicable diseases for this population.

Objective: The aim of this study was to develop a model for predicting blood uric acid values based on basic health checkup test results, dietary information, and sociodemographic characteristics using machine-learning algorithms. The prediction of health checkup test measurements can be very helpful to reduce health management costs.

Methods: Various machine-learning approaches were used in this study because clinical input data are not completely independent and exhibit complex interactions. Conventional statistical models have limitations to consider these complex interactions, whereas machine learning can consider all possible interactions among input data. We used boosted decision tree regression, decision forest regression, Bayesian linear regression, and linear regression to predict personalized blood uric acid based on basic health checkup test results, dietary information, and sociodemographic characteristics. We evaluated the performance of these five widely used machine-learning models using data collected from 271 employees in the Grameen Bank complex of Dhaka, Bangladesh.

Results: The mean uric acid level was 6.63 mg/dL, indicating a borderline result for the majority of the sample (normal range <7.0 mg/dL). Therefore, these individuals should be monitoring their uric acid regularly. The boosted decision tree regression model showed the best performance among the models tested based on the root mean squared error of 0.03, which is also better than that of any previously reported model.

Conclusions: A uric acid prediction model was developed based on personal characteristics, dietary information, and some basic health checkup measurements. This model will be useful for improving awareness among high-risk individuals and populations, which can help to save medical costs. A future study could include additional features (eg, work stress, daily physical activity, alcohol intake, eating red meat) in improving prediction.

(JMIR Med Inform 2020;8(10):e18331) doi:10.2196/18331

KEYWORDS

blood uric acid; urban corporate population; machine learning; noncommunicable diseases; Bangladesh; boosted decision tree regression model

Introduction

Background

Noncommunicable diseases such as cancer, diabetes, stroke, and cardiovascular diseases are the leading cause of death, disability, and morbidity worldwide. Surprisingly, the burden is particularly high in developing countries, accounting for 80% of deaths. In developing countries, 29% of noncommunicable disease–related deaths occur in the working-age population (aged <60 years) [1]. Therefore, noncommunicable diseases have become a major concern for developing countries and are also recognized as a threat for younger people [2]. Thus, reducing the incidence of noncommunicable diseases is one of the targets of sustainable development goals [3].

Uric acid is associated with several noncommunicable diseases such as cardiovascular disease and its risk factors, including chronic kidney disease, coronary artery disease, stroke, diabetes, metabolic syndrome, vascular dementia, and hypertension [4,5]. Uric acid is considered to be one of the predictors of various chronic diseases [6]. Hypertension showed positive correlations with uric acid levels among arsenic-endemic individuals in Bangladesh [7]. Another study found significant associations between uric acid and BMI, overweight, and waist circumference among the adult population of Bangladesh [8].

People working in urban areas, especially in private sectors, have significant workloads and remain seated for a long time to complete their tasks, and are thus more likely to develop noncommunicable diseases. In addition, there are few opportunities to engage in physical activities for the urban population of Bangladesh because of a lack of playgrounds, parks, walkable footpaths, and safe roads for cycling [9]. The prevalence of risk factors for developing noncommunicable diseases is also higher among urban than rural people in Bangladesh [9]. Therefore, it is important to control and prevent the severity of noncommunicable diseases by getting regular health checkups. However, most people are not interested in spending money and time on preventive health care services. Corporate people in Bangladesh lack health insurance and high health awareness, do not get routine mandatory health checkups, and are not habituated to use information and communications technology (ICT)-based health care services. Moreover, to get a checkup, they need to visit a hospital in traffic-congested areas and wait in a long, laborious queue [10].

The health status of an individual strongly depends on uric acid, which is considered to be a risk factor for the development of noncommunicable diseases [6,11]. Therefore, uric acid should be measured routinely at basic health checkups. As the reduction of noncommunicable diseases management cost is the main goal of health policies [12], studies are needed to determine blood uric acid regularly in a cost-effective manner. An accurate predictive model can help to identify a high-risk population without having to directly measure uric acid [13]. Using a prediction model designed by machine-learning approaches to test individual uric acid measurement rapidly will save costs and time of both doctors and patients.

However, to our knowledge, the application of machine-learning approaches for uric acid prediction in developing countries is very rare. In addition, different algorithms will work differently on different types of data with respect to various diseases such as different types of cancers and diabetes; therefore, separate investigations are needed for different types of data to identify the most accurate algorithms [14].

Machine-learning methods have not been practically established for clinical data from developing countries such as Bangladesh. There is also a lack of research on predicting blood uric acid based on basic clinical tests, dietary information, and sociodemographic characteristics using machine-learning approaches in Bangladesh, especially for the urban corporate population.

Therefore, the aim of the present study was to use machine-learning approaches to predict blood uric acid based on basic health checkup test results, dietary information, and sociodemographic characteristics. We tested several machine-learning approaches to evaluate the predictive power of these techniques and to best predict personalized uric acid measurement. Predicting health checkup test measurements is expected to be helpful in reducing health management costs.

Existing Related Studies

During the past few decades, the prevalence of hyperuricemia has been increasing rapidly all over the world [8]. Similar to the case of developed countries, hyperuricemia is also prevalent in developing countries [15,16]. A purine-enriched diet, obesity, and alcohol intake have been reported as the main predictors of hyperuricemia [17-19]. Approximately two-thirds of the uric acid is derived from the metabolism of endogenous purine, and the remainder is a result of eating purine-enriched foods [8,20,21]. Many previous studies identified relationships between uric acid and hypertension. For example, increasing levels of serum uric acid were associated with hypertension [4]. Serum uric acid was positively associated with incident hypertension [22] and the development of hypertension [23].

Several techniques have been proposed for the survivability analysis of various cancers [24]; however, the results of machine-learning algorithms may change due to different databases and for different measuring tools [25]. One study predicted lung cancer survival time using supervised machine-learning regression predictive techniques; although the root mean squared error (RMSE) value for each model was large (>15.30), it was unclear which predictive model would yield more predictive information for lung cancer survival time [26]. Another study also predicted hyperuricemia based on basic health checkup tests in Korea using machine-learning classification algorithms, which showed poor accuracy [6]. Targeting the prediction as a continuous target, rather than a classification into categories or levels, could help to improve

such predictions. Further, to make the prediction more accurate, it is necessary to incorporate more new features than traditionally used [27].

Most of the previous studies on uric acid have been conducted in selected White populations of North America and Europe or in entirely Black populations from South Africa [15]. Moreover, most of the previous machine learning–based research in health care has been conducted in developed countries [28]. However, there has been minimal application of supervised machine learning for medical data to predict diseases, survivability of diseases, and different types of health checkup test results using sample data from developing countries such as Bangladesh.

Study Objectives and Design

We used machine-learning approaches for development of a predictive model because clinical input data are not completely independent and complex interactions exist between them. Conventional statistical models have limitations to consider these complex interactions, whereas machine learning can consider all possible interactions among input data. Machine-learning prediction models can incorporate all of the input variables with marginal effect and variables with unknown associations with the targeted outcome variable. Machine-learning algorithms are used to identify patterns in datasets and to iteratively improve the performance of this identification with additional data [26]. Machine-learning algorithms have been extensively used in various domains such as in advertisement, agriculture, banking, online shopping, insurance, finance, social media, travel, tourism, marketing, consumer behavior, and fraud detection. These approaches are also used to analyze current and historical facts to make predictions about future events. Machine learning has also been used in the health care field for the prevention, diagnosis, and treatment phases of various diseases such as diabetes, cancer, cardiology, and mental health [29,30]. Through machine-learning prediction models, we incorporated both well-known risk factors of high uric acid such as age, BMI, and blood glucose, along with factors without clear associations to uric acid [6].

Methods

Sample

Data were collected from employees who work in the Grameen bank complex of Dhaka, Bangladesh. The Grameen bank complex comprises 18 different institutions such as Grameen Bank, Grameen Communications, other nongovernment organizations, and private companies, with more than 500 workers. We collected data from 271 employees who received human-assisted Portable Health Clinic (PHC) system services to predict blood uric acid. In general, a large sample size is required for machine-learning approaches. However, some studies have used a small sample size, including N=300 [27] and N=118 [31]. Of note, a small sample size has also been associated with higher classification accuracy [32].

Grameen Communications, Bangladesh and Kyushu University, Japan have jointly developed a human-assisted PHC system [33]. A PHC is an eHealth system that aims to provide affordable primary health care services to prevent the severity of or to control noncommunicable diseases. A PHC system has four modules: (1) a set of medical devices, (2) a software system to collect and archive medical records, (3) health care workers to make the clinical measurements and explain ePrescriptions, and (4) ICT-trained call center doctors. Consumers come to the service point and a health checkup is conducted by pretrained health care workers. If needed, the consumer is connected to the call center doctors for a consultation. The clinical measurements addressed by a PHC are as follows: (1) blood pressure; (2) pulse rate; (3) body temperature; (4) oxygenation of blood (SpO_2); (5) arrhythmia; (6) BMI; (7) waist, hip, and waist/hip ratio; (8) blood glucose; (9) blood cholesterol; (10) blood hemoglobin; (11) blood uric acid; (12) blood grouping; (13) urinary sugar; and (14) urinary protein.

These test items (except arrhythmia, blood cholesterol, blood hemoglobin, blood grouping, urinary sugar, and urinary protein because there were many missing cases in these measurements) in this PHC system were used as input factors for the present study, and uric acid measurement was set as an output factor.

Measurements

Clinical measurements were obtained through direct diagnosis using PHC instruments operated by well-trained nurses or health care professionals. Data on dietary information and sociodemographic characteristics were collected during interviews using a standard questionnaire.

Regression Predictive Modeling

As the targeted output variable of this study is a continuous variable, the regression predictive model was applied, and our objective was to predict the value of the blood uric acid of an individual. Among the multiple types of regression predictive models available, it is important to choose the best-suited models based on the type of independent and dependent variables, dimensionality in the data, and other essential characteristics of the data. We selected several algorithms that showed the best performance. Overall, no specific algorithm works best for every problem, which is especially true in the case of machine learning (ie, predictive modeling). For example, it cannot be stated that neural networks are always better than decision trees or vice versa. There are many factors at play, such as the size and structure of the dataset. Therefore, in this study, we used several machine-learning approaches, including boosted decision tree regression, decision forest regression, neural network, Bayesian linear regression, and linear regression, to predict personalized blood uric acid values based on basic health checkup test results, dietary information, and sociodemographic characteristics. We chose these five specific machine-learning algorithms because they are popular tools used to predict clinical data and they are widely used regression predictive models. These five models are also traditional machine-learning models, which perform well for regression tasks [26], and have been applied in other studies on biomedical data prediction [34].

Because a regression predictive model predicts a quantity, the performance of the model must be reported as an error in the predictions. Among the many evaluation criteria to estimate the

performance of a regression predictive model, the most common approach is to calculate the RMSE.

These five models were chosen for comparison in this study owing to their popularity in medical data prediction. Therefore, we compared these algorithms to see if the prediction accuracy can be further improved. Details of each model are described below.

Boosted Decision Tree Regression

Gradient boosting methods are a family of powerful machine-learning methods that have shown considerable success in a wide range of practical applications [35]. This model is particularly well suited for making predictions based on clinical data and exhibits high performance on clinical data [13,26,36,37]. Boosting is a popular machine-learning ensemble method [38]. Boosting means that each tree is dependent on prior trees. The algorithm learns by fitting the residual of the trees that preceded it; thus, boosting in a decision tree ensemble tends to improve accuracy with some small risk of less coverage. In the Azure Machine Learning platform, boosted decision trees use an efficient implementation of the MART gradient boosting algorithm. Gradient boosting is a machine-learning technique for regression problems. It builds each regression tree in a stepwise fashion, using a predefined loss function to measure the error in each step and correct for it in the next step. Thus, the prediction model is an ensemble of weaker prediction models. In regression problems, boosting builds a series of trees in a stepwise fashion, and then selects the optimal tree using an arbitrary differentiable loss function [39]. Similar to random forest, boosting uses many smaller, weaker models and brings them together into a final summed prediction. However, the idea of boosting is to add new models to the ensemble in a sequence for several sequences. In each iteration, a new weak model is trained with respect to the whole ensemble learned up to that new model. These new models, iteratively produced, are built to maximally correlate with the negative gradient of the loss function that is also associated with the ensemble as a whole. In this approach, a performance function is placed on the gradient boosting machine to find the point at which adding more iterations becomes negligible in benefit (ie, when adding more simple models, decision trees no longer reduce the error by a significant margin). It is at this point that the ensemble sums all of the predictions into a final overall prediction [26].

Decision Forest Regression

Decision forest or random forest has been employed in many biomedicine research applications [40-42]. In the regression problem, the decision forest output is the average value of the output of all decision trees [42-44]. Decision forests compare favorably to other techniques [45]. This regression model consists of an ensemble of decision trees. A collection of trees constitutes a forest. Each tree in a regression decision forest outputs a Gaussian distribution as a prediction. Aggregation is performed over the ensemble of trees to find a Gaussian distribution closest to the combined distribution for all trees in the model [45]. This technique generates several decision trees during training, which are allowed to split randomly from a seed point. This results in a “forest” of randomly generated decision trees whose outcomes are ensembled by the random forest algorithm to achieve more accurate prediction than possible with a single tree. One problem with a single decision tree is overfitting, making the predictions seem very good on the training data, but unreliable in future predictions [26]. By using decision forest regression, we can train a model with a relatively small number of samples and obtain good results.

Neural Network

Applying a neural network to the problem can provide much more prediction power compared to a traditional regression. Neural networks have the highest accuracy in predicting various health conditions such as heart attack and heart diseases [46,47], and have become widely used machine-learning algorithms. The neural network is a network of connected neurons. The neurons cannot operate without other neurons to which they are connected. Usually, these neurons are grouped in layers and process data in each layer, which are then passed forward to the next layers. The last layer of neurons makes decisions. The basic neural network, which is also known as multilayer perceptron, is used for comparison with one hidden layer of 500 neurons that is considered to be a reasonable number in neural network–based approaches [48].

Bayesian Linear Regression

Bayesian linear regression is the Bayesian approach to linear regression analysis. Bayesian regression methods are very powerful, as they not only provide point estimates of regression parameters but also deliver an entire distribution over these parameters. In recent years, Bayesian learning has been widely adopted and was even proven to be more powerful than other machine-learning techniques [49]. Bayesian linear regression follows a fairly natural mechanism to survive insufficient data or poorly distributed data by placing a prior on the coefficients and on the noise so that the priors can take over in the absence of data. Bayesian linear regression provides information about which parts of the model fit confidently to the data and which parts are very uncertain. The result of Bayesian linear regression is a distribution of possible model parameters based on the data and the prior. This enables quantifying the uncertainty about the model; if there are fewer data points, the posterior distribution will be more spread out.

Linear Regression

Linear regression is one of the most well-known and well-understood algorithms in statistics and machine learning. It is a fast yet simple algorithm to test, which is suitable for continuous dependent variables and can be fitted with a linear function (straight line). Linear regression models have been widely applied to predict medical data [50]. Linear regression is a very simple machine-learning method in which each data point consists of a pair of vectors: the input vector and the output vector. As the simplest, oldest, and most commonly used correlational method, linear regression fits a straight line to a set of data points using a series of coefficients multiplied to each input (ie, a weighting function) and an intercept. The weights are decided within the linear regression function in such a way that minimizes the mean error. These weight coefficients multiplied by the respective inputs, plus an intercept, give a

general function for the outcome (in this case, uric acid measurement). Thus, linear regression is easy to understand and quick to implement, even on larger datasets. The disadvantage of this method is that it is inherently linear and does not always fit real-world data [26].

Model Performance Comparison

In this study, we used five machine-learning algorithms that have been used in previous studies to predict several health conditions, including lung cancer, diabetes, heart attack, heart diseases, and breast cancer. Therefore, we considered the above five regression algorithms to be best suited for our study.

We used the Azure machine-learning platform, which is a cloud-based computing platform that allows for building, testing, and deploying predictive analytics solutions [51], to estimate the five machine-learning algorithms that are widely used to predict medical data.

For evaluating the performance of the models, RMSE values from each model were used. The RMSE of a model is the average distance between the model's prediction and the actual outcome [26], and is considered to be the prime evaluation criterion for examining the prediction performance of a continuous dependent variable through the regression predictive technique using machine-learning algorithms [34,52]. Therefore, as we are predicting the continuous value of blood uric acid, we used the regression predictive technique and evaluated the performance of models by using the RMSE. Like classification, the regression task is inductive, with the main difference being the continuous nature of the output [45].

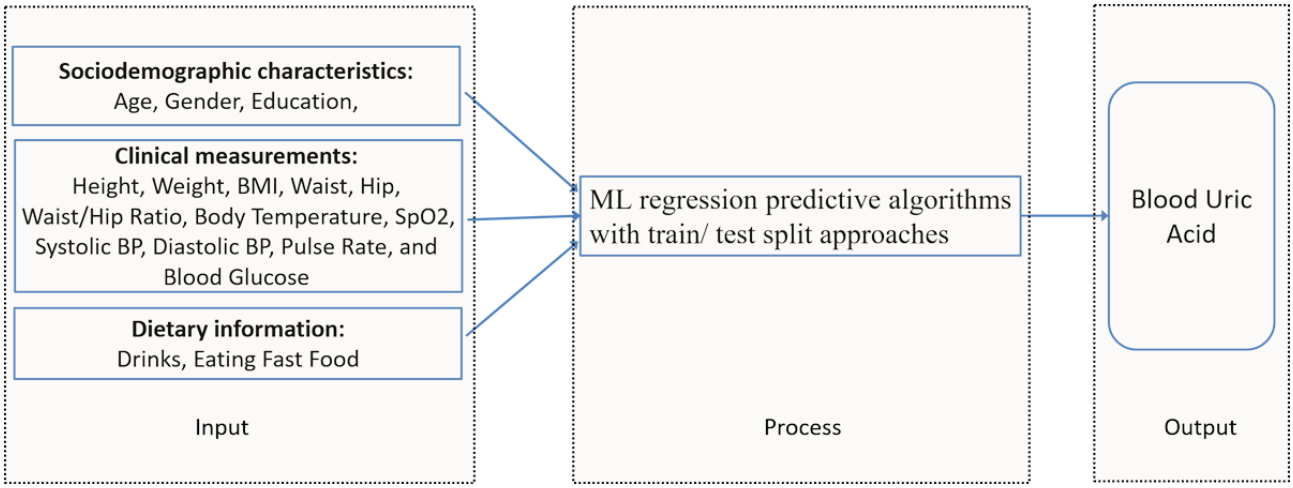
Many studies have used two validation methods to evaluate the capability of a model: the holdout method and k-fold cross-validation. According to the goal of each problem and the size of the data, different methods can be chosen to solve the problem. In the holdout method, as a popular validation method, the dataset is divided into two distinct parts: a training set and test set. The training set is used to train the machine-learning algorithm and the test set is used to evaluate the model [42,53]. The holdout method involves portioning the datasets into nonoverlapping subsets, where the first subset is entirely used for training and the rest for testing [54], and is often used instead of k-fold cross-validation [55-57]. When given no testing sample independent of the training sample, one can randomly select and hold out a portion of the training sample for testing, and construct a prediction with only the remaining sample. Typically, 30% of the training sample is set aside for testing and 70% is used for the training step [58-60].

In this study, the holdout method was used to evaluate the proposed model because it is more suitable for small sample sizes [61,62]. It is used in most of the machine-learning platforms, including the Azure machine learning studio [51] that was applied in our study. A random train-test split method is the recommended dataset split method, and machine-learning models in general yield more accurate results when trained with a greater amount of data points (70%:30%) [63]. Many previous studies also applied a 70%:30% random train-test split method in similar fields [63-65].

It is common practice to split the data into 70% as a training set and 30% as a testing set. This splitting ratio is large enough to yield statistically meaningful results. Train-test split is a simple and reliable validation approach. A portion of the data is split before any model development steps and is used only once to validate the developed model [32]. Therefore, in this study, each model was trained on a 70% training sample to ensure that each model was trained uniformly. We split the data according to a training set ratio of 0.7 and test set ratio of 0.3. We did not use the cross-validation method because k-fold cross-validation produces strongly biased performance estimates with small sample sizes [32].

The input-process-output model for predicting blood uric acid based on sociodemographic characteristics, dietary information, and some basic health checkup test results is shown in Figure 1.

Figure 1. The input-process-output model used for predicting uric acid after processing 17 input variables by machine-learning (ML) algorithms. BP: blood pressure.



Ethical Approval

We obtained ethical approval from the National Research Ethics Committee of the Bangladesh Medical Research Council (approval no. 18325022019).

Results

Characteristics of the Study Population

Data from a total of 271 employees of Grameen bank complex were collected during health checkups provided by the PHC service. The descriptive statistics of baseline characteristics of the participants are shown in Table 1.

The mean age of participants was 49.61 years. Most of the respondents had a BMI that put them in the category of overweight according to the World Health Organization criteria (range 25-29.9). The uric acid of the participants was borderline with a mean of 6.63 mg/dL, as the normal reference level is <7.0 mg/dL [11]. Therefore, the majority of the participants should be checking their uric acid regularly.

Table 1. Summary statistics of the selected continuous predictors (N=271).

Variables	Range	Mean (SD)
Age (years)	34-77	49.61 (7.39)
Height (cm)	140-184	163.05 (7.45)
Weight (kg)	44.20-114.40	67.52 (10.06)
BMI (kg/m^2)	18.39-40.53	25.37 (3.20)
Waist (cm)	63.60-118.00	90.24 (7.80)
Hip (cm)	80.00-127.00	94.54 (6.29)
Waist/hip ratio	0.64-1.11	0.96 (0.06)
Body temperature (°F)	92.12-99.64	96.07 (1.15)
Blood oxygenation (SpO_2) (%)	93-99	97.67 (1.17)
Systolic blood pressure (mmHg)	92-180	126.68 (14.88)
Diastolic blood pressure (mmHg)	59-108	81.71 (8.43)
Pulse rate (bpm)	51-123	80.27 (11.66)
Blood uric acid (mg/dL)	3.10-11.00	6.63 (1.54)
Blood glucose (mg/dL)	66.60-392.40	128.02 (56.92)

The lifestyle characteristics of the participants are summarized in Table 2. The majority of the respondents were male and had completed a college/university degree. Approximately 10% reported that they drink sugar-containing drinks 3 or more times a week and nearly 20% reported that they regularly eat fast food.

Table 2. Summary statistics of selected categorical predictors related to lifestyle factors (N=271).

Variable	n (%)
Gender	
Male	225 (83.0)
Female	46 (17.0)
Education	
No education	10 (3.7)
Primary school completed	30 (11.1)
Secondary school completed	11 (4.1)
High school completed	23 (8.5)
Vocation school completed	1 (0.4)
College/university completed	63 (23.2)
Higher education (master or doctorate degree) completed	133 (49.1)
Consumption of high-sugar drinks (eg, soda, fruit juice) ≥3 times a week	
Yes	26 (9.6)
No	245 (90.4)
Consumption of fast food such as pizza, hamburger, deep-fried foods (eg, singara, samosa, moglai parata) ≥3 times a week	
Yes	49 (18.1)
No	222 (81.9)

Prediction Performance

The RMSE was used to examine the prediction performance of the regression predictive technique with machine-learning algorithms. As shown in Table 3, the boosted decision tree regression model showed the best performance among the tested models.

Table 3. Comparison of modeling techniques ranked from best to worst based on root mean squared error (RMSE).

Model	RMSE[a]	Mean absolute error[b]	Coefficient of determination (R^2)
Boosted decision tree regression	0.03	0.01	0.99
Decision forest regression	0.75	0.53	0.75
Neural network	1.46	1.13	0.04
Bayesian linear regression	1.37	1.06	0.16
Linear regression	1.36	1.06	0.17

[a]Root mean squared error measures the average magnitude of the error by taking the square root of the average of squared differences between predicted and actual observations. That is, it measures how close the predicted value is to the actual vale. There is no cutoff or benchmark value; the smaller the value, the better the prediction.

[b]The mean absolute error is the sum of the absolute differences between predicted and actual values.

Score Model

The Score model represents the predicted value of the output or predicting variable. For regression models, the score model generates a predicted numeric value. The score model obtained using the boosted decision tree regression model is shown in Figure 2.

Figure 2. Partial view of the score model obtained by the boosted decision tree regression. Scored labels is the result column in this scoring result. The numbers are the predicted blood uric acid value for each individual.

BodyTemperature	SpO2	BloodPressuresys	BloodPressuredia	PulseRate	Blooduricacid	Bglucose	DrinksC	FastfoodC	Scored Labels
96.62	97	131	88	81	8.1	169.2	1	2	8.07499
96.26	98	126	88	91	7.5	90	1	1	7.488424
94.1	99	140	87	76	5.2	180	1	1	5.199734
96.62	99	124	82	68	6.1	163.8	1	2	6.123245
96.62	96	150	95	93	7.5	108	1	1	7.484407
92.48	99	129	84	78	7.1	151.2	1	1	7.108648
95.36	99	155	86	54	5.7	66.6	1	1	5.704509
98.24	96	129	84	105	6.3	131.4	1	1	6.29644
96.44	99	114	78	91	4.9	226.8	1	2	4.814224
95.54	97	119	71	80	6.3	117	1	1	6.296839

Discussion

Principal Findings

Machine-learning algorithms can identify the pattern in a dataset that may not be apparent directly. Thus, machine learning can provide useful information and support to medical staff by identifying patterns that may not be readily apparent [25]. There are several advantages of choosing machine-learning algorithms over conventional statistical methods for designing a prediction model. First, machine-learning algorithms can handle noisy information. Second, they can model complex, nonlinear relationships between variables without prior knowledge of a model [66], which enables including all information from the dataset during the analysis [6]. Finally, machine learning can consider all potential interactions between input variables, whereas conventional statistical analysis assumes that the input variables are independent [67]. Since many input variables are interrelated in complex ways, whether known or not, machine-learning algorithms can be used to identify high-risk individual cases and can help medical staff with clinical assessment [67].

Machine learning uses techniques that enable machines to use experience to improve at tasks. Through machine learning, data fed into an algorithm or model are used to train and test a model. The model is then deployed to conduct an automated rapid predictive task or to receive the predictions returned by the model. In many clinical studies, the gradient boosting machine-learning algorithm has been successfully used to predict cardiovascular diseases [13]. The gradient boosting decision tree method introduced by Friedman [68] predicted BMI with an accuracy of 0.91 [37]. In the current study, the boosted decision tree regression was found to be the best predictive model for uric acid, followed by decision forest regression. These are both popular ensemble learning methods.

In this study, a prediction model was designed for improving uric acid prediction by including not only well-known relevant factors of high uric acid such as age, gender, and BMI but also factors that have unknown associations with uric acid. The test items used in the PHC service were used as input factors, except for uric acid as the output factor. Therefore, a tool to predict uric acid was developed with good predictive performance based on the RMSE of 0.03; this RMSE is better than any previously reported in the literature in models related to biomedical data [26,35,69]. These results can provide useful insights for understanding the observed trend in population health and to inform future strategic decision making for improved health outcomes.

It is very important to compare the results of this study to previous related work. Most of the previous studies reported performance measurements as a function of classification accuracy, which may not be directly compared to this study with a regression approach to building a predictive model for a continuous variable (blood uric acid value).

A previous uric acid prediction study [6] that predicted uric acid levels based on health checkup data archived in a hospital in Korea used data that were collected from laboratory-quality devices in a very specific group of people who participated in an expensive, self-paid comprehensive health checkup program. The data were collected from 38,001 people, and the prediction sensitivity was 0.73 and 0.66 using naive Bayes classification and random forest classification models, respectively. They used a total of 25 variables available in their database. Our uric acid prediction model was developed using machine-learning approaches and included personal characteristics, dietary information, and basic clinical measurements. These data were collected using portable and cheap devices. Health records of 271 employees (aged 34-77 years with 83% men) were collected. We found that uric acid value can be predicted with an RMSE value of 0.03. Among the five machine-learning algorithms, boosted decision tree regression was found to be the most effective.

Contribution

This is the first study aimed at predicting laboratory test results of health measurements or health checkup items in Bangladesh. The ability to determine uric acid using the developed machine-learning prediction model would avoid the need for health care workers of PHC services to carry out uric acid

measurements. These findings can be helpful in achieving sustainable development goals and universal health coverage, and thus reducing overall morbidity and mortality. Using the prediction model designed by the machine-learning approaches to measure individual blood uric acid will save the cost and time of doctors as well as patients. This prediction model can also be applied to other institutions.

By inputting only 17 variables (12 basic clinical measurements, 3 sociodemographic characteristics, and 2 dietary characteristics) in the models, we were able to predict blood uric acid. In emergency situations such as floods, pandemics, tsunamis, and other contexts in which it is difficult to physically go to the clinic, blood uric acid can be predicted, therefore contributing to public health improvement. From the perspective of underdeveloped or developing countries such as Bangladesh, people do not check their blood uric acid frequently and do not know about the potential associated complications. However, people frequently measure the clinical variables that are included in the predictive models. By applying these machine-learning algorithms, we can also predict other health parameters such as blood glucose and SpO_2. Moreover, beyond the fields of health care and medical science, similar models can also be applied to agriculture, insurance and banking, online shopping, travel and tourism, marketing, and consumer behavior along with many other fields.

Conclusion and Prospects

This study provides a measure for reducing noncommunicable diseases, and hence can be a good component of national or global health plans. We developed a uric acid prediction model based on personal characteristics, dietary information, and some basic clinical measurements related to noncommunicable disease risk. Such a uric acid prediction model will be useful for improving awareness among high-risk individuals. The blood uric acid prediction model can further help to provide health services with the early detection and cost-effective management of noncommunicable diseases.

There are a few limitations of this study. First, the sample size was relatively small, which should be increased for training the prediction model in the future. Second, this study was limited to a particular area among a group of employees who work in a corporate setting. Our prediction model was not confirmed with data from other institutes. Although the framework achieved high performance on Grameen bank complex data, we believe that this model will also be suitable for predicting blood uric acid values in individuals that work in other types of corporate settings. Third, the included variables in the model were selected based on validated key features from previous studies rather than by using statistical approaches to identify the significant influence of factors on the output variable from the data. A future study could also include additional features (eg, work stress, everyday physical activity, eating red meat). Fourth, this study evaluated only five machine-learning algorithms among many other algorithms available. Finally, we applied only a random split method (train/test split method), although cross-validation is a good method for training and testing a dataset. We did not consider applying the cross-validation method in this case owing to the small dataset. Therefore, further study can be considered with an extended sample size and cross-validation method.

Despite these limitations, we conclude that this study represents a successful case to open discussions on further applications of this combined approach to wider regions and various types of health checkup measurements.

Acknowledgments

This research was supported by multiple organizations. Japan Society for the Promotion of Science (JSPS) KAKENHI (grant number 18K11529) and the Future Earth Research Fund (grant number 18-161009264) jointly financed the core research. The Institute of Decision Science for a Sustainable Society (IDS3), Kyushu University, Japan, provided travel expenses for data collection, and Grameen Communications, Bangladesh, provided technical assistance.

Conflicts of Interest

None declared.

References

1. Nohara Y, Kai E, Ghosh PP, Islam R, Ahmed A, Kuroda M, et al. Health checkup and telemedical intervention program for preventive medicine in developing countries: verification study. J Med Internet Res 2015 Jan 28;17(1):e2 [FREE Full text] [doi: 10.2196/jmir.3705] [Medline: 25630348]
2. Khalequzzaman M, Chiang C, Choudhury SR, Yatsuya H, Al-Mamun MA, Al-Shoaibi AAA, et al. Prevalence of non-communicable disease risk factors among poor shantytown residents in Dhaka, Bangladesh: a community-based cross-sectional survey. BMJ Open 2017 Nov 14;7(11):e014710. [doi: 10.1136/bmjopen-2016-014710] [Medline: 29138190]
3. Goal 3: ensure healthy lives and promote well-being for all at all ages. United Nations: Sustainable Development Goals. 2019. URL: https://www.un.org/sustainabledevelopment/health/ [accessed 2019-10-28]
4. Loeffler LF, Navas-Acien A, Brady TM, Miller ER, Fadrowski JJ. Uric acid level and elevated blood pressure in US adolescents: National Health and Nutrition Examination Survey, 1999-2006. Hypertension 2012 Apr;59(4):811-817 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.111.183244] [Medline: 22353609]

5. Feig DI, Kang D, Johnson RJ. Uric acid and cardiovascular risk. N Engl J Med 2008 Oct 23;359(17):1811-1821 [FREE Full text] [doi: 10.1056/NEJMra0800885] [Medline: 18946066]
6. Lee S, Choe E, Park B. Exploration of Machine Learning for Hyperuricemia Prediction Models Based on Basic Health Checkup Tests. J Clin Med 2019 Feb 02;8(2):172 [FREE Full text] [doi: 10.3390/jcm8020172] [Medline: 30717373]
7. Huda N, Hossain S, Rahman M, Karim MR, Islam K, Mamun AA, et al. Elevated levels of plasma uric acid and its relation to hypertension in arsenic-endemic human individuals in Bangladesh. Toxicol Appl Pharmacol 2014 Nov 15;281(1):11-18. [doi: 10.1016/j.taap.2014.09.011] [Medline: 25281834]
8. Ali N, Perveen R, Rahman S, Mahmood S, Rahman S, Islam S, et al. Prevalence of hyperuricemia and the relationship between serum uric acid and obesity: A study on Bangladeshi adults. PLoS One 2018 Nov 1;13(11):e0206850 [FREE Full text] [doi: 10.1371/journal.pone.0206850] [Medline: 30383816]
9. Zaman M, Rahman MM, Rahman MR, Bhuiyan M, Karim MN, Chowdhury MA. Prevalence of risk factors for non-communicable diseases in Bangladesh: Results from STEPS survey 2010. Indian J Public Health 2016;60(1):17. [doi: 10.4103/0019-557x.177290]
10. Sampa MB, Hossain MN, Hoque MR, Islam R, Yokota F, Nishikitani M, et al. Influence of Factors on the Adoption and Use of ICT-Based eHealth Technology by Urban Corporate People. JSSM 2020;13(01):1-19. [doi: 10.4236/jssm.2020.131001]
11. Kim S, Chang Y, Yun KE, Jung H, Lee S, Shin H, et al. Development of Nephrolithiasis in Asymptomatic Hyperuricemia: A Cohort Study. Am J Kidney Dis 2017 Aug;70(2):173-181. [doi: 10.1053/j.ajkd.2017.01.053] [Medline: 28410765]
12. Hunter DJ, Reddy KS. Noncommunicable diseases. N Engl J Med 2013 Oct 03;369(14):1336-1343. [doi: 10.1056/NEJMra1109345] [Medline: 24088093]
13. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevska O, written on behalf of AME Big-Data Clinical Trial Collaborative Group. Predictive analytics with gradient boosting in clinical medicine. Ann Transl Med 2019 Apr;7(7):152-152. [doi: 10.21037/atm.2019.03.29] [Medline: 31157273]
14. Noohi NA, Ahmadzadeh M, Fardaei M. Medical Data Mining and Predictive Model for Colon Cancer Survivability. Int J Innov Res Eng Sci 2013;2(2).
15. Conen D, Wietlisbach V, Bovet P, Shamlaye C, Riesen W, Paccaud F, et al. Prevalence of hyperuricemia and relation of serum uric acid with cardiovascular risk factors in a developing country. BMC Public Health 2004 Mar 25;4(1):9 [FREE Full text] [doi: 10.1186/1471-2458-4-9] [Medline: 15043756]
16. Chen L, Zhu W, Chen Z, Dai H, Ren J, Chen J, et al. Relationship between hyperuricemia and metabolic syndrome. J Zhejiang Univ Sci B 2007 Jul;8(8):593-598. [doi: 10.1631/jzus.2007.b0593]
17. Nakanishi N, Yoshida H, Nakamura K, Suzuki K, Tatara K. Predictors for development of hyperuricemia: an 8-year longitudinal study in middle-aged Japanese men. Metabolism 2001 Jun;50(6):621-626. [doi: 10.1053/meta.2001.24196] [Medline: 11398134]
18. Wortmann RL. Gout and hyperuricemia. Curr Opin Rheumatol 2002 May;14(3):281-286. [doi: 10.1097/00002281-200205000-00015] [Medline: 11981327]
19. Ogura T, Matsuura K, Matsumoto Y, Mimura Y, Kishida M, Otsuka F, et al. Recent trends of hyperuricemia and obesity in Japanese male adolescents, 1991 through 2002. Metabolism 2004 Apr;53(4):448-453. [doi: 10.1016/j.metabol.2003.11.017] [Medline: 15045690]
20. Schlesinger N. Dietary factors and hyperuricaemia. Curr Pharm Des 2005 Dec 01;11(32):4133-4138. [doi: 10.2174/138161205774913273] [Medline: 16375734]
21. Miao Z, Yan S, Wang J, Wang B, Li Y, Xing X, et al. Insulin resistance acts as an independent risk factor exacerbating high-purine diet induced renal injury and knee joint gouty lesions. Inflamm Res 2009 Oct 31;58(10):659-668. [doi: 10.1007/s00011-009-0031-9] [Medline: 19333726]
22. Mellen PB, Bleyer AJ, Erlinger TP, Evans GW, Nieto FJ, Wagenknecht LE, et al. Serum uric acid predicts incident hypertension in a biethnic cohort: the atherosclerosis risk in communities study. Hypertension 2006 Dec;48(6):1037-1042. [doi: 10.1161/01.HYP.0000249768.26560.66] [Medline: 17060502]
23. Perlstein TS, Gumieniak O, Williams GH, Sparrow D, Vokonas PS, Gaziano M, et al. Uric acid and the development of hypertension: the normative aging study. Hypertension 2006 Dec;48(6):1031-1036. [doi: 10.1161/01.HYP.0000248752.08807.4c] [Medline: 17060508]
24. Agrawal A, Misra S, Narayanan R, Polepeddi L, Choudhary A. Lung Cancer Survival Prediction using Ensemble Data Mining on Seer Data. Sci Program 2012;20(1):29-42. [doi: 10.1155/2012/920245]
25. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005 Jun;34(2):113-127. [doi: 10.1016/j.artmed.2004.07.002] [Medline: 15894176]
26. Lynch CM, Abdollahi B, Fuqua JD, de Carlo AR, Bartholomai JA, Balgemann RN, et al. Prediction of lung cancer patient survival via supervised machine learning classification techniques. Int J Med Inform 2017 Dec;108:1-8 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.09.013] [Medline: 29132615]
27. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform 2017 Jan;97:120-127 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.09.014] [Medline: 27919371]

28. Gu D, Li J, Li X, Liang C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. Int J Med Inform 2017 Feb;98:22-32. [doi: 10.1016/j.ijmedinf.2016.11.006] [Medline: 28034409]
29. Triantafyllidis AK, Tsanas A. Applications of Machine Learning in Real-Life Digital Health Interventions: Review of the Literature. J Med Internet Res 2019 Apr 05;21(4):e12286 [FREE Full text] [doi: 10.2196/12286] [Medline: 30950797]
30. Misawa D, Fukuyoshi J, Sengoku S. Cancer Prevention Using Machine Learning, Nudge Theory and Social Impact Bond. Int J Environ Res Public Health 2020 Jan 28;17(3):790 [FREE Full text] [doi: 10.3390/ijerph17030790] [Medline: 32012838]
31. Zelic I, Kononenko I, Lavrac N, Vuga V. Induction of decision trees and Bayesian classification applied to diagnosis of sport injuries. J Med Syst 1997 Dec;21(6):429-444. [doi: 10.1023/a:1022880431298] [Medline: 9555629]
32. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One 2019 Nov 7;14(11):e0224365 [FREE Full text] [doi: 10.1371/journal.pone.0224365] [Medline: 31697686]
33. Sampa MB, Hossain N, Hoque R, Islam R, Yokota F, Nishikitani M, et al. A Framework of Longitudinal Study to Understand Determinants of Actual Use of the Portable Health Clinic System. In: Streitz N, Konomi S, editors. Distributed, Ambient and Pervasive Interactions. HCII 2019. Lecture Notes in Computer Science, vol 11587. Cham: Springer; 2019:323-332.
34. 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]
35. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7:21. [doi: 10.3389/fnbot.2013.00021] [Medline: 24409142]
36. Afzal M, Hussain M, Malik KM, Lee S. Impact of Automatic Query Generation and Quality Recognition Using Deep Learning to Curate Evidence From Biomedical Literature: Empirical Study. JMIR Med Inform 2019 Dec 09;7(4):e13430 [FREE Full text] [doi: 10.2196/13430] [Medline: 31815673]
37. Hu M, Nohara Y, Wakata Y, Ahmed A, Nakashima N, Nakamura M. Machine Learning Based Prediction of Non-communicable Diseases to Improving Intervention Program in Bangladesh. Eur J Bioinformatics 2018;14(4):20-28. [doi: 10.24105/ejbi.2018.14.4.5]
38. Wu J, Roy J, Stewart WF. Prediction Modeling Using EHR Data. Medical Care 2010;48:S106-S113. [doi: 10.1097/mlr.0b013e3181de9e17]
39. Manna S, Biswas S, Kundu R, Rakshit S, Gupta P, Barman S. A statistical approach to predict flight delay using gradient boosted decision tree. : IEEE; 2017 Presented at: International Conference on Computational Intelligence in Data Science (ICCIDS); June 2-3, 2017; Chennai, India. [doi: 10.1109/iccids.2017.8272656]
40. Zhao X, Zou Q, Liu B, Liu X. Exploratory Predicting Protein Folding Model with Random Forest and Hybrid Features. Curr Proteomics 2015 Jan 21;11(4):289-299. [doi: 10.2174/1570164611104150121115154]
41. Liao Z, Ju Y, Zou Q. Prediction of G Protein-Coupled Receptors with SVM-Prot Features and Random Forest. Scientifica (Cairo) 2016;2016:8309253-8309210. [doi: 10.1155/2016/8309253] [Medline: 27529053]
42. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes Mellitus With Machine Learning Techniques. Front Genet 2018 Nov 6;9:515. [doi: 10.3389/fgene.2018.00515] [Medline: 30459809]
43. Liaw A, Wiener M. Classification and Regression by RandomForest. R News 2002;2(3):18-22.
44. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 2003 Nov;43(6):1947-1958. [doi: 10.1021/ci034160g] [Medline: 14632445]
45. Criminisi A, Shotton J, Konukoglu E. Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. In: Foundations and Trends in Computer Graphics and Vision. Boston, MA: NOW Publishers; 2012:81-227.
46. Yahyaie M, Tarokh MJ, Mahmoodyar MA. Use of Internet of Things to Provide a New Model for Remote Heart Attack Prediction. Telemed J E Health 2019 Jun;25(6):499-510. [doi: 10.1089/tmj.2018.0076] [Medline: 30256729]
47. Dangare CS, Apte SS. A data mining approach for prediction of heart disease using neural networks. Int J Comput Eng Technol 2012;3(3):30-40.
48. Li X, Ding Q, Sun J. Remaining useful life estimation in prognostics using deep convolution neural networks. Reliab Eng Syst Saf 2018 Apr;172:1-11. [doi: 10.1016/j.ress.2017.11.021]
49. Jihan N. Bayesian Learning for Machine Learning: Linear Regression (Part 2). DZone. 2019 May 09. URL: https://dzone.com/articles/bayesian-learning-for-machine-learning-part-ii-lin [accessed 2020-09-24]
50. Zarkogianni K, Mitsis K, Litsa E, Arredondo M, Fico G, Fioravanti A, et al. Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring. Med Biol Eng Comput 2015 Dec 7;53(12):1333-1343. [doi: 10.1007/s11517-015-1320-9] [Medline: 26049412]
51. Barga R, Fontama V, Tok WH. Predictive Analytics with Microsoft Azure Machine Learning. New York: Apress; 2015:21-43.
52. Cox DR. Two further applications of a model for binary regression. Biometrika 1958;45(3-4):562-565 [FREE Full text] [doi: 10.1093/biomet/45.3-4.562]
53. Kim J. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Comput Stat Data Analysis 2009 Sep;53(11):3735-3745. [doi: 10.1016/j.csda.2009.04.009]

54. Yadav S, Shukla S. Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. New York: Institute of Electrical and Electronics Engineers Inc; 2016 Presented at: Proceedings - 6th International Advanced Computing Conference, IACC 2016; February 27-28, 2016; Bhimavaram, India p. 78-83. [doi: 10.1109/iacc.2016.25]
55. Zarkogianni K, Litsa E, Vazeou A, Nikita KS. Personalized glucose-insulin metabolism model based on self-organizing maps for patients with Type 1 Diabetes Mellitus. Chania; 2013 Presented at: 13th IEEE International Conference on BioInformatics and BioEngineering, IEEE BIBE; 2013; Greece. [doi: 10.1109/bibe.2013.6701604]
56. Ruiz-Velázquez E, Alanis AY, Femat R, Quiroz G. Neural modeling of the blood glucose level for type 1 diabetes mellitus patients. : IEEE; 2011 Presented at: 2011 IEEE International Conference on Automation Science and Engineering; August 24-27, 2011; Trieste, Italy. [doi: 10.1109/CASE.2011.6042485]
57. Mirshekarian S, Bunescu R, Marling C, Schwartz F. Using LSTMs to learn physiological models of blood glucose behavior. New York: IEEE; 2017 Presented at: 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS; 2017; Seogwipo, South Korea p. 2887-2891. [doi: 10.1109/embc.2017.8037460]
58. Ben Ali J, Hamdi T, Fnaiech N, Di Costanzo V, Fnaiech F, Ginoux J. Continuous blood glucose level prediction of Type 1 Diabetes based on Artificial Neural Network. Biocybern Biomed Eng 2018;38(4):828-840. [doi: 10.1016/j.bbe.2018.06.005]
59. Hamdi T, Ben Ali J, Di Costanzo V, Fnaiech F, Moreau E, Ginoux J. Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm. Biocybern Biomed Eng 2018;38(2):362-372. [doi: 10.1016/j.bbe.2018.02.005]
60. Li J, Xu Q, Shah N, Mackey TK. A Machine Learning Approach for the Detection and Characterization of Illicit Drug Dealers on Instagram: Model Evaluation Study. J Med Internet Res 2019 Jun 15;21(6):e13803 [FREE Full text] [doi: 10.2196/13803] [Medline: 31199298]
61. Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage 2018 Oct 15;180(Pt A):68-77. [doi: 10.1016/j.neuroimage.2017.06.061] [Medline: 28655633]
62. Isaksson A, Wallman M, Göransson H, Gustafsson M. Cross-validation and bootstrapping are unreliable in small sample classification. Pattern Recogn Lett 2008 Oct;29(14):1960-1965. [doi: 10.1016/j.patrec.2008.06.018]
63. Livera A, Theristis M, Makrides G, Ransome S, Sutterlueti J, Georghiou GE. Optimal development of location and technology independent machine learning photovoltaic performance predictive models. : IEEE; 2019 Presented at: 46th IEEE Photovoltaic Specialists Conference (IEEE PVSC); June 16-21, 2019; Chicago, IL. [doi: 10.1109/pvsc40753.2019.8980474]
64. Polat K, Akdemir B, Güneş S. Computer aided diagnosis of ECG data on the least square support vector machine. Dig Sign Process 2008 Jan;18(1):25-32. [doi: 10.1016/j.dsp.2007.05.006]
65. Soman T, Bobbie PO. Classification of arrhythmia using machine learning techniques. 2005 Presented at: 4th International Conference on System Science and Engineering (ICOSSE); April 25-27, 2005; Rio de Janeiro, Brazil.
66. Perai A, Nassiri Moghaddam H, Asadpour S, Bahrampour J, Mansoori G. A comparison of artificial neural networks with other statistical approaches for the prediction of true metabolizable energy of meat and bone meal. Poult Sci 2010 Jul;89(7):1562-1568 [FREE Full text] [doi: 10.3382/ps.2010-00639] [Medline: 20548088]
67. Singal AG, Mukherjee A, Elmunzer JB, Higgins PDR, Lok AS, Zhu J, et al. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. Am J Gastroenterol 2013 Nov;108(11):1723-1730 [FREE Full text] [doi: 10.1038/ajg.2013.332] [Medline: 24169273]
68. Friedman JH. Greedy function approximation: A gradient boosting machine. Annal Stat 2001;29(5):1189-1232. [doi: 10.1007/978-1-4842-3564-5_6]
69. Luo W, Nguyen T, Nichols M, Tran T, Rana S, Gupta S, et al. Is demography destiny? Application of machine learning techniques to accurately predict population health outcomes from a minimal demographic dataset. PLoS One 2015;10(5):e0125602 [FREE Full text] [doi: 10.1371/journal.pone.0125602] [Medline: 25938675]

Abbreviations

ICT: information and communications technology
PHC: Portable Health Clinic
RMSE: root mean squared error
SpO_2: blood oxygen level

Edited by G Eysenbach; submitted 20.02.20; peer-reviewed by D Carvalho, SK Lee; comments to author 28.05.20; revised version received 16.07.20; accepted 10.08.20; published 08.10.20.

Please cite as:
Sampa MB, Hossain MN, Hoque MR, Islam R, Yokota F, Nishikitani M, Ahmed A
Blood Uric Acid Prediction With Machine Learning: Model Development and Performance Comparison
JMIR Med Inform 2020;8(10):e18331
URL: https://medinform.jmir.org/2020/10/e18331
doi:10.2196/18331
PMID:33030442

Original Paper

AutoScore: A Machine Learning–Based Automatic Clinical Score Generator and Its Application to Mortality Prediction Using Electronic Health Records

Feng Xie[1], BSc; Bibhas Chakraborty[1,2,3], PhD; Marcus Eng Hock Ong[1,4,5], MBBS, MPH; Benjamin Alan Goldstein[1,3], PhD; Nan Liu[1,5,6], PhD

[1]Programme in Health Services and Systems Research, Duke-NUS Medical School, Singapore, Singapore
[2]Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore
[3]Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, United States
[4]Department of Emergency Medicine, Singapore General Hospital, Singapore, Singapore
[5]Health Services Research Centre, Singapore Health Services, Singapore, Singapore
[6]Institute of Data Science, National University of Singapore, Singapore, Singapore

Corresponding Author:
Nan Liu, PhD
Programme in Health Services and Systems Research
Duke-NUS Medical School
8 College Road
Singapore, 169857
Singapore
Phone: 65 65767372
Email: liu.nan@duke-nus.edu.sg

Abstract

Background: Risk scores can be useful in clinical risk stratification and accurate allocations of medical resources, helping health providers improve patient care. Point-based scores are more understandable and explainable than other complex models and are now widely used in clinical decision making. However, the development of the risk scoring model is nontrivial and has not yet been systematically presented, with few studies investigating methods of clinical score generation using electronic health records.

Objective: This study aims to propose AutoScore, a machine learning–based automatic clinical score generator consisting of 6 modules for developing interpretable point-based scores. Future users can employ the AutoScore framework to create clinical scores effortlessly in various clinical applications.

Methods: We proposed the AutoScore framework comprising 6 modules that included variable ranking, variable transformation, score derivation, model selection, score fine-tuning, and model evaluation. To demonstrate the performance of AutoScore, we used data from the Beth Israel Deaconess Medical Center to build a scoring model for mortality prediction and then compared the data with other baseline models using the receiver operating characteristic analysis. A software package in R 3.5.3 (R Foundation) was also developed to demonstrate the implementation of AutoScore.

Results: Implemented on the data set with 44,918 individual admission episodes of intensive care, the AutoScore-created scoring models performed comparably well as other standard methods (ie, logistic regression, stepwise regression, least absolute shrinkage and selection operator, and random forest) in terms of predictive accuracy and model calibration but required fewer predictors and presented high interpretability and accessibility. The nine-variable, AutoScore-created, point-based scoring model achieved an area under the curve (AUC) of 0.780 (95% CI 0.764-0.798), whereas the model of logistic regression with 24 variables had an AUC of 0.778 (95% CI 0.760-0.795). Moreover, the AutoScore framework also drives the clinical research continuum and automation with its integration of all necessary modules.

Conclusions: We developed an easy-to-use, machine learning–based automatic clinical score generator, AutoScore; systematically presented its structure; and demonstrated its superiority (predictive performance and interpretability) over other conventional methods using a benchmark database. AutoScore will emerge as a potential scoring tool in various medical applications.

(JMIR Med Inform 2020;8(10):e21798) doi:10.2196/21798

KEYWORDS

clinical decision making; machine learning; prognosis; clinical prediction rule; electronic health records

Introduction

Risk-scoring models are sparse models with integer point scores, which are used pervasively throughout medicine for risk stratification [1]. Risk-scoring models have been developed to determine which patients are at most risk of adverse events or worsening health conditions. Accurate identification of patients at risk can be useful for appropriate allocations of medical resources [2-4]. Risk-scoring models have been traditionally developed in 1 of 2 ways: through expert opinions or consensus, such as the Sepsis-related Organ Failure Assessment [5] score and the National Early Warning Score [6], and through the analysis of conventional cohort studies, such as the History, Electrocardiogram, Age, Risk factors, and Troponin score [7] and the Charlson Comorbidity Index [8]. Both approaches are labor-intensive and are not easy to update over time, which reveals the need for a flexible and fast approach to deriving risk-scoring models.

At present, the increasing popularity of electronic health records (EHRs) [9] creates an opportunity to take advantage of its growing quantity and diversity of data for creating novel risk models with both domain expert–curated approaches and advanced machine learning solutions. Although EHRs are rich data sources, numerous data items are collected in a nonsystematic way related to clinical use, leading to a bevy of irrelevant and redundant information. Therefore, variable selection, the process of determining a subset of relevant and discriminative variables for model development [10], plays an essential role in the development of a risk model. In risk models, more variables do not necessarily lead to better performance [11]. Moreover, irrelevant and redundant information can adversely affect model interpretability and accessibility, especially in the clinical context. A typical but time-intensive approach for variable selection uses domain knowledge obtained from literature reviews and consultation with experts; however, the literature may not always be available, and the expert's interpretation could be biased. Analytic approaches exist, such as stepwise methods (eg, forward and backward) and regularization (eg, the least absolute shrinkage and selection operator [LASSO]). However, when data sets are large enough, these methods do not often achieve a sparse solution. Thus, there is an unmet need to develop a parsimonious model with easy access to validation in the context of EHRs.

Model complexity not only affects model efficiency but also impacts transparency and interpretability [12] in clinical practice. Although machine learning often has greater predictive accuracy than simpler models, it has 2 key shortcomings. First, machine learning is harder to implement in real-world settings where many EHR systems can only accept regression or point-based approaches [13,14]. Second, it has lower explainability due to its black box nature. Clinicians may not accept black box models due to various reasons such as lack of external validation and the involvement of complex mathematical computation. Sullivan et al [15] suggested that the multivariable mathematical models are relatively complex, and the calculation should be simplified to allow application of models even without a computer, making these complex statistical models useful to clinical practitioners. Churpek et al [4] also suggested that a simple and parsimonious model can be applied at the bedside and easily validated across different hospitals. Thus, point-based scoring models are more favored in the medical context and are still widely used in clinical decision making. However, as developing a scoring model is nontrivial, there is a need to automate the process of score generation to cater to the increasingly diversified patient population and large-scale EHRs.

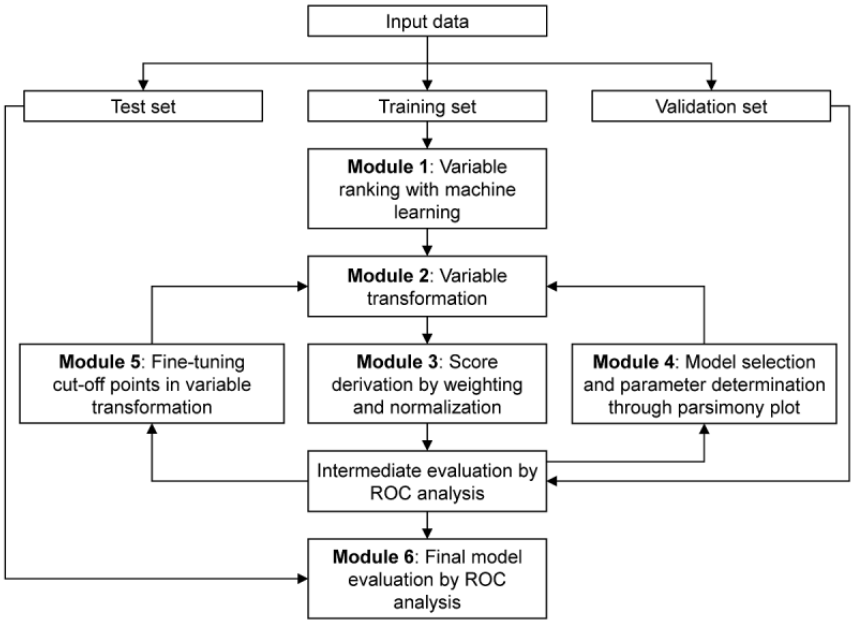
To tackle these problems and systematically present a robust and generic method for developing risk-scoring models, we proposed AutoScore, an automatic clinical score generator, by combining machine learning and regression modeling. The proposed AutoScore framework can automatically generate parsimonious sparse-score risk models (ie, risk scores), which can be easily implemented and validated in clinical practice. In this study, we implemented our proposed AutoScore framework to build an actual risk-scoring model for inpatient mortality prediction.

Methods

AutoScore for Automatic Score Generation

In this paper, we proposed the AutoScore, a novel framework for automating the development of a clinical scoring model for predefined outcomes and systematically presented its structure. AutoScore consists of 6 modules: variable ranking with machine learning, variable transformation, score derivation, model selection, domain knowledge–based score fine-tuning, and model evaluation. In our demonstration, the full data set was randomly split into a nonoverlapping training set (70%), validation set (10%; if downstream parameter tuning is needed), and test set (20%). The training set was used to derive the scores. The validation set was used for intermediate performance evaluation and parameter selection, which were elaborated in Module 4. The test set acted as an unseen data set and was used to generate the metrics of final model performance in Module 6. In real-world clinical applications, users can set up training, validation, and test sets accordingly instead of random splitting. Figure 1 illustrates the framework of AutoScore, and details of its 6 modules are elaborated as follows.

Figure 1. Flowchart of the AutoScore framework. ROC: receiver operating characteristic.



Module 1: Variable Ranking With Machine Learning

The first step in the AutoScore framework is variable ranking. We use random forest (RF) [16,17], an ensemble machine learning algorithm, to identify the top-ranking predictors for subsequent score generation. RF consists of multiple tree-structured classifiers (decision trees). Each of the trees is grown using a classification and regression tree [18] to maximum size, without pruning, and trained on a bootstrap sample and a random subset of all variables. Each tree sees only a subset of variables and part of the observations by resampling, which guarantees that the trees are decorrelated and, therefore, less prone to overfitting [19]. For the classification task, the Gini index is used to determine the optimal split. For each node [redacted] of a decision tree [redacted], the Gini index can be defined as follows:

[redacted]

where p_r refers to the fraction of training samples from the r^{th} class in the node [redacted] and $R=2$ in binary classification. In addition to outcome prediction, RF ranks variables on the basis of their predictive importance [20]. The mean decrease impurity is the measurement of variable importance, calculated by the total decrease in node impurities from splitting on the variable. The importance measurement of a variable X_m is the weighted total of impurity decreases $w([redacted])\Delta Gini([redacted])$ for all nodes [redacted], averaged over all trees [21]:

[redacted]

Where $w([redacted])$ is the proportional weight $N[redacted]/N$ of samples reaching node [redacted], v([redacted]) is the variable in the split of the node [redacted], $\Delta Gini([redacted])$ is the total impurity decrease after the split of the node [redacted]; and $N[redacted]$ is the number of decision trees in the RF model. Then, $Imp(X_m)$ will be used for variable ranking for each X_m.

An advantage of using RF as the variable ranker over other methods such as backward stepwise regression or LASSO is that as a nonparametric model, RF is able to rank variables on the basis of their nonlinear and heterogeneous effects. In the AutoScore framework, the final list of variables is decided by the ranking, in addition to the parameter m, which is the number of final selected variables. Parameter m can be chosen case by case in accordance with clinical preference, expert knowledge, or the needs of real-world applications. Moreover, an optimized number of variables can be determined through grid search and performance validation, which will be elaborated in Module 4.

Module 2: Variable Transformation

After variable selection, all selected variables are preprocessed for variable transformation, that is, continuous variables are converted into categorical variables. Creating categorical variables allows for the modeling of nonlinear effects. In AutoScore, the maximum number of categories (eg, K=5) for each variable is predefined to ensure its usability. For a categorical variable, if the original number of categories (L) exceeds the predefined maximum number (ie, $L>K$), several excess categories need to be combined, and K' is the number of categories of the transformed variable where $K' \leq L$. Unlike categorical variables, to develop a point-based score, continuous variables will be stratified by specific quantiles into K categories (in our study, K=5). We set the quantiles as 0%, k_1%, k_2%, k_3%, k_4%, and 100%. The values of k_1, k_2, k_3, and k_4 can be set in accordance with the distributions of the variables. In our study, we set the default values as follows: k_1=5, k_2=20, k_3=80, and k_4=95, which were appropriate for most variables (such as common vital signs and laboratory test results), especially those with normal or near-normal distributions.

Module 3: Score Derivation by Weighting and Normalization

With the selected and transformed variables, we created a risk score to predict the outcome, in which each category of an individual variable is weighted and given an integer point. As the default setting, we used logistic regression for score weighting, with which the points can be easily interpreted.

Where β_0 is the intercept, $\beta_1 \ldots \beta_m$ are the coefficients for each category, $X_1 \ldots X_m$ are the predictive variables, and Y is the binary outcome.

Multivariable logistic regression is performed to determine regression coefficients. On the basis of the results, the category of each variable with the lowest β coefficient is set as the reference. Next, multivariable logistic regression is performed again with adjusted reference categories to ensure that there are no negative coefficients. Subsequently, all coefficients β obtained from the second-round logistic regression are divided by the lowest β of all variables to ensure that all of the points are larger than one, that is, $\beta_{new} = \beta/\beta_{lowest}$. The final weighted points for each category were rounded as $\beta_{score} = round(\beta_{new})$. With β_{score}, we can obtain a scoring table where each category of a variable is given certain points. The total score is computed by summing up all points. To satisfy the need for specific clinical applications, we can set the ceiling value for the total score and normalize the score breakdowns, divided by a common denominator.

Module 4: Model Selection and Parameter Determination

The number of variables (*m*) is a critical parameter for controlling model complexity in the scoring model. A model is considered parsimonious when it is both sparse (using the least number of variables possible) and possesses a good prediction accuracy. To cope with the trade-off between accuracy and complexity, different parameter *m* will be examined on the validation set and a parsimony plot (ie, model performance vs complexity) will be plotted, to which the user can refer for deciding the trade-off in deriving the risk scores. The best parameter *m* is determined when *m* continues to increment and the prediction performance is no longer improving significantly, as shown in the parsimony plot. After confirming the parameter *m*, the final list of variables will be determined on the basis of the ranking obtained from Module 1. Modules 2 and 3 will be reimplemented to generate the initial scoring model.

Module 5: Fine-Tuning Cutoff Points in the Variable Transformation

Domain knowledge is essential in guiding risk model development. For continuous variables, the variable transformation (Module 2) is a data-driven process, in which domain knowledge is not integrated. In this module, the automatically generated cutoff values for each continuous variable can be fine-tuned by combining, rounding, and adjusting according to the standard clinical norm. The fine-tuning process endows the final risk scores with orderliness, professionality, and acceptability. After adjusting the cutoffs to convert continuous variables into categorical variables, Modules 2 and 3 will be implemented again to create an updated score table.

Module 6: Predictive Performance Evaluation

The performance of the score is evaluated on the basis of the receiver operating characteristic (ROC) analysis. The intermediate evaluation based on the validation set provides information for model optimization (eg, Modules 4 and 5). For the final model evaluation based on the unseen test set, the area under the ROC curve (AUC) acts as the primary metric. In addition, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) are calculated under the optimal cutoffs, defined as the points nearest to the upper-left corner of the ROC curves. Performance metrics under different cutoffs are also compared to evaluate the predictive performance. In the demonstration, we included cutoffs, by which the sensitivity or specificity could reach about 95% to satisfy certain needs in clinical settings.

Software Package

We have introduced all the 6 major modules of the AutoScore framework, with which clinical risk scores can be developed using specific patient cohorts and outcomes. We further created the AutoScore software suite [22] (Multimedia Appendices 1 and 2) under the R 3.5.3 (R Foundation) programming environment to demonstrate its capability and to facilitate its implementation and validation in other applications. Given a new data set, the AutoScore tool can be conveniently implemented to generate a point-based clinical scoring model to predict the outcome, with the minimum manual processes for data processing, parameter tuning, and model fine-tuning.

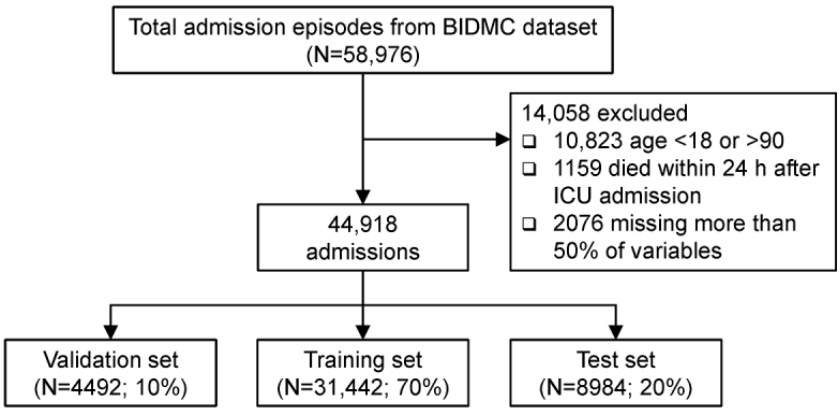
Clinical Study Design

We conducted a retrospective analysis of data from the Beth Israel Deaconess Medical Center (BIDMC) to demonstrate the usability of our proposed AutoScore framework. BIDMC is a teaching hospital at the Harvard Medical School in Boston. It has 673 inpatient beds and receives about 55,000 emergency department visits annually. We aimed to implement AutoScore to automatically generate point-based scores for risk prediction of inpatient mortality and compared AutoScore-created scoring models with several baseline models.

Data Collection and Cohort

The BIDMC data set was obtained from the Medical Information Mart for Intensive Care III [23] database compiled by the Massachusetts Institute of Technology Laboratory for Computational Physiology. A total of 58,976 BIDMC admission encounters from 2001 to 2012 were recorded in this database. All inpatient encounters for which the patient aged 18 to 90 years were included in our study cohort. The admission episodes during which patients died within 24 hours after the intensive care unit (ICU) admission or missed more than 50% of the features were excluded. A flowchart of cohort formation is shown in Figure 2.

Figure 2. Flowchart of the study cohort formation. BIDMC: Beth Israel Deaconess Medical Center.



Variables and Clinical Outcome

The primary outcome in this study was inpatient mortality, defined as deaths that occurred during the hospital stay. In the BIDMC data set, we extracted patients' first-day variables during their ICU stay. We previously demonstrated that demographic features, vital signs, and laboratory tests were highly related to inpatient mortality [24]. Similar results were also reported in other studies [25]. Thus, the predictor variables included age, sex, race, type of insurance, heart rate (beats/min), respiration rate (breaths/min), peripheral capillary oxygen saturation (SpO_2; %), diastolic blood pressure (mm Hg), systolic blood pressure (mm Hg), mean arterial pressure (MAP; mm Hg), temperature (°C), bicarbonate (mmol/L), creatinine (μmol/L), potassium (mmol/L), sodium (mmol/L), hemoglobin (g/dL), glucose (mg/dL), blood urea nitrogen (BUN; mg/dL), platelet (thousand per microliter), lactate (mmol/L), anion gap (mEq/L), hematocrit (%), chloride (mEq/L), and white blood cells (thousand per microliter). As there were multiple sets of vital signs or laboratory data collected in the ICU, the mean values were used in this study.

Baseline Models Versus AutoScore

To evaluate the performance of AutoScore, we compared it with several standard predictive models. The first model was built with logistic regression by using all available variables from the training data set without variable selection. The second model was built using stepwise multivariable logistic regression [26]. It built a regression model with variable selection using the Akaike information criterion (AIC). Backward selection began with all the variables and removed the least significant one at each step following the declined AIC until none met the criterion. It penalized models with a large number of variables for a simple and parsimonious model. The third baseline model was built with LASSO [27], which is another popular method used in clinical modeling. It is a regression-based method that performs regularization for variable selection to improve both the predictive accuracy and interpretability of the statistical model. Its regularization rate was optimized through 10-fold cross-validation in our study. The last two baseline models were built using RF. We created both a full RF model using all available variables and an RF model using the AutoScore-selected variables. The parameters were selected according to the suggestions in previous literature [28,29], where *ntree*=100 and $mtry=m^{1/2}$ (*ntree*: the number of trees grown; *mtry*: the number of variables randomly sampled as candidates at each split).

Statistical Analysis and Model Evaluation

Data were analyzed using R 3.5.3 (R Foundation). The baseline characteristics of the data set are described. In the descriptive summaries, frequencies and percentages were reported for categorical variables, whereas means and SDs were reported for continuous variables. We compared patients with and without inpatient mortality using a two-tailed Student *t* test for continuous variables and the χ^2 test for categorical variables. During the analysis, values of vital signs or laboratory tests were considered as outliers if they were beyond the normal range on the basis of domain knowledge. All detected outliers were set as missing values, which were subsequently imputed with the median values that were computed from the training set.

We compared the AutoScore-created scoring model with several baseline models to evaluate their predictive accuracy and interpretability. The test set was used to generate the metrics of model performance, and its bootstrapped samples were applied to calculate 95% CIs. Predictive accuracy was compared on the basis of ROC analysis and AUC values. Model interpretability was assessed by its complexity (eg, the number of variables included and the level of model nonlinearity) and its inherent explainability of the internal interaction. Model calibration was evaluated using the calibration belt plot test [30]. In addition, the distribution and observed mortality rate for each aggregated score were plotted for displaying its discriminative power.

Results

Baseline Characteristics of the Study Cohort

In this study, a total of 44,918 individual ICU admission episodes from the BIDMC data set were selected (Figure 2). Of all eligible episodes, 8.8% (3958/44,918) of the episodes had an outcome, that is, inpatient mortality. Summary baseline characteristics are shown in Table 1, and the distributions of other clinical continuous variables are shown in Table 2. In this cohort, the mean age was 62.5 (SD 16.5) years, 57.4% (25,788/44,918) were male, 84.9% (38,138/44,918) admissions were emergent, and the ethnic compositions were complex (31,889/44,918, 71.0% White; 4399/44,918, 9.8% African;

1625/44,918, 3.6% Hispanic; 1034/44,918, 2.3% Asian; and 5971/44,918, 13.3% others or unknown). We noticed that patients were admitted into different ICUs, which included Coronary Care Unit (CCU), Cardiac Surgery Recovery Unit (CSRU), Medical Intensive Care Unit (MICU), Surgical Intensive Care Unit (SICU), and Trauma Surgical Intensive Care Unit (TSICU). The average length of stay for all admission episodes was 4.19 (SD 6.11) days. Compared with the patients who survived to discharge, patients who died in hospitals were older, had a higher chance of emergency admission, had a longer length of stay, and a higher probability of being admitted to the MICU and paying by Medicare.

Table 1. Description of the study cohort (N=44,918).

Variables	All episodes (N=44,918)	Live discharged (n=40,960)	Inpatient mortality (n=3958)	*P* value
Age (years), mean (SD)	62.5 (16.5)	62.0 (16.6)	68.5 (14.7)	<.001
Gender, n (%)				.04
Male	25,788 (57.4)	23,578 (57.6)	2210 (55.8)	
Female	19,130 (42.6)	17,382 (42.4)	1748 (44.2)	
Admission type, n (%)				<.001
Emergency	38,138 (84.9)	34,339 (83.8)	3799 (96.0)	
Elective	6780 (15.1)	6621 (16.2)	159 (4.0)	
Ethnicity, n (%)				<.001
White	31,889 (71.0)	29,148 (71.2)	2741 (69.3)	
Hispanic	1625 (3.6)	1539 (3.8)	86 (2.2)	
Asian	1034 (2.3)	933 (2.3)	101 (2.6)	
African	4399 (9.8)	4110 (10.0)	289 (7.3)	
Others or unknown	5971 (13.3)	5230 (12.8)	741 (18.7)	
Insurance, n (%)				<.001
Government	1326 (3.0)	1258 (3.1)	68 (1.7)	
Medicaid	4176 (9.3)	3896 (9.5)	280 (7.1)	
Medicare	23,878 (53.2)	21,283 (52.0)	2595 (65.6)	
Private	15,031 (33.5)	14,063 (34.3)	968 (24.5)	
Self-pay	507 (1.1)	460 (1.1)	47 (1.2)	
ICU[a] type, n (%)				<.001
CCU[b]	6445 (14.3)	5907 (14.4)	538 (13.6)	
CSRU[c]	8284 (18.4)	8031 (19.6)	253 (6.4)	
MICU[d]	17,490 (38.9)	15,420 (37.6)	2070 (52.3)	
SICU[e]	7320 (16.3)	6649 (16.2)	671 (17.0)	
TSICU[f]	5379 (12.0)	4953 (12.1)	426 (10.8)	
Length of stay (days), mean (SD)	4.19 (6.11)	3.87 (5.75)	7.57 (8.36)	<.001

[a]ICU: intensive care unit.

[b]CCU: coronary care unit.

[c]CSRU: cardiac surgery recovery unit.

[d]MICU: medical intensive care unit.

[e]SICU: surgical intensive care unit.

[f]TSICU: trauma surgical intensive care unit.

Table 2. Distribution of clinical variables in the study cohort.

Variables	Values, median (IQR)
Age (years)	64.4 (51.9-75.9)
Heart rate (beats/min)	84.4 (74.5- 95.2)
Systolic blood pressure (mm Hg)	116.7 (107.1-129.5)
Diastolic blood pressure (mm Hg)	60 (53.7-67.4)
Mean arterial pressure (mm Hg)	76.9 (70.7-84.9)
Respiration rate (breaths/min)	18.0 (15.9-20.6)
Temperature (°C)	36.8 (36.5-37.2)
Peripheral capillary oxygen saturation (SpO_2; %)	97.6 (96.2-98.7)
Glucose (mg/dL)	129.0 (111.3-154.3)
Anion gap (mEq/L)	13.5 (12-16)
Bicarbonate (mmol/L)	24.0 (21.5-26.5)
Creatinine (μmol/L)	0.95 (0.7-1.4)
Chloride (mEq/L)	105 (101.5-108)
Lactate (mmol/L)	1.8 (1.7-2.0)
Hemoglobin (g/dL)	10.9 (9.6-12.3)
Hematocrit (%)	32.3 (28.8-36.4)
Platelet (thousand per microliter)	208.5 (153.5-276.5)
Potassium (mmol/L)	4.2 (3.8-4.5)
Blood urea nitrogen (mg/dL)	18.0 (12.5-29.5)
Sodium (mmol/L)	138.5 (136-140.5)
White blood cells (thousand per microliter)	10.7 (8.0-14.3)

Comparison of Selected Variables

Table 3 depicts the comparison of selected variables in the final model with different methods. The stepwise regression selected 22 variables, whereas the LASSO algorithm selected 17 variables after parameter tuning by 10-fold cross-validation. AutoScore selected a predefined number (*m*) of variables, and parameter *m* was optimized by a parsimony plot (ie, model performance vs complexity) on the validation set. As shown in part (a) of Figure 3, we chose 9 variables as the parsimonious choice as it achieved a good balance in the parsimony plot. When more variables were added to the scoring model, the performance was not markedly improved. Nine and 12 were selected as the number of variables in the demonstration. Users can also choose another parameter *m* if other restrictions or clinical preferences exist in real-life application scenarios. As seen from Table 3, the selected variables of AutoScore mostly coincided with those of the stepwise regression and LASSO. Notably, AutoScore generated a more parsimonious selection and sparse solution, catering to user preference and practical need.

Table 3. Selected variables by AutoScore and other baseline models.

Variables	Stepwise	LASSO	AutoScore (m=12)[a]	AutoScore (m=9)[a]
Age (years)	✔[b]	✔	✔	✔
Ethnicity	✔	✔	—[c]	—
Insurance	✔	✔	—	—
Gender	—	—	—	—
Heart rate	✔	✔	✔	✔
Systolic blood pressure	✔	✔	✔	✔
Diastolic blood pressure	✔	—	—	—
Mean arterial pressure	✔	✔	—	—
Respiration rate	✔	✔	✔	✔
Temperature	✔	✔	✔	✔
SpO_2[d]	✔	✔	✔	✔
Glucose	✔	✔	✔	—
Anion gap	✔	—	—	—
Bicarbonate	✔	✔	✔	—
Creatinine	✔	—	—	—
Chloride	✔	✔	—	—
Hematocrit	✔	✔	—	—
Hemoglobin	✔	—	—	—
Lactate	✔	✔	✔	✔
Platelet	✔	✔	✔	✔
Potassium	✔	✔	—	—
BUN[e]	✔	—	✔	✔
Sodium	—	✔	—	—
White blood cells	✔	—	✔	—

[a]Parameter m is the number of variables included in the AutoScore model.

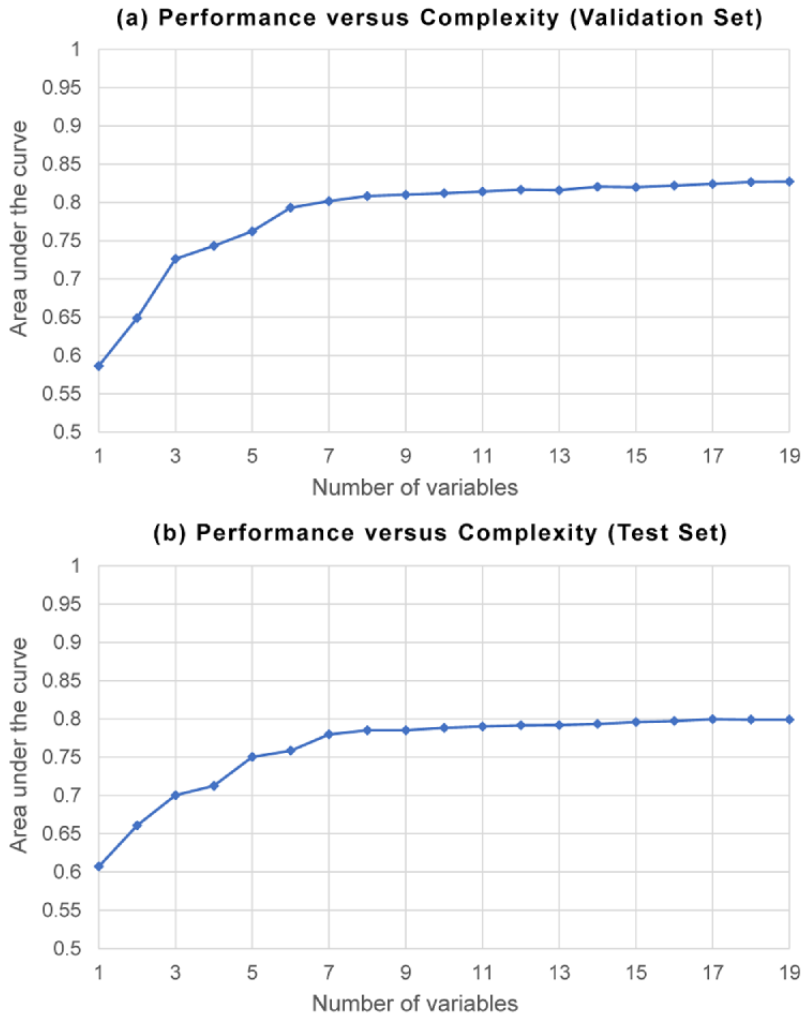
[b]Tick mark represents that this variable is included by the corresponding method.

[c]This variable is not included by the corresponding method.

[d]SpO_2: peripheral capillary oxygen saturation.

[e]BUN: blood urea nitrogen.

Figure 3. Model performance versus complexity for the implementation of the AutoScore on (a) the validation set and (b) the test set. The area under the curve reflects the discrimination performance, whereas the number of variables represents the complexity of the model.



Scoring Models by AutoScore

The nine-variable AutoScore-created scoring model of inpatient mortality for the BIDMC data set is tabulated in Table 4. Age, heart rate, respiration rate, systolic blood pressure, SpO_2, temperature, BUN, platelet, and lactate levels were selected into the final models. The final score summed up from 9 breakdowns ranged from 0 to 162. We used the test set to evaluate the property of this nine-variable point-based score. Part (a) of Figure 4 depicts the distribution of episodes at different score intervals, which is a near-normal distribution. Most patients had a risk score from 21 to 50, and very few patients had scores under 10 or above 80. As seen in part (b) of Figure 4, the observed mortality rate increased as our risk scores grew on the test set. The observed mortality rate was about 10% for a score of 50, whereas the mortality rate was over 50% for scores above 90. In terms of different breakdowns of the score, when age was lower than 30 years, its corresponding risk was the lowest; when it was higher than 85 years, the risk was the highest. Similarly, when the reported temperature was between 36.5°C and 37.5°C, the corresponding risk was the lowest, and when it was lower than 36°C, the risk was the highest. In addition, some variables, such as age, SpO_2, and BUN, have larger score values, indicating more significant contributions to the risk.

Table 4. A nine-variable AutoScore-created scoring model for inpatient mortality.

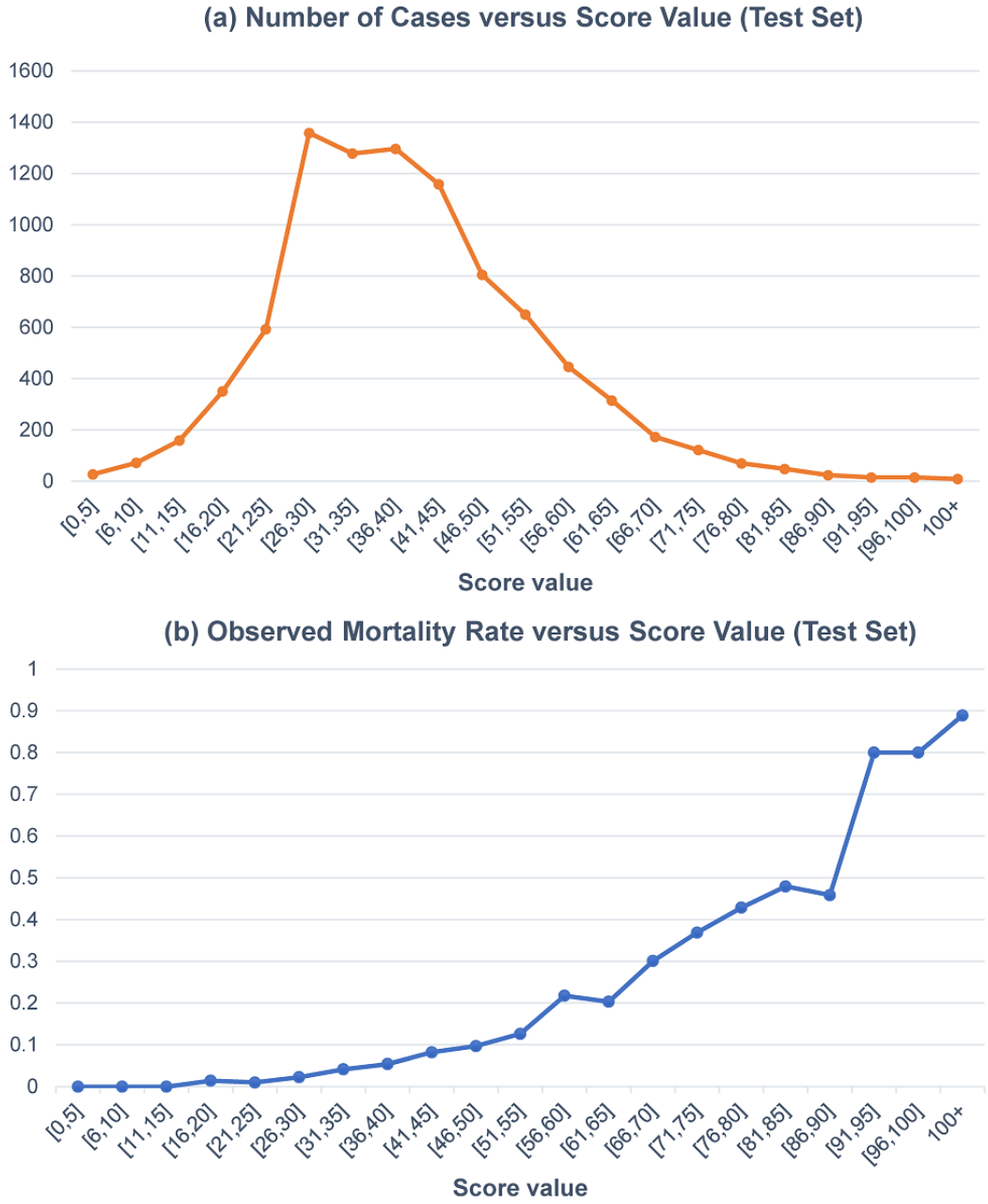
Variables and interval[a]	Point
Age (years)	
<30	0
30-48	5
48-78	14
78-85	22
≥85	24
Heart rate (beats/min)	
<62	1
62-72	0
72-98	1
98-112	8
≥112	13
Respiration rate (breaths/min)	
<12	3
12-16	0
16-22	4
≥22	12
Systolic blood pressure (mm Hg)	
<90	15
90-100	8
100-130	0
130-150	1
≥150	3
Temperature (°C)	
<36	12
36-36.5	3
36.5-37.5	0
37.5-38	5
≥38	9
SpO_2[b](%)	
<85	25
85-90	13
90-95	4
≥95	0
Platelet (thousand per microliter)	
<80	17
80-150	3
150-300	0
300-450	3
≥450	5
Blood urea nitrogen (mg/dL)	

Variables and interval[a]	Point
<7.5	0
7.5-12	2
12-35	9
35-70	19
≥70	23
Lactate (mmol/L)	
<1	0
1-2.5	2
2.5-4	8
≥4	21

[a]Interval (q_1-q_2) represents $q_1 \leq x < q_2$.

[b]SpO_2: peripheral capillary oxygen saturation.

Figure 4. (a) Number of cases and (b) observed mortality rate, versus different score intervals obtained by the nine-variable AutoScore model.



Comparison of Predictive Performance

The results of mortality prediction, as assessed by the ROC analysis on the unseen test set, are reported in Table 5. The scoring models generated by AutoScore showed promising discriminatory capability in predicting inpatient mortality. The 12-variable AutoScore model achieved an AUC of 0.789 (95% CI 0.773-0.802) with a sensitivity of 71.7% (95% CI 68.5%-74.7%) and a specificity of 71.7% (95% CI 70.7%-72.7%) under the optimal threshold (score=130). When we compromised on accuracy for parsimony, the nine-variable AutoScore model achieved a slightly lower AUC of 0.780 (95% CI 0.764-0.798) with a sensitivity of 63.7% (95% CI 60.3%-67.1%) and a specificity of 77.2% (95% CI 76.3%-78.2%) under the optimal threshold (score=48). In comparison, the performance of the 24-variable full logistic regression, the 22-variable stepwise regression, the 17-variable LASSO models, the nine-variable RF model, and the 24-variable full RF model achieved AUC values of 0.778 (95% CI 0.760-0.795), 0.778 (95% CI 0.760-0.795), 0.772 (95% CI 0.755-0.790), 0.785 (95% CI 0.768-0.801), and 0.809 (95% CI 0.794-0.825), respectively. Table 5 presents the performance metrics that were calculated under different score cutoffs. Besides the optimal cutoffs, other cutoffs by which the sensitivity or specificity could reach approximately 95% were also evaluated.

Table 5. Performance of the AutoScore and other baseline models.

Methods, AUC[a] (95% CI)	*m* [b]	Threshold	Sensitivity (%), 95% CI	Specificity (%), 95% CI	PPV[c] (%), 95% CI	NPV[d] (%), 95% CI
AutoScore (*m*[b]=9)						
0.780 (0.764-0.798)	9	48[e]	63.7 (60.3-67.1)	77.2 (76.3-78.2)	20.9 (19.8-22.0)	95.8 (95.4-96.1)
N/A[f]	N/A	30[g]	95.7 (94.3-97.2)	25.1 (24.2-26.0)	10.8 (10.6-10.9)	98.4 (97.9-98.9)
N/A	N/A	64[h]	28.8 (25.7-32.0)	95.5 (95.0-95.9)	37.6 (34.2-41.0)	93.4 (93.2-93.7)
AutoScore (*m*[b]=12)						
0.789 (0.773-0.802)	12	130[e]	71.7 (68.5-74.7)	71.7 (70.7-72.7)	19.3 (18.4-20.1)	96.4 (96.0-96.8)
N/A	N/A	95[g]	93.7 (92.0-95.3)	34.5 (33.4-35.6)	11.9 (11.6-12.1)	98.3 (97.8-98.7)
N/A	N/A	180[h]	32.0 (28.8-35.3)	94.8 (94.3-95.2)	36.6 (33.4-39.9)	93.7 (93.4-94.0)
Full logistic regression						
0.778 (0.760-0.795)	24	0.085[e]	68.6 (65.4-71.8)	72.8 (71.8-73.7)	19.2 (18.3-20.1)	96.1 (95.7-96.5)
N/A	N/A	0.028[g]	95.2 (93.5-96.6)	25.3 (24.4-26.3)	10.7 (10.5-10.9)	98.3 (97.7-98.8)
N/A	N/A	0.24[h]	27.9 (24.5-31.3)	95.1 (94.7-95.6)	35.0 (31.7-38.6)	93.3 (93.0-93.6)
Stepwise regression						
0.778 (0.760-0.795)	22	0.096[e]	65.0 (61.6-68.5)	76.9 (76.0-77.8)	21.0 (19.9-22.0)	95.9 (95.5-96.3)
N/A	N/A	0.028[g]	95.1 (93.5-96.5)	25.0 (24.1-26.1)	10.7 (10.5-10.9)	98.2 (97.6-98.7)
N/A	N/A	0.24[h]	28.4 (25.1-31.7)	95.2 (94.7-95.6)	35.7 (32.2-39.1)	93.4 (93.1-93.7)
LASSO[i]						
0.772 (0.755-0.790)	17	–2.47[e]	73.4 (70.2-76.4)	68.1 (67.1-69.2)	17.8 (17.0-18.6)	96.4 (96.0-96.8)
N/A	N/A	–3.34[g]	95.2 (93.7-96.5)	25.1 (24.1-26.1)	10.7 (10.5-10.9)	98.2 (97.7-98.7)
N/A	N/A	–1.27[h]	28.4 (25.2-31.8)	95.2 (94.7-95.7)	36.0 (32.6-39.5)	93.4 (93.1-93.7)
Random forest(*m*[b]=9)[j]						
0.785 (0.768-0.801)	9	0.085[e]	74.2 (71.1-77.0)	69.4 (68.4-70.4)	18.6 (17.8-19.4)	96.6 (96.2-97.0)
N/A	N/A	0.015[g]	94.2 (92.5-95.7)	30.1 (29.1-31.1)	11.3 (11.1-11.5)	98.2 (97.7-98.7)
N/A	N/A	0.3[h]	30.5 (27.4-34.0)	94.8 (94.4-95.3)	35.7 (32.5-39.0)	93.5 (93.3-93.8)
Full random forest						
0.809 (0.794-0.825)	24	0.115[e]	73.1 (69.9-76.2)	75.4 (74.5-76.3)	21.9 (20.9-22.9)	96.8 (96.4-97.1)
N/A	N/A	0.025[g]	94.4 (92.8-95.9)	37.9 (36.9-38.9)	12.5 (12.3-12.8)	98.6 (98.2-99.0)
N/A	N/A	0.285[h]	34.1 (30.6-37.5)	95.1 (94.6-95.5)	39.4 (36.2-42.9)	93.9 (93.6-94.2)

[a]AUC: the area under the ROC curve.

[b]Number of variables in the model.

[c]PPV: positive predictive value.

[d]NPV: negative predictive value.

[e]Optimal cutoff values, defined as the points nearest to the upper-left corner of the ROC curves.

[f]N/A: not applicable.

[g]Cutoff values by which the sensitivity could reach about 95%.

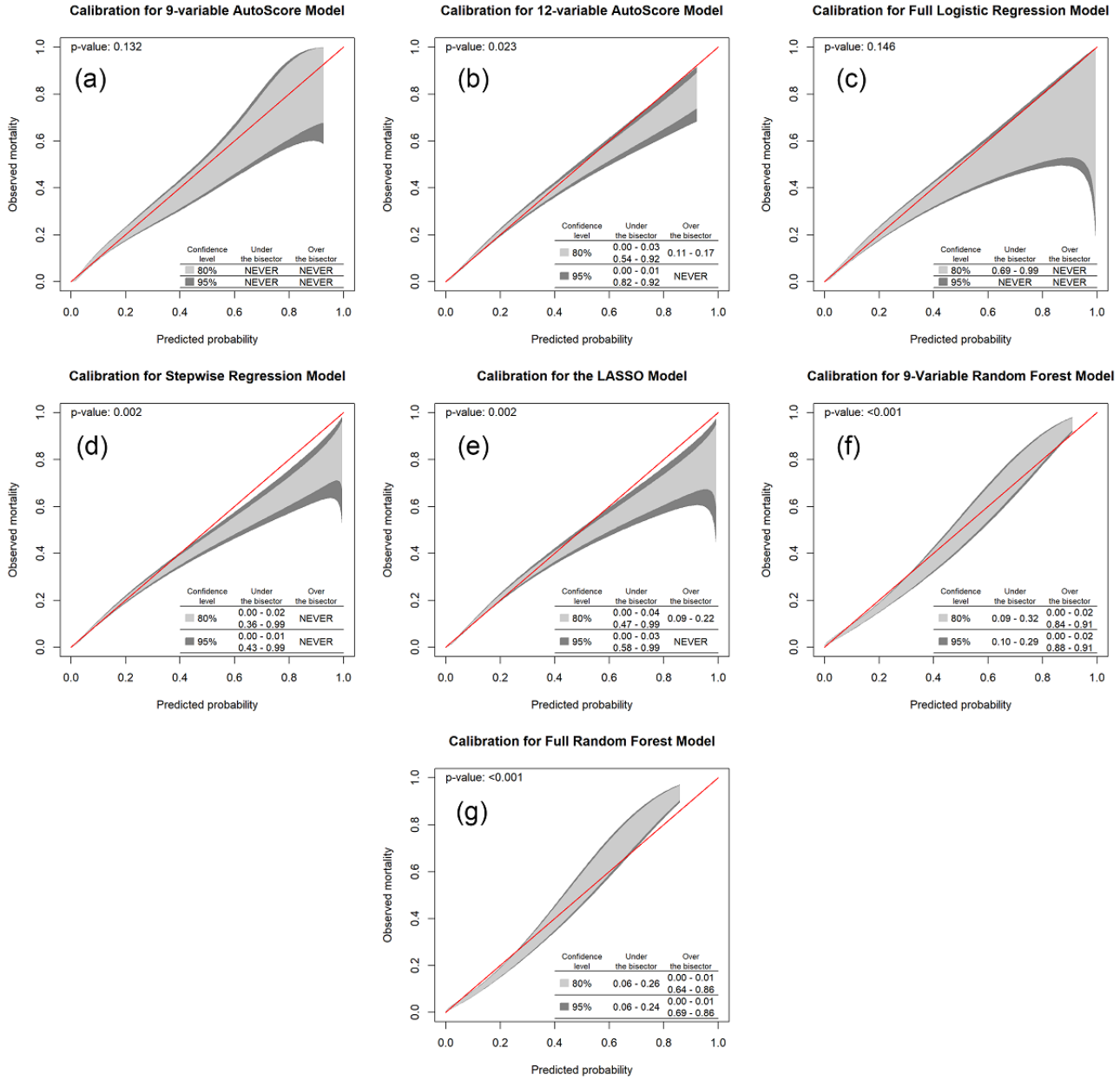
[h]Cutoff values by which the specificity could reach about 95%.

[i]LASSO: least absolute shrinkage and selection operator.

[j]AutoScore-based variable selection was implemented beforehand, where the same set of variables were selected as the AutoScore (*m*=9).

As illustrated in Figure 5, our nine-variable AutoScore model remained well calibrated, and all parts of the calibration belt showed a good fit under both 80% and 95% CIs. In comparison, other models displayed varying degrees of underestimation or overestimation. Two RF models performed the worst in the calibration test, followed by the stepwise regression and LASSO models. On the contrary, the AutoScore and logistic regression perform relatively well in terms of model calibration.

Figure 5. Calibration belts (at 80% and 95% confidence levels) for (a) a nine-variable AutoScore-created model, (b) a 12-variable AutoScore-created model, (c) a full logistic regression model, (d) a stepwise regression model, (e) the LASSO model, (f) a nine-variable random forest model, and (g) a full random forest model.



Discussion

Principal Findings

In this study, we developed AutoScore, a framework of automatic clinical score creation, and tested it in a large clinical data set. The scoring models generated by AutoScore were comparable with other standard methods (ie, logistic regression, stepwise regression, LASSO, RF model) in terms of predictive performance and model calibration. More importantly, the AutoScore-created scoring models showed superiority in interpretability and accessibility, as they were point-based scores with fewer variables used. In clinical practice, point-based scores have the advantage of easy implementation and, thus, can be widely utilized and validated in different circumstances and health care settings. The novelty of our study was the development of a generic, scalable, and robust methodology for automatically generating a point-based scoring model, which has been demonstrated by deriving an actual scoring model of inpatient mortality with a large benchmark EHR data set.

The proposed AutoScore has several advantages in creating risk prediction models. First, the machine learning–based variable ranking or selection can efficiently filter out redundant

information. The importance of including variable selection in the development of predictive models has been demonstrated in many studies. In a study by Zhao et al [31], variable selection removed noninformative variables from the clinical predictive model. Bagherzadeh-Khiabani et al [32] demonstrated that the use of variable selection could improve the performance of clinical prediction models. Sanchez-Pinto et al [11] also provided evidence of modern tree-based methods of variable selection with better parsimony in large data sets. Liu et al [33] demonstrated that machine learning–based variable selection was promising for discovering a few relevant and significant variables in the prediction of adverse cardiac events. Second, the module of variable transformation could improve the fit of models. Several studies [34,35] have reported U-shaped nonlinearity between continuous variables and health-related outcomes. According to expert opinion, the value of vital signs or laboratory tests is usually considered as an abnormal value if it is beyond a healthy normal range. Besides, the categorization of continuous variables remains to be a dominant practice in epidemiological studies [36]. Discretizing features requires a smaller memory footprint, simplifies model interpretation, and can be applied directly by a human expert in routine care [37]. In addition, categorization creates a natural way to handle missing values, where the missing values can be treated as an extra category. This missing-indicator method has the appealing property that all available information can be used in the analyses [38]. Third, we use a parsimony plot (model performance vs complexity) to determine the appropriate number of variables (*m*), balancing the trade-off between performance and sparsity [39,40]. We value the model parsimony as the most desirable characteristic, as there is a real-world cost associated with mapping numerous variables, maintaining complex algorithms, and replicating it in different settings. This parsimony-driven parameter tuning process can be performed in an independent validation set (ie, 10% randomly selected samples from the entire data set in this study), as shown in Figure 3. It also shows a similar trend on the basis of the unseen test set, illustrating the effectiveness and consistency of parsimony-driven tuning for determining the number of necessary variables.

Furthermore, the scoring models created by the AutoScore framework are interpretable and clinically practical. The output of AutoScore is a point-based scoring model, based on addition, subtraction, and multiplication of a few sparse numbers, facilitating quick stratification without the need for a computing system. Doctors can easily understand how risk models make predictions in a transparent manner. Although numerous machine learning models, such as neural networks [41,42] and ensemble learning models [43,44], have been developed to complement traditional regression models, most of them are black boxes that do not explain their predictions in a way that humans can understand. In our study, the nine-variable RF model was performed as accurately as our nine-variable AutoScore (AUC 0.785 vs 0.780). However, it is challenging to explain the prediction made by the RF model, which consists of 100 different decision trees together. The lack of transparency of predictive models could lead to severe consequences in patient care. Vellido [12] suggested that these models with low explainability are unlikely to become part of routine clinical and health care practice as providing care is a highly sensitive task. Rudin [45] also suggested designing models that are inherently interpretable rather than explaining black box models and doubted the blind belief in the myth of the accuracy-interpretability trade-off.

Relationship With Previous Work

Researchers have previously created several scoring models for predicting mortality, such as the Modified Early Warning Score [46], the VitalPAC Early Warning Score [47], and the Acute Physiology And Chronic Health Evaluation [48], mainly utilizing vital signs to predict mortality for hospitalized patients. However, they were designed by hand subjectively from expert opinions and domain knowledge, which hindered their generalization and dynamic evolution. Considering the disparate EHR systems among various health care settings, these scoring models may not work well because of the diversity among routinely collected information. As the characteristics of the population evolve, the adjustment and updating of risk scores are needed, which are time-consuming and inflexible [49]. In contrast, our AutoScore framework is adaptive and flexible; it can generate scoring models automatically, given an evolving EHR system. A user-friendly and easy-to-use R package of AutoScore [22] has been developed to facilitate the creation of scoring systems in diverse contexts, satisfying the increasing need for the development of specific predictive scores in various health care settings.

Similar to our AutoScore framework, Zhang et al [50] presented a tutorial on building a scoring system from several steps. However, the tutorial did not integrate some vital components such as variable ranking or selection and several crucial tuning processes inherently into the process of score generation. In comparison, our AutoScore framework includes all essential modules, driving the clinical continuum of 6 modules and realizing the automation. Although users may benefit from the built-in automation of AutoScore for developing a clinical score, domain knowledge is equally important in building the scoring models, as suggested in many studies [10,51]. In AutoScore, domain knowledge can be involved in 2 ways: (1) the variable can be preselected by expert opinion before implementing the AutoScore and (2) domain knowledge can be used to fine-tune the risk scores and determine clinically valid cutoff values in variable transformation.

Future Research and Limitations

Although the proposed AutoScore framework is comprehensively and systematically presented, improvements can still be made. Each module of the AutoScore can be improved using advanced algorithms and enhanced methodologies. For example, in the module of variable ranking, various established machine learning methods can potentially be integrated into the AutoScore framework. In variable transformation, the means of categorization may be customized according to its distribution, provided a handful of clinical variables such as SpO_2 that may not be subject to a near-normal distribution. Furthermore, the application of AutoScore is not limited to its application to large-scale EHR data [24,52]. AutoScore can be readily implemented in small-scale observational cohort studies. Beyond health care applications,

AutoScore is potentially applicable to other high-stakes prediction applications such as criminal justice and finance, where highly interpretable predictive models are needed.

This study has several limitations. First, the data set used in this study was on the basis of EHR data with routinely collected vital and laboratory test variables. Some relevant variables were not available in this analysis. For example, health utilization, such as intubation and resuscitation, has been proven to be predictive of overall mortality. Second, given the limitation in data availability, the clinical scores built with AutoScore in this study are not perfect for real-world implementation. This clinical study was primarily designed to demonstrate the effectiveness of the AutoScore framework in building risk scores. Third, this was a retrospective analysis. To further prove its clinical practicability, prospective validation of the scoring model is needed. Finally, this was the initial development of AutoScore, where only selected methods were integrated into the framework, leaving opportunities for further development with more sophisticated and state-of-the-art algorithms.

Conclusions

We developed an easy-to-use, machine learning–based automatic clinical score generator, AutoScore, to conveniently build scoring models and demonstrated its usability with a clinical study on mortality prediction. Using a benchmark data set, we showed that the scoring models derived with the AutoScore framework achieved satisfactory predictive performance and proved its superiority over several conventional methods for risk model development. The AutoScore framework integrates both the advantage of machine learning in strong discriminative power and the merit of point-based scores in its excellent accessibility and interpretability. Our proposed AutoScore framework can be readily used to generate clinical scores in various medical applications, such as early warning systems and risk predictions of mortality, hospital readmissions, and adverse cardiac events. In the future, advanced machine learning algorithms and methodologies could improve individual modules of AutoScore and provide AutoScore with more robust predictive capability or broader applicability in various types of data.

Acknowledgments

This study was supported by the Singapore National Medical Research Council under the PULSES Center Grant.

Authors' Contributions

NL conceived and supervised the study. FX and NL developed the AutoScore algorithm and wrote the first draft of the manuscript. FX analyzed the data. FX, BC, MO, BG, and NL made substantial contributions to the interpretation of results, algorithm improvement, and critical revision of the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
AutoScore R package (version 0.1).
[ZIP File (Zip Archive), 943 KB - medinform_v8i10e21798_app1.zip]

Multimedia Appendix 2
AutoScore R package (version 0.1) source codes.
[ZIP File (Zip Archive), 687 KB - medinform_v8i10e21798_app2.zip]

References

1. Smith ME, Chiovaro JC, O'Neil M, Kansagara D, Quiñones AR, Freeman M, et al. Early warning system scores for clinical deterioration in hospitalized patients: a systematic review. Ann Am Thorac Soc 2014 Nov;11(9):1454-1465. [doi: 10.1513/AnnalsATS.201403-102OC] [Medline: 25296111]
2. Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the national early warning score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation 2013 Apr;84(4):465-470. [doi: 10.1016/j.resuscitation.2012.12.016] [Medline: 23295778]
3. Brady W, de Souza K. The HEART score: a guide to its application in the emergency department. Turk J Emerg Med 2018 Jun;18(2):47-51 [FREE Full text] [doi: 10.1016/j.tjem.2018.04.004] [Medline: 29922729]
4. Churpek MM, Yuen TC, Park SY, Meltzer DO, Hall JB, Edelson DP. Derivation of a cardiac arrest prediction model using ward vital signs. Crit Care Med 2012 Jul;40(7):2102-2108 [FREE Full text] [doi: 10.1097/CCM.0b013e318250aa5a] [Medline: 22584764]
5. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. Intensive Care Med 1996 Jul;22(7):707-710. [doi: 10.1007/BF01709751] [Medline: 8844239]

6. Jones M. NEWSDIG: The national early warning score development and implementation group. Clin Med (Lond) 2012 Dec;12(6):501-503 [FREE Full text] [doi: 10.7861/clinmedicine.12-6-501] [Medline: 23342400]
7. Six AJ, Backus BE, Kelder JC. Chest pain in the emergency room: value of the HEART score. Neth Heart J 2008 Jun;16(6):191-196 [FREE Full text] [doi: 10.1007/BF03086144] [Medline: 18665203]
8. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373-383. [doi: 10.1016/0021-9681(87)90171-8] [Medline: 3558716]
9. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):198-208 [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]
10. Heinze G, Wallisch C, Dunkler D. Variable selection - a review and recommendations for the practicing statistician. Biom J 2018 May;60(3):431-449 [FREE Full text] [doi: 10.1002/bimj.201700067] [Medline: 29292533]
11. Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. Int J Med Inform 2018 Aug;116:10-17 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.05.006] [Medline: 29887230]
12. Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Applic 2019 Feb 4:- epub ahead of print. [doi: 10.1007/s00521-019-04051-w]
13. O'Brien C, Goldstein BA, Shen Y, Phelan M, Lambert C, Bedoya AD, et al. Development, implementation, and evaluation of an in-hospital optimized early warning score for patient deterioration. MDM Policy Pract 2020;5(1):2381468319899663. [doi: 10.1177/2381468319899663] [Medline: 31976373]
14. Low LL, Liu N, Lee KH, Ong ME, Wang S, Jing X, et al. FAM-FACE-SG: a score for risk stratification of frequent hospital admitters. BMC Med Inform Decis Mak 2017 Apr 8;17(1):35 [FREE Full text] [doi: 10.1186/s12911-017-0441-5] [Medline: 28390405]
15. Sullivan LM, Massaro JM, D'Agostino RB. Presentation of multivariate data for clinical use: the Framingham study risk score functions. Stat Med 2004 May 30;23(10):1631-1660. [doi: 10.1002/sim.1742] [Medline: 15122742]
16. Tin KH. Random Decision Forests. 1995 Presented at: International Conference on Document Analysis and Recognition; August 14-16, 1995; Montreal, Quebec, Canada. [doi: 10.1109/icdar.1995.598994]
17. Flaxman AD, Vahdatpour A, Green S, James SL, Murray CJ, Population Health Metrics Research Consortium (PHMRC). Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. Popul Health Metr 2011 Aug 4;9:29 [FREE Full text] [doi: 10.1186/1478-7954-9-29] [Medline: 21816105]
18. Breiman L, Friedman J, Stone C, Olshen R. Classification and Regression Trees. New York, USA: Taylor & Francis; 1984.
19. Biau G. Analysis of a random forests model. J Mach Learn Res 2012;13:1063-1095.
20. Genuer R, Poggi J, Tuleau-Malot C. Variable selection using random forests. Pattern Recogn Letters 2010 Oct;31(14):2225-2236. [doi: 10.1016/j.patrec.2010.03.014]
21. Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding Variable Importances in Forests of Randomized Trees. In: Advances in Neural Information Processing Systems. 2013 Presented at: NIPS'13; December 5-10, 2013; Lake Tahoe, Nevada p. 431-439.
22. AutoScore: A Machine Learning-Based Automatic Clinical Score Generator. GitHub. URL: https://github.com/nliulab/AutoScore [accessed 2020-09-28]
23. 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]
24. Xie F, Liu N, Wu SX, Ang Y, Low LL, Ho AF, et al. Novel model for predicting inpatient mortality after emergency admission to hospital in Singapore: retrospective observational study. BMJ Open 2019 Sep 26;9(9):e031382. [doi: 10.1136/bmjopen-2019-031382] [Medline: 31558458]
25. Redfern OC, Pimentel MA, Prytherch D, Meredith P, Clifton DA, Tarassenko L, et al. Predicting in-hospital mortality and unanticipated admissions to the intensive care unit using routinely collected blood tests and vital signs: development and validation of a multivariable model. Resuscitation 2018 Dec;133:75-81 [FREE Full text] [doi: 10.1016/j.resuscitation.2018.09.021] [Medline: 30253229]
26. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R. New York, USA: Springer New York; 2013.
27. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 2018 Dec 5;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
28. Oshiro T, Perez P, Baranauskas J. How Many Trees in a Random Forest? Machine Learning and Data Mining in Pattern Recognition 2012;Berlin, Heidelberg:154-168. [doi: 10.1007/978-3-642-31537-4_13]
29. Probst P. To tune or not to tune the number of trees in random forest. The Journal of Machine Learning Research 2017;18(1):6673-6690.
30. Nattino G, Finazzi S, Bertolini G. A new test and graphical tool to assess the goodness of fit of logistic regression models. Stat Med 2016 Feb 28;35(5):709-720. [doi: 10.1002/sim.6744] [Medline: 26439593]

31. 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]
32. Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg EW, Khalili D. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. J Clin Epidemiol 2016 Mar;71:76-85. [doi: 10.1016/j.jclinepi.2015.10.002] [Medline: 26475568]
33. Liu N, Koh ZX, Goh J, Lin Z, Haaland B, Ting BP, et al. Prediction of adverse cardiac events in emergency department patients with chest pain using machine learning for variable selection. BMC Med Inform Decis Mak 2014 Aug 23;14:75 [FREE Full text] [doi: 10.1186/1472-6947-14-75] [Medline: 25150702]
34. Salonia A, Abdollah F, Capitanio U, Suardi N, Briganti A, Gallina A, et al. Serum sex steroids depict a nonlinear u-shaped association with high-risk prostate cancer at radical prostatectomy. Clin Cancer Res 2012 Jul 1;18(13):3648-3657. [doi: 10.1158/1078-0432.CCR-11-2799] [Medline: 22589393]
35. Chen Y, Huang J, He X, Gao Y, Mahara G, Lin Z, et al. A novel approach to determine two optimal cut-points of a continuous predictor with a U-shaped relationship to hazard ratio in survival data: simulation and application. BMC Med Res Methodol 2019 May 9;19(1):96 [FREE Full text] [doi: 10.1186/s12874-019-0738-4] [Medline: 31072334]
36. Mabikwa OV, Greenwood DC, Baxter PD, Fleming SJ. Assessing the reporting of categorised quantitative variables in observational epidemiological studies. BMC Health Serv Res 2017 Mar 14;17(1):201 [FREE Full text] [doi: 10.1186/s12913-017-2137-z] [Medline: 28288628]
37. Schellingerhout JM, Heymans MW, de Vet HC, Koes BW, Verhagen AP. Categorizing continuous variables resulted in different predictors in a prognostic model for nonspecific neck pain. J Clin Epidemiol 2009 Aug;62(8):868-874. [doi: 10.1016/j.jclinepi.2008.10.010] [Medline: 19230604]
38. Donders A, van der Heijden GJ, Stijnen T, Moons K. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006 Oct;59(10):1087-1091. [doi: 10.1016/j.jclinepi.2006.01.014] [Medline: 16980149]
39. Murtaugh PA. Methods of variable selection in regression modeling. Commun Stat - Simul Comput 2010 Dec 23;27(3):711-734. [doi: 10.1080/03610919808813505]
40. Austin PC, Tu JV. Bootstrap methods for developing predictive models. Am Stat 2004 May;58(2):131-137. [doi: 10.1198/0003130043277]
41. 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]
42. Thorsen-Meyer H, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Digit Health 2020 Apr;2(4):e179-e191. [doi: 10.1016/s2589-7500(20)30018-2]
43. Klug M, Barash Y, Bechler S, Resheff YS, Tron T, Ironi A, et al. A gradient boosting machine learning model for predicting early mortality in the emergency department triage: devising a nine-point triage score. J Gen Intern Med 2020 Jan;35(1):220-227. [doi: 10.1007/s11606-019-05512-7] [Medline: 31677104]
44. Spangler D, Hermansson T, Smekal D, Blomberg H. A validation of machine learning-based risk scores in the prehospital setting. PLoS One 2019;14(12):e0226518 [FREE Full text] [doi: 10.1371/journal.pone.0226518] [Medline: 31834920]
45. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019 May 13;1(5):206-215. [doi: 10.1038/s42256-019-0048-x]
46. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified early warning score in medical admissions. QJM 2001 Oct;94(10):521-526. [doi: 10.1093/qjmed/94.10.521] [Medline: 11588210]
47. Prytherch DR, Smith GB, Schmidt PE, Featherstone PI. ViEWS--towards a national early warning score for detecting adult inpatient deterioration. Resuscitation 2010 Aug;81(8):932-937. [doi: 10.1016/j.resuscitation.2010.04.014] [Medline: 20637974]
48. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 2006 May;34(5):1297-1310. [doi: 10.1097/01.CCM.0000215112.84523.F0] [Medline: 16540951]
49. Hendriksen JM, Geersing GJ, Moons KG, de Groot JA. Diagnostic and prognostic prediction models. J Thromb Haemost 2013 Jun;11(Suppl 1):129-141 [FREE Full text] [doi: 10.1111/jth.12262] [Medline: 23809117]
50. Zhang Z, Zhang H, Khanal MK. Development of scoring system for risk stratification in clinical medicine: a step-by-step tutorial. Ann Transl Med 2017 Nov;5(21):436 [FREE Full text] [doi: 10.21037/atm.2017.08.22] [Medline: 29201888]
51. Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. J Clin Epidemiol 2004 Nov;57(11):1138-1146. [doi: 10.1016/j.jclinepi.2004.04.003] [Medline: 15567629]
52. Mahmoudi E, Kamdar N, Kim N, Gonzales G, Singh K, Waljee AK. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. Br Med J 2020 Apr 8;369:m958. [doi: 10.1136/bmj.m958] [Medline: 32269037]

Abbreviations

AIC: Akaike information criterion
AUC: area under the curve
BIDMC: Beth Israel Deaconess Medical Center
BUN: blood urea nitrogen
CCU: coronary care unit
CSRU: cardiac surgery recovery unit
EHR: electronic health record
ICU: intensive care unit
LASSO: least absolute shrinkage and selection operator
MAP: mean arterial pressure
MICU: medical intensive care unit
NPV: negative predictive value
PPV: positive predictive value
RF: random forest
ROC: receiver operating characteristic
SICU: surgical intensive care unit
TSICU: trauma surgical intensive care unit

Edited by G Eysenbach; submitted 25.06.20; peer-reviewed by J Li, M Adly, A Adly; comments to author 14.07.20; revised version received 25.07.20; accepted 27.07.20; published 21.10.20.

Please cite as:
Xie F, Chakraborty B, Ong MEH, Goldstein BA, Liu N
AutoScore: A Machine Learning–Based Automatic Clinical Score Generator and Its Application to Mortality Prediction Using Electronic Health Records
JMIR Med Inform 2020;8(10):e21798
URL: http://medinform.jmir.org/2020/10/e21798/
doi:10.2196/21798
PMID:33084589

Original Paper

Institution-Specific Machine Learning Models for Prehospital Assessment to Predict Hospital Admission: Prediction Model Development Study

Toru Shirakawa[1,2], MD; Tomohiro Sonoo[2,3], MD; Kentaro Ogura[2,4]; Ryo Fujimori[2,4]; Konan Hara[2,5], MD, PhD; Tadahiro Goto[2,6], MD, MPH; Hideki Hashimoto[3], MD; Yuji Takahashi[3], MD; Hiromu Naraba[3], MD; Kensuke Nakamura[3,7], MD, PhD

[1]Department of Public Health, Graduate School of Medicine, Osaka University, Suita, Japan

[2]TXP Medical Co, Ltd, Chuo-ku, Japan

[3]Department of Emergency Medicine, Hitachi General Hospital, Hitachi, Japan

[4]Faculty of Medicine, The University of Tokyo, Bunkyo-ku, Japan

[5]Department of Public Health, The University of Tokyo, Bunkyo-ku, Japan

[6]Department of Clinical Epidemiology and Health Economics, School of Public Health, The University of Tokyo, Bunkyo-ku, Japan

[7]Department of Emergency Medicine, The University of Tokyo, Bunkyo-ku, Japan

Corresponding Author:
Tomohiro Sonoo, MD
Department of Emergency Medicine
Hitachi General Hospital
Jounan-cho 2-1-1
Hitachi, 317-0077
Japan
Phone: 81 294 23 1111
Email: tomohiro.sonoo@txpmedical.com

Abstract

Background: Although multiple prediction models have been developed to predict hospital admission to emergency departments (EDs) to address overcrowding and patient safety, only a few studies have examined prediction models for prehospital use. Development of institution-specific prediction models is feasible in this age of data science, provided that predictor-related information is readily collectable.

Objective: We aimed to develop a hospital admission prediction model based on patient information that is commonly available during ambulance transport before hospitalization.

Methods: Patients transported by ambulance to our ED from April 2018 through March 2019 were enrolled. Candidate predictors were age, sex, chief complaint, vital signs, and patient medical history, all of which were recorded by emergency medical teams during ambulance transport. Patients were divided into two cohorts for derivation (3601/5145, 70.0%) and validation (1544/5145, 30.0%). For statistical models, logistic regression, logistic lasso, random forest, and gradient boosting machine were used. Prediction models were developed in the derivation cohort. Model performance was assessed by area under the receiver operating characteristic curve (AUROC) and association measures in the validation cohort.

Results: Of 5145 patients transported by ambulance, including deaths in the ED and hospital transfers, 2699 (52.5%) required hospital admission. Prediction performance was higher with the addition of predictive factors, attaining the best performance with an AUROC of 0.818 (95% CI 0.792-0.839) with a machine learning model and predictive factors of age, sex, chief complaint, and vital signs. Sensitivity and specificity of this model were 0.744 (95% CI 0.716-0.773) and 0.745 (95% CI 0.709-0.776), respectively.

Conclusions: For patients transferred to EDs, we developed a well-performing hospital admission prediction model based on routinely collected prehospital information including chief complaints.

(JMIR Med Inform 2020;8(10):e20324) doi:10.2196/20324

KEYWORDS

prehospital; prediction; hospital admission; emergency medicine; machine learning; data science

Introduction

For patients being transported to an emergency department (ED), predicting hospital admission is important for providing high-quality care. Choosing the appropriate destination hospital with available beds can enhance efficient resource utilization in the context of integrated community health care [1]. Furthermore, accurate risk stratification during transportation can be expected to curb the risk of ED overcrowding and reduce ambulance turnaround times when implemented at hospitals [2].

Although multiple prediction models have been developed to predict hospital admission for ED use [3-11] to address overcrowding and patient safety [12-15], few studies have examined prediction models for prehospital use. Previously reported prehospital prediction models have been limited to patients with a specific disease or to models predicting critically ill conditions or mortality [16-23]. Several studies in the United States and United Kingdom have demonstrated the predictive performance of ED disposition, including hospitalization for general patients transferred by ambulance [24-26]. Nevertheless, these studies were not based on statistical models but on subjective prediction by ambulance staff. Therefore, they have limited generalizability across emergency medical systems and countries. Another study, conducted in Sweden, assessed a prehospital prediction model of hospital admission [27]. However, its predictors included more than 1000 distinct question and answer combinations recorded in a clinical decision support system used at a dispatch center. Therefore, its scalability might not be readily achievable.

Given this context, we aimed to develop prehospital prediction models of hospital admission using machine learning techniques and conventional logistic regression, based on replicable measurements such as chief complaints, vital signs, and past medical histories, which can all be collected routinely in an ambulance in any country. Our goal was to develop an institution-specific model based on readily collectable data with sufficient predictive performance, not a universal model that has broad generalizability.

Methods

Study Design and Setting

This prognostic study used data obtained at a tertiary care hospital in Japan from April 2018 to March 2019. The hospital covers approximately 3 million local residents. Annually, the hospital has about 20,000-25,000 visits, including 5500-6500 ambulance visits. The study protocol was approved by the Ethics Committee of the hospital. They waived informed consent because of the characteristics of the retrospective study design.

Study Participants

We enrolled patients who had been transported to our ED by ambulance. We excluded children aged 6 years or younger because of the difficulty in taking chief complaints and measuring vital signs such as blood pressure. Patients with cardiopulmonary arrest were not excluded from analyses, thereby facilitating comparison with earlier studies that included patients with cardiopulmonary arrest and examined the predictive performance of ambulance staff [24-26].

Patient Information in the Prehospital Setting

Vital signs were measured at the scene when the patient was placed in the ambulance. After emergency medical service (EMS) staff members recorded patient information and conditions during transportation, they transmitted the information via telephone to ED staff members at the destination hospital. This information was input into an ED database through the Next Stage ER system (TXP Medical Co, Ltd), which structures information related to the chief complaint and past medical history with flexible input templates and a minor natural language processing algorithm [28]. The recorded chief complaint was translated automatically into 231 chief complaint categories based on the Japan Triage and Acuity Scale (JTAS) [29], which was developed based on the Canadian Triage and Acuity System [30]. Past medical histories were encoded corresponding to the International Statistical Classification of Diseases, 10th Revision (ICD-10) codes [31].

Candidate Predictors

Candidate predictors were age, sex, chief complaints, prehospital vital signs, and past medical histories. Although chief complaints were grouped into 231 categories based on JTAS, 75 complaints were not observed (ie, none of the included patients presented with these complaints). Therefore, 156 complaints were used. Vital signs include the level of consciousness, systolic blood pressure, diastolic blood pressure, pulse rate, respiratory rate, body temperature, and oxygen saturation with oxygen administration during transportation. The level of consciousness was assessed according to the Japan Coma Scale, which can be summarized briefly into four categories of alert, possible eye opening but not lucid, possible eye opening upon stimulation, and no eye opening and coma [32]. Past medical histories were grouped using the first 3 characters (1 alphabet letter and 2 digits) of the ICD-10 code. The 156 chief complaints and 505 past medical histories observed in our study were encoded to dummy variables. In all, 832 predictors were identified as candidate predictors.

Outcome Measures

The primary outcome was the composite of hospitalization, transfer to other care facilities, and death at the ED. These outcomes were recorded at the time patients left the ED. Sensitivity analysis was performed by excluding mortality from the hospitalization outcomes.

Data Analysis

Model Development

To predict hospital admission, we developed four models using candidate predictors as explained above: (1) logistic regression, (2) logistic regression with lasso penalization (logistic lasso),

(3) random forest [33], and (4) gradient boosting machine (GBM) [34]. For the GBM model, we used the extreme gradient boosting (XGBoost) implementation [35]. For each model, to evaluate the incremental benefit of adding each predictor, we further developed four models according to the predictors. Model 1 consists of age and sex only. Model 2 further includes 156 chief complaints. Model 3 further includes vital signs. Model 4 further includes 505 past medical histories. These modalities were designed according to the typical temporal order of information collection processing: call by a patient or bystander, arrival of an emergency medical team, and examination in the ambulance.

Feature Processing

To account for potential nonlinear relations between continuous features and the risk of hospital admission, we categorized the values of age and vital signs into deciles for logistic regression and logistic lasso. Since random forest and GBM can accommodate the nonlinear relations, we used continuous age and vital signs in those models.

Study Cohorts and Missing Values

We used 70.0% (3601/5145) of the available data for the derivation cohort. The remaining 30.0% (1544/5145) of data were used for the validation cohort. We divided patients into the two groups by random allocation. Hyperparameters for machine learning models were determined using a grid search with 5-fold cross-validation in the derivation cohort. Among the 5145 patients, frequencies (proportions) of missing values were 25 (0.5%) for sex, 552 (10.7%) for orientation, 593 (11.5%) for systolic blood pressure, 647 (12.6%) for diastolic blood pressure, 511 (9.9%) for pulse rate, 1152 (22.4%) for respiratory rate, 1040 (20.2%) for oxygen saturation, and 1086 (21.1%) for body temperature. The number of patients with at least one missing vital sign was 2174 (42.3%). To address the missing data, we used a missing indicator for logistic regression and lasso, assigned 0 for random forest, and left missing data in GBM, for which XGBoost can accommodate missing values.

Model Validation

In the validation cohort, we examined the prediction ability of the models by calculating the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). Calibration of the models was depicted by plotting predicted probabilities and the observed admission rates according to deciles of the predicted probabilities. Sensitivity, specificity, positive predictive value, negative predictive value, and accuracy were estimated with predictors in the most accurate model at the threshold probability that maximizes the Youden indices [36].

For comparison with earlier studies of hospital admission prediction in the ED including walk-in patients and those transported by ambulance [3-11], we evaluated the prediction performance of the model described above including walk-in patients.

All analyses were conducted using Python 3.7 with scikit-learn [37], XGBoost [35], and tableone [38] packages. We used 200 bootstrap samples to calculate 95% confidence intervals for performance measures. Two-tailed *P* values of <.05 were inferred as statistically significant.

Results

During the study period, 5530 patients were transported to our ED by ambulance. From these, we excluded 385 visits by patients aged 6 years or younger. In all, 5145 visits were included in the analyses. Among the 5145 visits with ambulance transport, 2507 visits (48.7%) led to hospital admission, 96 visits (1.9%) led to death in the ED, and 96 visits (1.9%) required hospital transfer. The number of patients who required hospital admission, died in the ED, or required hospital transfer was 1889 of 3601 patients (52.5%) in the derivation cohort and 810 of 1544 patients (52.5%) in the validation cohort. Compared to patients who were not admitted to the hospital, patients who were admitted to the hospital (including those who died or were transferred) had worse vital signs (eg, lower level of consciousness, lower blood pressure). Moreover, they were older, were likely to have altered mental status or fever, and were likely to have a history of circulatory and respiratory system symptoms (Table 1).

Table 1. Baseline characteristics of patients according to hospital admission status.

Characteristic		Admission		*P* value
		No (n=2446)	Yes (n=2699)	
Age (years), mean (SD)		63.0 (23.7)	73.4 (16.2)	<.001
Male sex, n (%)		1308 (53.5)	1591 (58.9)	.25
Selected chief complaint, n (%)[a]				
Altered mental status		220 (9.0)	373 (13.8)	<.001
Dyspnea		176 (7.2)	403 (14.9)	.75
Chest pain		188 (7.7)	200 (7.4)	.54
Abdominal pain		169 (6.9)	174 (6.4)	<.001
Fever		97 (4.0)	175 (6.5)	<.001
Vital signs				
Level of consciousness, n (%)				
Alert		1619 (66.2)	1338 (49.6)	<.001
Possible eye opening, not lucid		427 (17.5)	642 (23.8)	<.001
Possible eye opening upon stimulation		85 (3.5)	228 (8.4)	<.001
No eye opening and coma		36 (1.5)	218 (8.1)	<.001
Systolic blood pressure (mm Hg), mean (SD)		148.0 (33.2)	140.7 (43.8)	<.001
Diastolic blood pressure (mm Hg), mean (SD)		83.5 (21.2)	80.8 (27.5)	<.001
Pulse rate (bpm), mean (SD)		88.7 (21.3)	92.7 (24.7)	<.001
Respiratory rate (bpm), mean (SD)		21.5 (5.3)	22.8 (6.2)	<.001
Body temperature (°C), mean (SD)		36.8 (7.6)	36.9 (2.7)	.76
Oxygen saturation (%), mean (SD)		97.0 (3.1)	93.8 (7.7)	<.001
Oxygen administration during transportation, n (%)		287 (11.7)	947 (35.1)	<.001
Selected past medical history, n (%)[a]				
R09	Other symptoms and signs involving the circulatory and respiratory systems	624 (25.5)	805 (29.8)	<.001
E11	Type 2 diabetes mellitus	419 (17.1)	582 (21.6)	<.001
I63	Cerebral infarction	202 (8.3)	303 (11.2)	.001
E78	Disorders of lipoprotein metabolism and other lipidemias	151 (6.2)	212 (7.9)	.02
I51	Complications and ill-defined descriptions of heart disease	159 (6.5)	180 (6.7)	.85

[a]The five most frequent chief complaints and past medical history items are shown.

Overall, the GBM model achieved the highest AUROCs and AUPRCs in models 3 and 4 (Tables 2 and 3). The most accurate model was GBM in model 3, with AUROC of 0.818 (95% CI 0.792-0.839), AUPRC of 0.831 (95% CI 0.804-0.855), sensitivity of 0.744 (95% CI 0.716-0.773), and specificity of 0.745 (95% CI 0.709-0.776) (Tables 2-4). The highest AUROC of logistic regression was 0.805 (95% CI 0.782-0.827) in model 3. It was lower in model 4: 0.750 (95% CI 0.720-0.774) (Figure 1). In models 2-4, precision-recall curve analysis showed superior performance of machine learning models compared to that of logistic regression among patients with higher risk of hospital admission (Figure 2). The lasso and GBM showed good calibration in all models (Figure 3). Hyperparameters of machine learning models are shown in Table S1 in Multimedia Appendix 1. The exclusion of mortality at the ED showed slightly lower predictive performance, with AUROC of 0.803 (95% CI 0.775-0.823) for GBM in model 3 (Tables S2-S4 in Multimedia Appendix 1).

Table 2. Areas under the receiver operating characteristic curve and 95% confidence intervals of hospital admission prediction models according to machine learning methods and prediction models.

Model type	Model 1[a]	Model 2[b]	Model 3[c]	Model 4[d]
Logistic regression	0.631 (0.602-0.657)	0.750 (0.723-0.774)	0.805 (0.782-0.827)	0.750 (0.720-0.774)
Lasso	0.631 (0.602-0.657)	0.755 (0.730-0.779)	0.817 (0.793-0.839)	0.811 (0.787-0.832)
Random forest	0.594 (0.567-0.619)	0.735 (0.710-0.763)	0.813 (0.786-0.834)	0.814 (0.786-0.833)
Gradient boosting machine	0.624 (0.598-0.652)	0.758 (0.734-0.783)	0.818 (0.792-0.839)	0.815 (0.788-0.833)

[a]Model 1: Age and sex.

[b]Model 2: Age, sex, and chief complaint.

[c]Model 3: Age, sex, chief complaint, and vital signs.

[d]Model 4: Age, sex, chief complaint, vital signs, and past medical history.

Table 3. Areas under the precision-recall curve and 95% confidence intervals of hospital admission prediction models according to machine learning models and predictor modalities.

Model type	Model 1[a]	Model 2[b]	Model 3[c]	Model 4[d]
Logistic regression	0.614 (0.578-0.653)	0.729 (0.700-0.766)	0.794 (0.764-0.827)	0.709 (0.667-0.744)
Lasso	0.614 (0.578-0.654)	0.766 (0.737-0.795)	0.829 (0.805-0.853)	0.820 (0.793-0.845)
Random forest	0.580 (0.550-0.615)	0.734 (0.703-0.770)	0.828 (0.802-0.853)	0.828 (0.801-0.851)
Gradient boosting machine	0.609 (0.580-0.647)	0.766 (0.734-0.799)	0.831 (0.804-0.855)	0.828 (0.803-0.852)

[a]Model 1: Age and sex.

[b]Model 2: Age, sex, and chief complaint.

[c]Model 3: Age, sex, chief complaint, and vital signs.

[d]Model 4: Age, sex, chief complaint, vital signs, and past medical history.

Table 4. Measures of predictive performance and 95% confidence intervals for prediction model 3 at optimal thresholds[a].

Model type	Sensitivity	Specificity	PPV[b]	NPV[c]	Accuracy
Logistic regression	0.760 (0.724-0.786)	0.731 (0.698-0.766)	0.752 (0.728-0.783)	0.737 (0.700-0.771)	0.746 (0.723-0.770)
Lasso	0.724 (0.684-0.751)	0.774 (0.742-0.800)	0.776 (0.748-0.803)	0.722 (0.683-0.752)	0.749 (0.723-0.767)
Random forest	0.720 (0.687-0.749)	0.777 (0.742-0.804)	0.776 (0.745-0.807)	0.718 (0.685-0.748)	0.745 (0.720-0.768)
Gradient boosting machine	0.736 (0.696-0.768)	0.743 (0.712-0.772)	0.757 (0.726-0.785)	0.721 (0.680-0.756)	0.739 (0.713-0.765)

[a]Predictors were age, sex, chief complaint, and vital signs.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

Figure 1. ROC curves of hospital admission prediction models. ROC curves for the three machine learning models are similar to those of logistic regression in models 1, 2, and 3, but superior to those of logistic regression in model 4. ROC: receiver operating characteristic.

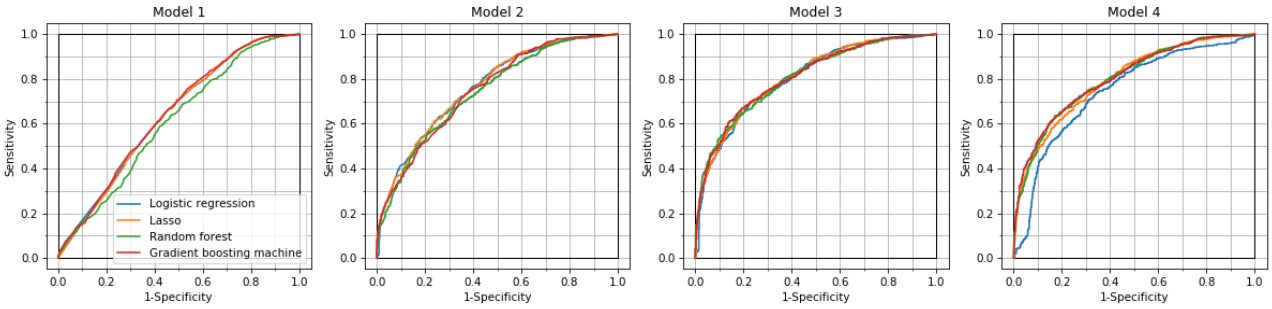


Figure 2. Precision-recall curves of hospital admission prediction models. Precision-recall curves of the three machine learning models are similar. Logistic regression model showed inferior performance for patients with higher predicted probabilities (left side on the horizontal axis) in models 2, 3, and 4.

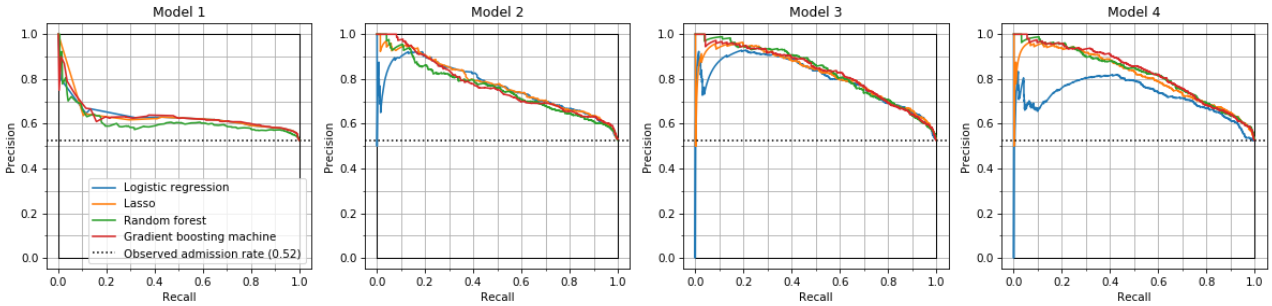
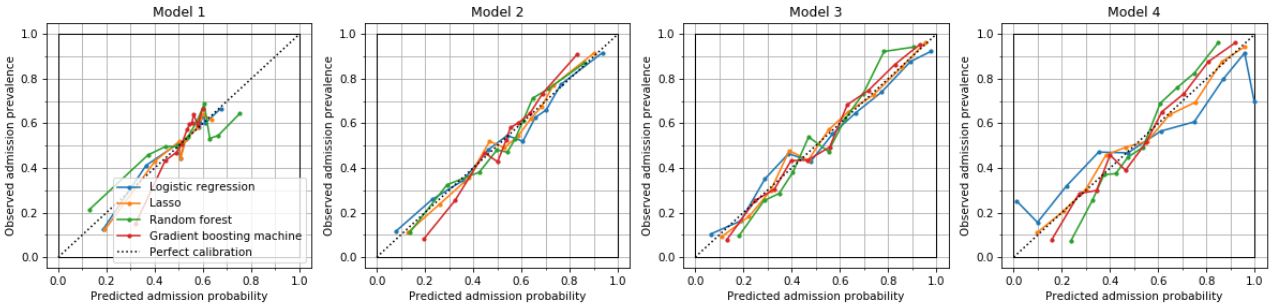


Figure 3. Calibration curves of hospital admission prediction models. Lasso and gradient boosting machine showed good calibration in all models. Logistic regression was ill-calibrated for patients with the lowest and the highest deciles of predicted probability in model 4.



A GBM model with data of both walk-in and ambulance visits to our ED during the study period (n=16,857) demonstrated higher performance than that for patients transported by ambulance, with AUROC of 0.873 (95% CI 0.860-0.883), sensitivity of 0.830 (95% CI 0.807-0.850), and specificity of 0.743 (95% CI 0.712-0.772) in the validation set.

Discussion

Principal Findings

To our knowledge, this report is the first of a study developing and validating prediction models for hospital admission based on common prehospital information for patients transported to EDs by ambulance. Information used for this study was collected in prehospital settings within a routine clinical practice. Therefore, the method of the prediction model development is readily applicable to other facilities that support clinical decision making by EMSs.

Our results are comparable to those presented in earlier reports describing the performance of subjective prediction by ambulance staff for patients they transported. A prospective study in the United Kingdom revealed a response rate of 99.7% (396/397). Analyses of 396 cases demonstrated sensitivity of 0.717 (95% CI 0.65-0.78) and specificity of 0.770 (95% CI 0.71-0.81) [24]. Another prospective study conducted in the United States found a response rate of 24.6% (101/411) from the cases analyzed [25]. Sensitivity of prediction by EMS staff members was 0.733 (95% CI 0.658-0.798), and the specificity was 0.850 (95% CI 0.798-0.891). Another study in the United States examined 932 transports to a hospital and reported the performance of EMS staff prediction of hospitalization as 0.62 (95% CI 0.54-0.68) for sensitivity and 0.89 (95% CI 0.86-0.91) for specificity [26]. However, prediction by EMS staff in this study was done at the time they left the ED. The results might be affected by incorporation bias because of observation or direct discussion with physicians and nurses in the ED. Therefore, the true performance might be lower. These studies

are based on the impressions of paramedics. Therefore, their performance in other emergency medical systems remains unknown. However, our method relies on common prehospital measurements, which present the benefit of applicability to other standard emergency medical systems.

The AUROC achieved using the proposed model was lower than those reported from earlier studies for patients after arrival at the ED, reporting values of 0.80-0.87 [3-11]. However, these earlier models included both walk-in and ambulance patients. Because our prediction model was restricted to patients transferred by ambulance, the target population might be more severely affected by health issues than walk-in patients, making it difficult to discriminate patients who need inpatient care and patients who do not. Indeed, prediction performance including both walk-in and ambulance visits to our ED demonstrated comparable performance to that of an AUROC of 0.873.

The logistic regression model demonstrated comparable performance to that obtained with other machine learning models, with <0.02 difference in AUROCs in models 1-3 and lower performance in model 4. Two recent reports have described similar predictive performance in logistic regression and machine learning models for predicting hospital admission after ED visits [39,40]. However, the ratios of the number of variables to the number of patients were smaller in those studies than in this study: previous studies reported 972 variables to 560,486 patients [39], and 111 variables to 1,721,294 patients [40], whereas this study reported 832 variables to 5145 patients. The lower predictive performance of logistic regression can be attributed to overfitting. By selecting important predictors by lasso or other methods, a logistic regression model might be built with comparable performance to those of other machine learning models, as suggested by our result obtained for lasso, which virtually reduces the number of variables in logistic regression.

Limitations

First, hospital admission might reflect not only the medical conditions, but also the social context. Performance can be improved by adding socioeconomic factors such as activities of daily living, education, income, type of insurance, family structure, and marital status, or neurological characteristics such as cognitive function and depressive symptoms, especially for elderly people [41-43]. Second, because the models were developed from data from a single institution, the external validity of our model is uncertain. For generalization of our results to other hospitals, assessments similar to ours are expected to be necessary. However, data used for this study can be collected automatically in daily routine practice. Therefore, development of a hospital-specific prediction model is feasible. For small hospitals with ED volume that is too small to generate a model, privacy-preserving federated learning [44,45] might provide a solution. Third, information on past medical history might be affected by information bias because it is collected in a critical situation. Nonsignificant incremental benefits of adding past medical history information in this study can be partially attributable to this bias. Accurate data collection of past medical history, for example, linkage to personal health care records in an integrated community health care network, might improve the model's predictive performance. Fourth, we did not have detailed information related to the accurate time of measurement of vital signs. Taking the best or worst value of vital signs may increase the predictive ability of our proposed models.

Conclusions

We developed a model of hospital admission prediction for patients transferred by ambulance using common prehospital information that performed well. The methodology used for this study can be extended to multicenter settings to facilitate efficient medical resource use in communities.

Acknowledgments

The authors are grateful to Dr Kei Taneishi at RIKEN Advanced Institute for Computational Science for valuable comments and discussion.

Conflicts of Interest

T Shirakawa, T Sonoo, KO, RF, TG, and KH are owners or employees of TXP Medical Co, Ltd.

Multimedia Appendix 1
Supplementary tables.
[DOCX File , 17 KB - medinform_v8i10e20324_app1.docx]

References

1. EMS Agenda 2050. U.S. Department of Transportation NHTSA. 2019. URL: https://www.ems.gov/projects/ems-agenda-2050.html [accessed 2020-10-15]
2. Asplin BR, Magid DJ, Rhodes KV, Solberg LI, Lurie N, Camargo CA. A conceptual model of emergency department crowding. Ann Emerg Med 2003 Aug;42(2):173-180. [doi: 10.1067/mem.2003.302] [Medline: 12883504]
3. Parker CA, Liu N, Wu SX, Shen Y, Lam SSW, Ong MEH. Predicting hospital admission at the emergency department triage: A novel prediction model. Am J Emerg Med 2019 Aug;37(8):1498-1504 [FREE Full text] [doi: 10.1016/j.ajem.2018.10.060] [Medline: 30413365]

4. Kraaijvanger N, Rijpsma D, Roovers L, van Leeuwen H, Kaasjager K, van den Brand L, et al. Development and validation of an admission prediction tool for emergency departments in the Netherlands. Emerg Med J 2018 Aug;35(8):464-470. [doi: 10.1136/emermed-2017-206673] [Medline: 29627769]
5. Barak-Corren Y, Israelit SH, Reis BY. Progressive prediction of hospitalisation in the emergency department: uncovering hidden patterns to improve patient flow. Emerg Med J 2017 May;34(5):308-314. [doi: 10.1136/emermed-2014-203819] [Medline: 28188202]
6. Sun Y, Heng BH, Tay SY, Seow E. Predicting hospital admissions at emergency department triage using routine administrative data. Acad Emerg Med 2011 Aug;18(8):844-850. [doi: 10.1111/j.1553-2712.2011.01125.x] [Medline: 21843220]
7. Lucke JA, de Gelder J, Clarijs F, Heringhaus C, de Craen AJM, Fogteloo AJ, et al. Early prediction of hospital admission for emergency department patients: a comparison between patients younger or older than 70 years. Emerg Med J 2018 Jan;35(1):18-27. [doi: 10.1136/emermed-2016-205846] [Medline: 28814479]
8. Rahimian F, Salimi-Khorshidi G, Payberah AH, Tran J, Ayala Solares R, Raimondi F, et al. Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. PLoS Med 2018 Nov;15(11):e1002695 [FREE Full text] [doi: 10.1371/journal.pmed.1002695] [Medline: 30458006]
9. Cameron A, Rodgers K, Ireland A, Jamdar R, McKay GA. A simple tool to predict admission at the time of triage. Emerg Med J 2015 Mar;32(3):174-179. [doi: 10.1136/emermed-2013-203200] [Medline: 24421344]
10. Peck JS, Gaehde SA, Nightingale DJ, Gelman DY, Huckins DS, Lemons MF, et al. Generalizability of a simple approach for predicting hospital admission from an emergency department. Acad Emerg Med 2013 Nov;20(11):1156-1163 [FREE Full text] [doi: 10.1111/acem.12244] [Medline: 24238319]
11. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care 2019 Feb 22;23(1):64 [FREE Full text] [doi: 10.1186/s13054-019-2351-7] [Medline: 30795786]
12. Bernstein SL, 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 [FREE Full text] [doi: 10.1111/j.1553-2712.2008.00295.x] [Medline: 19007346]
13. Fatovich DM, Nagree Y, Sprivulis P. Access block causes emergency department overcrowding and ambulance diversion in Perth, Western Australia. Emerg Med J 2005 May;22(5):351-354. [doi: 10.1136/emj.2004.018002] [Medline: 15843704]
14. McCarthy ML, Zeger SL, Ding R, Levin SR, Desmond JS, Lee J, et al. Crowding delays treatment and lengthens emergency department length of stay, even among high-acuity patients. Ann Emerg Med 2009 Oct;54(4):492-503.e4. [doi: 10.1016/j.annemergmed.2009.03.006] [Medline: 19423188]
15. Schull MJ, Kiss A, Szalai J. The effect of low-complexity patients on emergency department waiting times. Ann Emerg Med 2007 Mar;49(3):257-64, 264.e1. [doi: 10.1016/j.annemergmed.2006.06.027] [Medline: 17049408]
16. Uchida K, Yoshimura S, Hiyama N, Oki Y, Matsumoto T, Tokuda R, et al. Clinical Prediction Rules to Classify Types of Stroke at Prehospital Stage. Stroke 2018 Aug;49(8):1820-1827 [FREE Full text] [doi: 10.1161/STROKEAHA.118.021794] [Medline: 30002147]
17. Peltan ID, Rowhani-Rahbar A, Vande Vusse LK, Caldwell E, Rea TD, Maier RV, et al. Development and validation of a prehospital prediction model for acute traumatic coagulopathy. Crit Care 2016 Nov 16;20(1):371 [FREE Full text] [doi: 10.1186/s13054-016-1541-9] [Medline: 27846895]
18. Pirneskoski J, Kuisma M, Olkkola KT, Nurmi J. Prehospital National Early Warning Score predicts early mortality. Acta Anaesthesiol Scand 2019 May;63(5):676-683. [doi: 10.1111/aas.13310] [Medline: 30623422]
19. Koyama S, Yamaguchi Y, Gibo K, Nakayama I, Ueda S. Use of prehospital qSOFA in predicting in-hospital mortality in patients with suspected infection: A retrospective cohort study. PLoS One 2019;14(5):e0216560 [FREE Full text] [doi: 10.1371/journal.pone.0216560] [Medline: 31063494]
20. Williams TA, Tohira H, Finn J, Perkins GD, Ho KM. The ability of early warning scores (EWS) to detect critical illness in the prehospital setting: A systematic review. Resuscitation 2016 May;102:35-43. [doi: 10.1016/j.resuscitation.2016.02.011] [Medline: 26905389]
21. Kievlan DR, Martin-Gill C, Kahn JM, Callaway CW, Yealy DM, Angus DC, et al. External validation of a prehospital risk score for critical illness. Crit Care 2016 Aug 11;20(1):255 [FREE Full text] [doi: 10.1186/s13054-016-1408-0] [Medline: 27515164]
22. Jouffroy R, Saade A, Ellouze S, Carpentier A, Michaloux M, Carli P, et al. Prehospital triage of septic patients at the SAMU regulation: Comparison of qSOFA, MRST, MEWS and PRESEP scores. Am J Emerg Med 2018 May;36(5):820-824. [doi: 10.1016/j.ajem.2017.10.030] [Medline: 29056391]
23. van Rein EAJ, van der Sluijs R, Voskens FJ, Lansink KWW, Houwert RM, Lichtveld RA, et al. Development and Validation of a Prediction Model for Prehospital Triage of Trauma Patients. JAMA Surg 2019 May 01;154(5):421-429 [FREE Full text] [doi: 10.1001/jamasurg.2018.4752] [Medline: 30725101]
24. Price TG, Hooker EA, Neubauer J. Prehospital provider prediction of emergency department disposition: implications for selective diversion. Prehosp Emerg Care 2005;9(3):322-325. [doi: 10.1080/10903120590962012] [Medline: 16147483]

25. Levine SD, Colwell CB, Pons PT, Gravitz C, Haukoos JS, McVaney KE. How well do paramedics predict admission to the hospital? A prospective study. J Emerg Med 2006 Jul;31(1):1-5. [doi: 10.1016/j.jemermed.2005.08.007] [Medline: 16798145]
26. Clesham K, Mason S, Gray J, Walters S, Cooke V. Can emergency medical service staff predict the disposition of patients they are transporting? Emerg Med J 2008 Oct;25(10):691-694. [doi: 10.1136/emj.2007.054924] [Medline: 18843076]
27. Spangler D, Hermansson T, Smekal D, Blomberg H. A validation of machine learning-based risk scores in the prehospital setting. PLoS One 2019;14(12):e0226518 [FREE Full text] [doi: 10.1371/journal.pone.0226518] [Medline: 31834920]
28. Sonoo T, Naraba H, Hashimotom H, Nakamura K, Morimura N. Development and evaluation of computer system that enables emergency department efficiency improvement and collection of coded data during normal clinical workflow. J Japanese Assoc Acute Med 2018;29(2):45-55. [doi: 10.1002/jja2.12276]
29. Kuriyama A, Ikegami T, Kaihara T, Fukuoka T, Nakayama T. Validity of the Japan Acuity and Triage Scale in adults: a cohort study. Emerg Med J 2018 Jun;35(6):384-388. [doi: 10.1136/emermed-2017-207214] [Medline: 29535086]
30. Bullard MJ, Musgrave E, Warren D, Unger B, Skeldon T, Grierson R, et al. Revisions to the Canadian Emergency Department Triage and Acuity Scale (CTAS) Guidelines 2016. CJEM 2017 Jul;19(S2):S18-S27. [doi: 10.1017/cem.2017.365] [Medline: 28756800]
31. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. World Health Organization. 1992. URL: http://www.who.int/classifications/icd/en/bluebook.pdf [accessed 2020-10-15]
32. Shigematsu K, Nakano H, Watanabe Y. The eye response test alone is sufficient to predict stroke outcome--reintroduction of Japan Coma Scale: a cohort study. BMJ Open 2013;3(4). [doi: 10.1136/bmjopen-2013-002736] [Medline: 23633419]
33. Breiman L. Random forests. Mach Learn 2001;45(1):5-32. [doi: 10.1023/A:1010933404324]
34. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat 2001 Oct;29(5):1189-1232. [doi: 10.1214/aos/1013203451]
35. 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. New York, New York, USA: Association for Computing Machinery; 2016 Presented at: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, California, USA p. 785-794. [doi: 10.1145/2939672.2939785]
36. Youden WJ. Index for rating diagnostic tests. Cancer 1950 Jan;3(1):32-35. [Medline: 15405679]
37. 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]
38. Pollard TJ, Johnson AEW, Raffa JD, Mark RG. tableone: An open source Python package for producing summary statistics for research papers. JAMIA Open 2018 Jul;1(1):26-31 [FREE Full text] [doi: 10.1093/jamiaopen/ooy012] [Medline: 31984317]
39. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. PLoS One 2018;13(7):e0201016 [FREE Full text] [doi: 10.1371/journal.pone.0201016] [Medline: 30028888]
40. Rendell K, Koprinska I, Kyme A, Ebker-White AA, Dinh MM. The Sydney Triage to Admission Risk Tool (START2) using machine learning techniques to support disposition decision-making. Emerg Med Australas 2019 Jun;31(3):429-435. [doi: 10.1111/1742-6723.13199] [Medline: 30469164]
41. Landi F, Onder G, Cesari M, Barillaro C, Lattanzio F, Carbonin PU, et al. Comorbidity and social factors predicted hospitalization in frail elderly patients. J Clin Epidemiol 2004 Aug;57(8):832-836. [doi: 10.1016/j.jclinepi.2004.01.013] [Medline: 15551473]
42. Clay OJ, Roth DL, Safford MM, Sawyer PL, Allman RM. Predictors of overnight hospital admission in older African American and Caucasian Medicare beneficiaries. J Gerontol A Biol Sci Med Sci 2011 Aug;66(8):910-916 [FREE Full text] [doi: 10.1093/gerona/glr082] [Medline: 21565981]
43. Amegbor PM, Plumb KB, Rosenberg MW. Determinants of Overnight Stay in Health Centres and Length of Admission: A Study of Canadian Seniors. Can J Aging 2020 Feb 24:1-12. [doi: 10.1017/S0714980819000771] [Medline: 32089138]
44. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med 2010 Nov 10;2(57):57cm29. [doi: 10.1126/scitranslmed.3001456] [Medline: 21068440]
45. Konečný J, McMahan H, Ramage D, Richtárik P. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. arXiv. 2016. URL: http://arxiv.org/abs/1610.02527 [accessed 2020-10-15]

Abbreviations

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
ED: emergency department
EMS: emergency medical service
GBM: gradient boosting machine
ICD-10: International Statistical Classification of Diseases, 10th Revision
JTAS: Japan Triage and Acuity Scale

XGBoost: extreme gradient boosting

Edited by G Eysenbach; submitted 04.06.20; peer-reviewed by R Lieu, Y Katayama; comments to author 29.06.20; revised version received 24.08.20; accepted 16.09.20; published 27.10.20.

Please cite as:
Shirakawa T, Sonoo T, Ogura K, Fujimori R, Hara K, Goto T, Hashimoto H, Takahashi Y, Naraba H, Nakamura K
Institution-Specific Machine Learning Models for Prehospital Assessment to Predict Hospital Admission: Prediction Model Development Study
JMIR Med Inform 2020;8(10):e20324
URL: http://medinform.jmir.org/2020/10/e20324/
doi:10.2196/20324
PMID:33107830

Original Paper

Prognostic Machine Learning Models for First-Year Mortality in Incident Hemodialysis Patients: Development and Validation Study

Kaixiang Sheng[1*], MD; Ping Zhang[1*], MD; Xi Yao[1], MD; Jiawei Li[1], BA; Yongchun He[1], BA; Jianghua Chen[1], MD

Kidney Disease Center, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Jianghua Chen, MD
Kidney Disease Center
The First Affiliated Hospital, Zhejiang University School of Medicine
#79 Qingchun Road
Hangzhou, 310003
China
Phone: 86 57187236992
Email: zjukidney@zju.edu.cn

Abstract

Background: The first-year survival rate among patients undergoing hemodialysis remains poor. Current mortality risk scores for patients undergoing hemodialysis employ regression techniques and have limited applicability and robustness.

Objective: We aimed to develop a machine learning model utilizing clinical factors to predict first-year mortality in patients undergoing hemodialysis that could assist physicians in classifying high-risk patients.

Methods: Training and testing cohorts consisted of 5351 patients from a single center and 5828 patients from 97 renal centers undergoing hemodialysis (incident only). The outcome was all-cause mortality during the first year of dialysis. Extreme gradient boosting was used for algorithm training and validation. Two models were established based on the data obtained at dialysis initiation (model 1) and data 0-3 months after dialysis initiation (model 2), and 10-fold cross-validation was applied to each model. The area under the curve (AUC), sensitivity (recall), specificity, precision, balanced accuracy, and F1 score were used to assess the predictive ability of the models.

Results: In the training and testing cohorts, 585 (10.93%) and 764 (13.11%) patients, respectively, died during the first-year follow-up. Of 42 candidate features, the 15 most important features were selected. The performance of model 1 (AUC 0.83, 95% CI 0.78-0.84) was similar to that of model 2 (AUC 0.85, 95% CI 0.81-0.86).

Conclusions: We developed and validated 2 machine learning models to predict first-year mortality in patients undergoing hemodialysis. Both models could be used to stratify high-risk patients at the early stages of dialysis.

(JMIR Med Inform 2020;8(10):e20578) doi:10.2196/20578

KEYWORDS

machine learning; hemodialysis; XGBoost; prediction model

Introduction

Background

The overall prevalence of chronic kidney disease is 10.8% in China and 15% in the United States, which has brought significant economic, social, and medical burdens on patients and society [1-3]. According to the United States Renal Data System, there are approximately 120,000 patients with end-stage renal disease starting chronic renal replacement therapy every year [2]. However, survival among incident hemodialysis patients remains poor, especially in the first year of the initiation of dialysis [4,5].

End-stage renal disease is a complex disease state with multiple associated comorbidities. Patients initiating hemodialysis often have acute complications, and some of them suffer from major comorbid conditions that are associated with poor short-term prognoses [6]. It is essential to stratify the risk of mortality according to clinical and laboratory findings of patients undergoing hemodialysis; therefore, the identification of patients undergoing hemodialysis who are at high risk of first-year mortality is of great clinical significance. It can inform patients

of their survival prognosis in the early stages of dialysis and allow clinicians to make targeted intervention strategies to improve first-year outcomes. Previous studies [7-11] have identified many risk factors for early dialysis mortality, such as old age, chronic heart failure, catheter use, low albumin, low hemoglobin, and high estimated glomerular filtration rate at dialysis initiation. However, because of the heterogeneity of primary disorders and broad comorbidities, these risk factors are not enough to be used for conclusive decision making. In recent years, a number of clinical risk models have been developed to predict early mortality in the dialysis population, and most are based on linear models (logistic or Cox model) [12-16]. The performances of these models were not good enough in either the original population or the external validation—area under the curve (AUC) of these models ranged from 0.710 to 0.752 [17]. In addition, no study compared models based on predialysis data with models based on data after dialysis.

In recent years, machine learning has been proven to be a very powerful method by researchers in medical fields [18-21]. Machine learning is useful in identifying the most important factors and for developing predictive models with the best performance. A recent study [22] reported on a random forest machine learning model used to predict first-year survival of incident hemodialysis patients. The model's AUC was 0.749 (95% CI 0.742-0.755), which was superior to those of traditional risk prediction models; however, this is not accurate enough for clinical application.

Objective

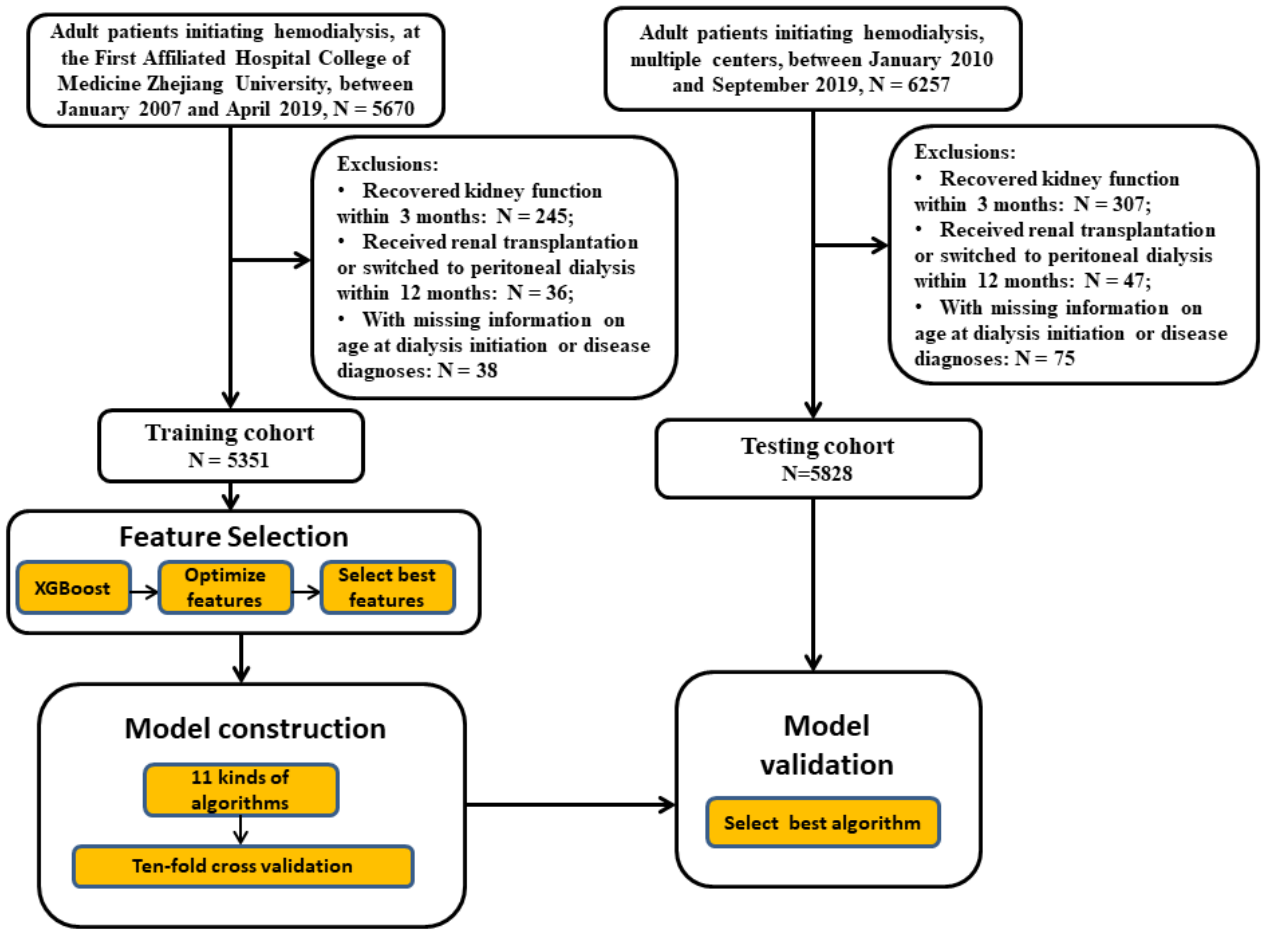
Therefore, in this study, we sought to develop and validate sufficiently accurate models based on machine learning techniques, utilizing readily available clinical factors to predict first-year mortality in incident dialysis patients.

Methods

Study Design

This study retrospectively collected data from Zhejiang Dialysis System. Zhejiang Dialysis System is a database of hemodialysis and peritoneal dialysis patients in East China. Training data were retrieved from the First Affiliated Hospital College of Medicine Zhejiang University between January 2007 and April 2019 (Figure 1). Testing data were collected from 97 renal centers between January 2010 and August 2018 for external validation (Figure 1). All follow-up data were updated to August 2019.

Figure 1. A workflow to develop the prediction models for first-year mortality in incident hemodialysis patients. XGBoost: extreme gradient boosting.



Adult patients (aged ≥18 years) with end-stage renal disease and with follow-up exceeding 12 months who started maintenance hemodialysis were included. Patients who died within 12 months of follow-up were also included.

The exclusion criteria were as follows: patients with a history of previous renal replacement therapy, patients whose kidney function recovered within 3 months, patients who received renal transplantation or switched to peritoneal dialysis within 12 months after dialysis initiation. We also excluded patients with missing information on disease diagnoses or age at dialysis initiation.

This study followed the tenets of the Declaration of Helsinki and was approved by the ethics committee of the First Affiliated Hospital of Zhejiang University (IIT20200088A) in Hangzhou, China. Written informed consent was obtained from each participant.

Outcome and Predictors

The outcome of this study was all-cause mortality during the first year of dialysis. Outcome status and potential candidate variables for the prediction tool, including demographic information, disease diagnoses, comorbidities, and laboratory test results, were obtained from the Zhejiang Dialysis System.

Demographic information and type of vascular access were collected at the start of dialysis. Disease diagnoses, comorbid information, and laboratory test results were collected 0-3 months after dialysis initiation. The most recent serum creatinine measurements prior to the index date were used to estimate the glomerular filtration rate using the Chronic Kidney Disease Epidemiology Collaboration equation [23].

A total of 42 variables were included as candidate features based on review of relevant literature and clinical experience. Only BMI and ferritin had missing data, and both instances of missing data were less than 6% (Table 1).

Table 1. Baseline characteristics of the training and testing cohorts.

Characteristics	At dialysis initiation		0-3 months	
	Training cohort (n=5351)	Testing cohort (n=5828)	Training cohort (n=4425)	Testing cohort (n=3729)
Sex, n (%)				
Male	3295 (61.58)	3524 (60.47)	2744 (62.01)	2264 (60.71)
Female	2056 (38.42)	2304 (39.53)	1681 (37.99)	1465 (39.29)
Body mass index (kg/m^2), mean (SD)[a]	22.09 (3.29)	21.73 (3.07)	22.19 (3.39)	21.83 (3.04)
Age at dialysis initiation (years), mean (SD)	51.67 (16.48)	62.53 (16.20)	52.61 (16.59)	62.45 (15.9)
Systolic pressure (mmHg), mean (SD)	137.49 (22.93)	146.18 (24.58)	138.52 (23.15)	146.33 (24.68)
Diastolic pressure (mmHg), mean (SD)	77.76 (12.26)	78.95 (15.52)	80.45 (12.15)	79.02 (15.45)
Chronic kidney disease etiology, n (%)				
Chronic glomerulonephritis	2823 (52.76)	3015 (51.73)	2445 (55.25)	2064 (55.35)
Diabetic nephropathy	1120 (20.93)	1191 (20.44)	895 (20.23)	818 (21.94)
Hypertensive nephropathy	262 (4.90)	557 (9.56)	218 (4.93)	370 (9.92)
Lupus nephritis	68 (1.27)	50 (0.86)	57 (1.29)	29 (0.78)
ANCA-associated[b] vasculitis	57 (1.07)	64 (1.10)	53 (1.20)	33 (0.88)
Gouty nephropathy	32 (0.60)	125 (2.14)	26 (0.59)	72 (1.93)
Polycystic kidney disease	286 (5.34)	214 (3.67)	220 (4.97)	150 (4.02)
Other	703 (13.14)	612 (11.07)	511 (11.54)	204 (5)
Comorbid conditions, n (%)				
Cirrhosis	86 (1.61)	90 (1.54)	81 (1.83)	60 (1.61)
Multiple myeloma	46 (0.86)	90 (1.54)	46 (1.04)	51 (1.37)
Atrial fibrillation	108 (2.02)	109 (1.87)	85 (1.92)	72 (1.93)
Congestive heart failure	969 (18.11)	999 (17.14)	794 (17.94)	605 (16.22)
Ischemic heart disease	1476 (27.58)	1578 (27.08)	1206 (27.25)	983 (26.36)
Metastatic cancer	86 (1.61)	91 (1.56)	74 (1.67)	38 (1.02)
Lymphoma	7 (0.13)	7 (0.12)	6 (0.14)	1 (0.03)
Chronic obstructive pulmonary disease	241 (4.50)	165 (2.83)	169 (3.82)	78 (2.09)
Cerebrovascular disease	322 (6.02)	411 (7.05)	244 (5.51)	271 (7.27)
Laboratory data				
Leukocyte (10^9/L), mean (SD)	7.32 (2.95)	7.71 (3.79)	7.40 (3.09)	6.90 (3.22)
Neutrophil (10^9/L), mean (SD)	5.23 (2.68)	5.06 (3.32)	5.36 (2.78)	4.22 (2.57
Hemoglobin (g/L), mean (SD)	94.82 (23.30)	83.09 (19.12)	91.05 (21.68)	86.50 (14.67)
Platelet (10^9/L), mean (SD)	193.28 (93.47)	182.47 (83.70)	190.84 (88.13)	184.36 (71.39)
Albumin (g/L), mean (SD)	36.01 (6.75)	33.27 (5.99)	36.80 (6.59)	33.98 (5.54)
Phosphorus (mmol/L), mean (SD)	1.81 (0.62)	1.70 (0.66)	1.66 (0.52)	1.54 (0.50)
Calcium (mmol/L), mean (SD)	2.15 (0.28)	2.02 (0.30)	2.14 (0.22)	2.08 (0.23)
Potassium (mmol/L)	4.87 (1.11)	4.52 (0.91)	4.76 (0.96)	4.42 (0.69)
Parathyroid hormone (pg/ml), mean (SD)	334.71 (292.07)	246.95 (193.61)	315.98 (291.84)	241.26 (206.48)
Creatinine (μmol/L), mean (SD)	807.11 (352.04)	718.84 (336.47)	755.28 (315.95)	661.5 (268.48)
Urea nitrogen (mmol/L), mean (SD)	22.65 (12.07)	23.61 (11.77)	19.87 (8.72)	20.01 (8.13)
Uric acid (μmol/L), mean (SD)	436.84 (147.54)	450.27 (157.44)	392.87 (126.48)	402.19 (113.46)

Characteristics	At dialysis initiation		0-3 months	
	Training cohort (n=5351)	Testing cohort (n=5828)	Training cohort (n=4425)	Testing cohort (n=3729)
C-reactive protein, mean (SD)	40.84 (44.09)	25.65 (44.46)	18.52 (35.01)	20.23 (31.22)
Cholesterol (mmol/L), mean (SD)	4.34 (1.30)	4.30 (1.42)	4.27 (1.23)	4.34 (1.25)
Triglycerides (mmol/L), mean (SD)	1.56 (1.00)	1.60 (1.03)	1.58 (0.96)	1.63 (0.97)
High-density lipoprotein, (mmol/L), mean (SD)	1.14 (0.42)	1.11 (0.43)	1.12 (0.39)	1.15 (0.38)
Low-density lipoprotein (mmol/L), mean (SD)	2.36 (1.10)	2.37 (1.02	2.31 (1.04)	2.35 (0.92)
Very low-density lipoprotein (mmol/L), mean (SD)	1.65 (1.55)	2.11 (1.35)	1.63 (1.54)	1.60 (0.93)
Ferritin (ng/mL), mean (SD)[c]	174.59 (126.34)	328.25 (295.78)	144.34 (144.87)	305.42 (278.73)
eGFR[d] (mL/min/1.73m^2), mean (SD)	6.75 (3.79)	7.28 (3.93)	7.23 (3.85)	7.58 (3.44)
Vascular access at dialysis initiation, n (%)				
Nontunneled catheter	3295 (61.58)	3388 (58.13)	2495 (56.38)	1893 (50.76)
Tunneled catheter	1068 (19.96)	1266 (21.72)	1005 (22.71)	938 (25.15)
Fistula or graft	988 (18.46)	1174 (20.14)	925 (20.90)	898 (24.08)
Death at 1-year follow-up, n (%)	585 (10.93)	764 (13.11)	437 (9.88)	477 (12.79)

[a]The missing rates of body mass index in the 4 cohorts were 270 (5.04%), 298 (5.11%), 210 (4.74%), and 168 (4.50%), respectively.

[b]ANCA: antineutrophil cytoplasmic antibody.

[c]The missing rates of ferritin in the 4 cohorts were 0.36%, 3.00%, 0.36%, and 2.13%, respectively.

[d]eGFR: estimated glomerular filtration rate.

Data Preprocessing

Before the baseline model was developed, missing data were imputed with the mean value for continuous variables and the mode value for categorical variables. By using one-hot encoding, all categorical features were transformed into numerical features. Box-Cox transformation was performed to normalize numerical features that were highly skewed [24].

Algorithm Development and Validation

An extreme gradient boosting machine learning algorithm was employed to build a model to predict the correlation between features and the outcome. Extreme gradient boosting is an integrated learning algorithm based on gradient boosted decision trees [25]. Using the Gini impurity index [26], we estimated the feature importance scores of candidate features after going through the training process. The feature importance scores showed how valuable each feature was in the construction of the boosted decision trees within the model.

The extreme gradient boosting algorithm was employed because (1) it has high efficiency and accuracy, (2) it can prevent overfitting via regularization, (3) it provides feature importance, and (4) it allows the use of a wide variety of computing environments.

Other popular machine learning algorithms—adaptive boosting, light gradient boosting machine, logistic regression, linear discriminant analysis, random forest, extra trees, gradient boosting, multiple layers perception, k-nearest neighbor, and decision trees—were compared with extreme gradient boosting.

We developed 2 models that were based on the data obtained at dialysis initiation (model 1) and data 0-3 months after dialysis initiation (model 2); 10-fold cross-validation was used to avoid overfitting and to validate each model [27]. We measured AUC, sensitivity (recall), specificity, precision, balanced accuracy, and F1 score to assess the predictive ability of each model. The balanced accuracy was calculated as follows: balanced accuracy = (sensitivity + specificity) / 2. The F1 score were calculated as follows: F1 score = (2 × precision × recall) / (precision + recall). Shapley additive explanation (SHAP) values were used to measure the marginal contribution of each feature to the models [28].

Results

Demographic and Clinical Characteristics

The demographic and clinical characteristics of the training and testing cohorts indicated that most characteristics were similarly distributed (Table 1). All patients were Chinese. The mean ages at dialysis initiation were 51.67 years (SD 16.48) in the training cohort and 62.53 years (SD 16.20) in the testing cohort; 61.58% of the patients (3295/5351) in the training cohort and 60.47% of the patients (3524/5828) in the testing cohort were men; out of 5351 patients, 585 (10.93%) deaths were reported in the training cohort, and out of 5828 patients, 764 (13.11%) deaths were reported in the testing cohort.

Model Performance

The ranks of features selected after training the extreme gradient boosting models are shown in Multimedia Appendix 1 and Multimedia Appendix 2. The same 15 most important features were chosen for both model 1 and model 2: age at dialysis initiation, vascular access, metastatic cancer, diabetic nephropathy, congestive heart failure, ischemic heart disease,

cerebrovascular disease, albumin, hemoglobin, neutrophil, C-reactive protein, creatinine, estimated glomerular filtration rate, systolic blood pressure, and BMI.

Among the 11 algorithms applied (Table 2), the extreme gradient boosting algorithm had the best generalized performance for both model 1 (AUC 0.83, 95% CI 0.78-0.84; balanced accuracy 84.52%; F1 score 0.75) and model 2 (AUC 0.85, 95% CI 0.81-0.86, balanced accuracy 89.21%, F1 score 0.78). As shown in Figure 2, the receiver operating characteristic curves of both models were similar.

SHAP value results are shown in Figure 3 (model 1) and Figure 4 (model 2). Each point represents a data sample for the feature. History of congestive heart failure, albumin level, C-reactive protein level, and age at dialysis initiation were the most important factors affecting the prediction for first-year mortality in both model 1 and model 2. Figure 5 shows an example using model 2 that shows how features contribute to the probability for a single participant. This participant had a history of congestive heart failure, low creatinine level, a high C-reactive protein level, high neutrophil count, and old age at dialysis initiation, which contributed to a higher probability of mortality in the first year, although he had normal BMI and slightly high systolic blood pressure levels.

Table 2. Performance of different algorithms trained on the testing data set.

Models	Precision, %	Sensitivity, %	Specificity, %	F1 score	Balanced accuracy, %	AUC[a] (95% CI)	Accuracy, %
Model 1							
Adaptive boosting	43.34	55.37	89.29	0.4862	72.33	0.81 (0.77-0.82)	84.92
Decision tree	68.61	35.47	97.55	0.4676	66.51	0.78 (0.76-0.80)	89.41
Extra trees	78.56	59.95	97.53	0.6800	78.74	0.83 (0.77-0.83)	92.60
Gradient boosting	52.58	49.35	93.29	0.5091	71.32	0.82 (0.77-0.83)	87.53
k-nearest neighbor	47.32	50.92	91.45	0.4905	71.18	0.76 (0.76-0.84)	86.14
Linear discriminant analysis	14.02	82.46	23.74	0.2397	53.10	0.75 (0.74-0.84)	31.43
Light gradient boosting	91.76	62.70	99.15	0.7449	80.92	0.82 (0.77-0.83)	94.37
Logistic regression	14.16	85.47	21.84	0.2430	53.66	0.68 (0.68-0.85)	30.18
Multiple layers perception	16.64	78.80	40.44	0.2748	59.62	0.80 (0.68-0.85)	45.47
Random forest	90.62	40.45	99.37	0.5593	69.91	0.81 (0.78-0.83)	91.64
Extreme gradient boosting	79.34	71.86	97.18	0.7541	84.52	0.83 (0.78-0.84)	93.86
Model 2							
Adaptive boosting	61.83	72.33	93.45	0.6667	82.89	0.83 (0.80-0.84)	90.75
Decision tree	78.50	63.52	97.45	0.7022	80.48	0.81 (0.80-0.82)	93.11
Extra trees	74.48	60.59	96.96	0.6682	78.77	0.84 (0.80-0.85)	92.30
Gradient boosting	83.08	67.92	97.97	0.7474	82.95	0.84 (0.82-0.85)	94.13
k-nearest neighbor	87.37	52.20	98.89	0.6535	75.55	0.82 (0.81-0.86)	92.92
Linear discriminant analysis	16.33	82.81	37.76	0.2728	60.29	0.76 ()0.76-0.86	43.52
Light gradient boosting	77.97	75.68	96.86	0.7681	86.27	0.85 (0.80-0.85)	94.15
Logistic regression	16.12	81.76	37.58	0.2692	59.67	0.73 (0.73-0.86)	43.23
Multiple layers perception	16.19	80.08	39.21	0.2694	59.65	0.71 (0.71-0.86)	44.44
Random forest	66.67	70.02	94.86	0.6830	82.44	0.82 (0.80-0.85)	91.69
Extreme gradient boosting	78.95	78.62	96.92	0.7878	87.77	0.85 (0.81-0.86)	94.58

[a]AUC: area under the curve.

Figure 2. Receiver-operating characteristic curves of model 1 and model 2. AUC: the area under the curve.

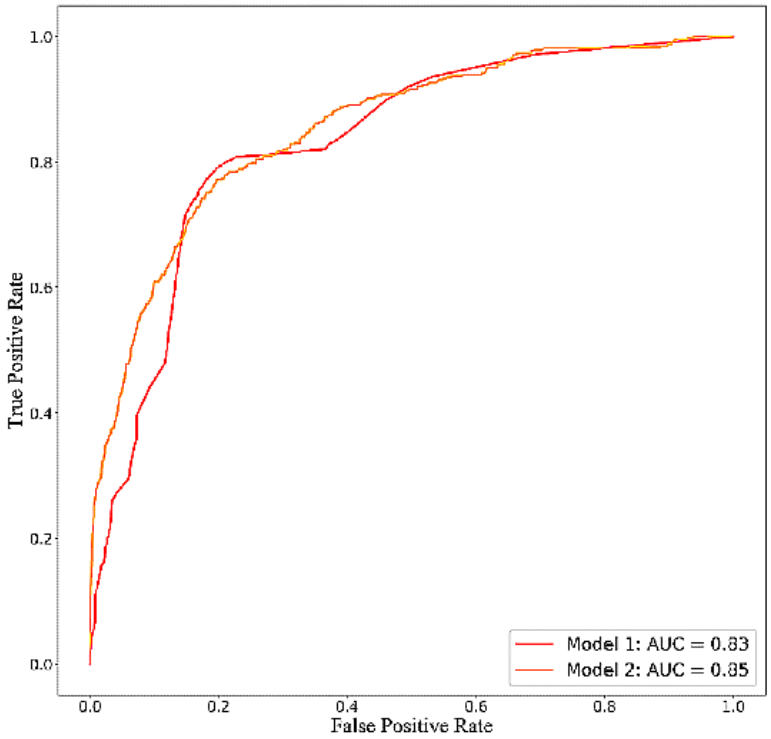


Figure 3. SHAP values illustrating how features contribute to model 1. Blue shows a negative contribution, and red shows a positive contribution. SHAP: Shapley additive explanation.

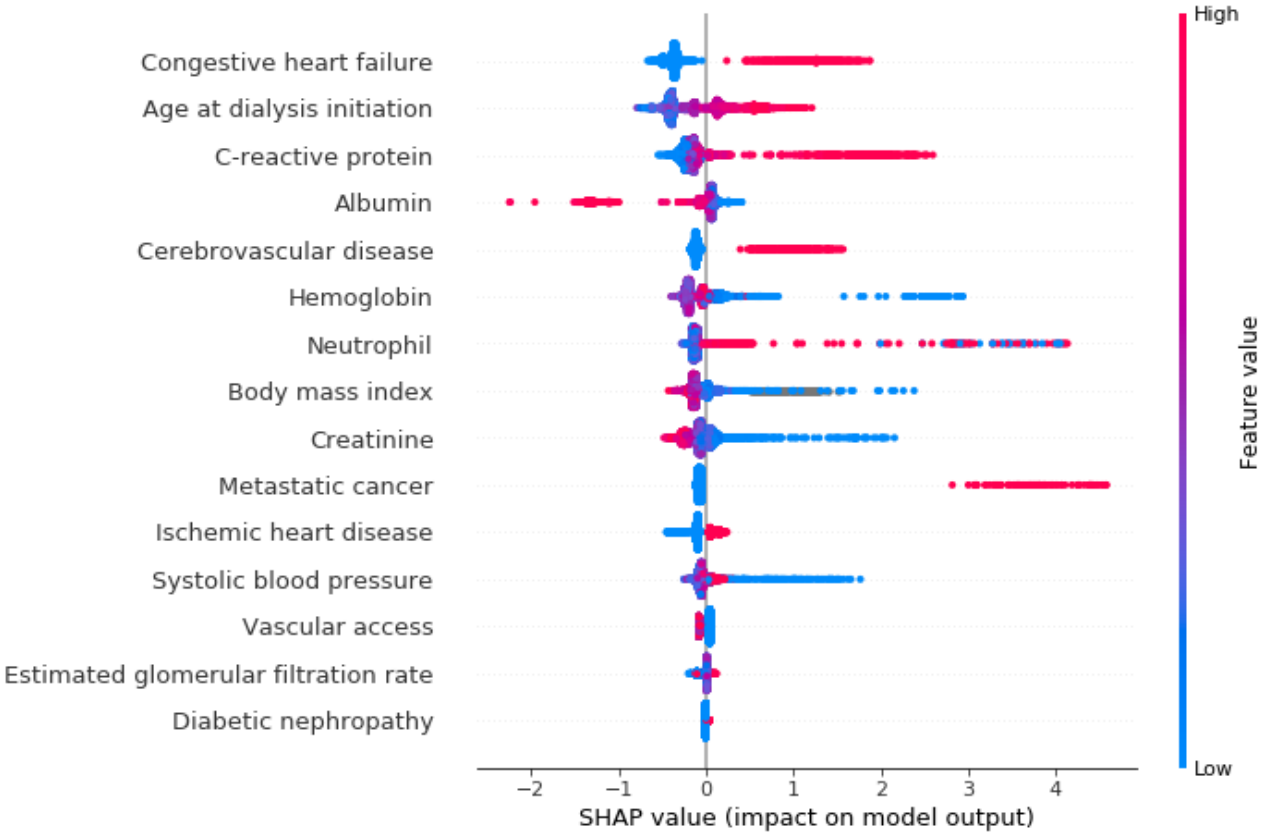


Figure 4. SHAP values illustrating how features contribute to model 2. Blue shows a negative contribution, and red shows a positive contribution. SHAP: Shapley additive explanation.

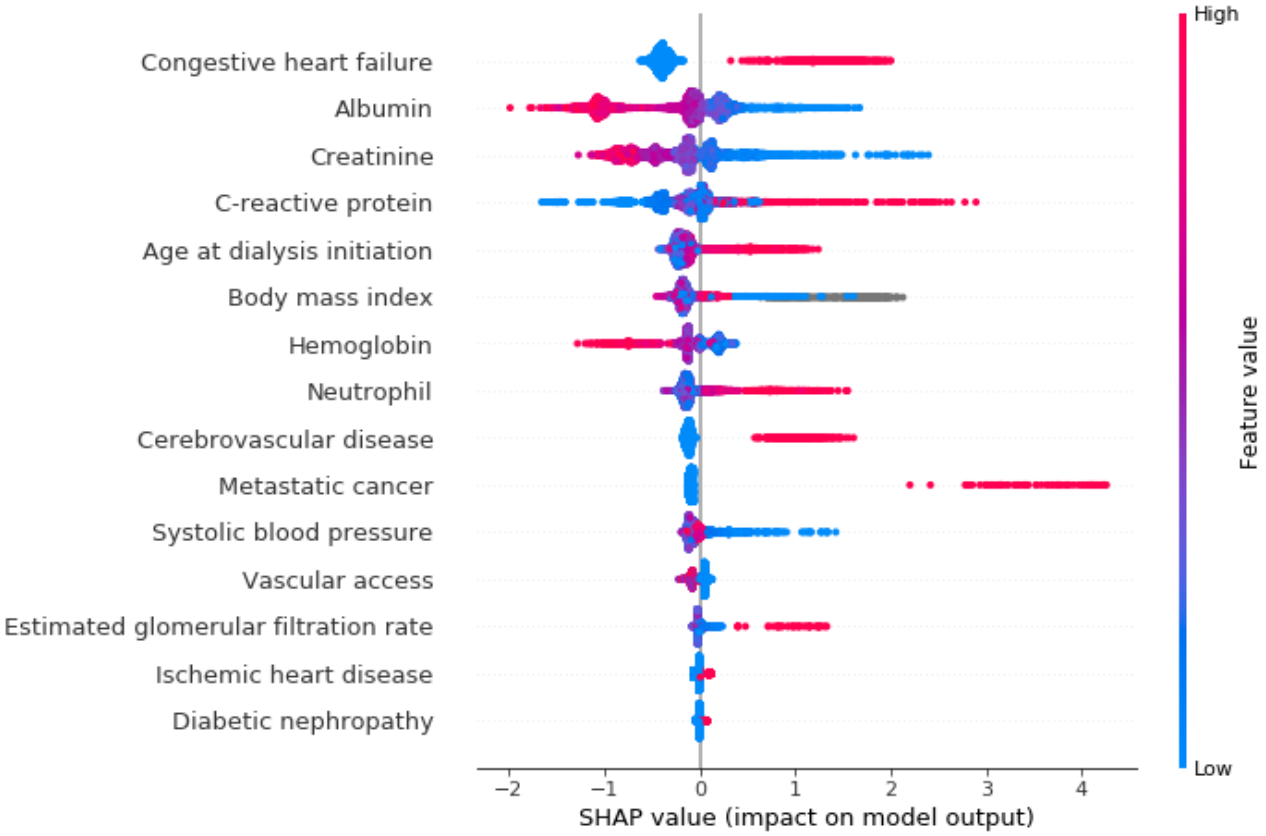
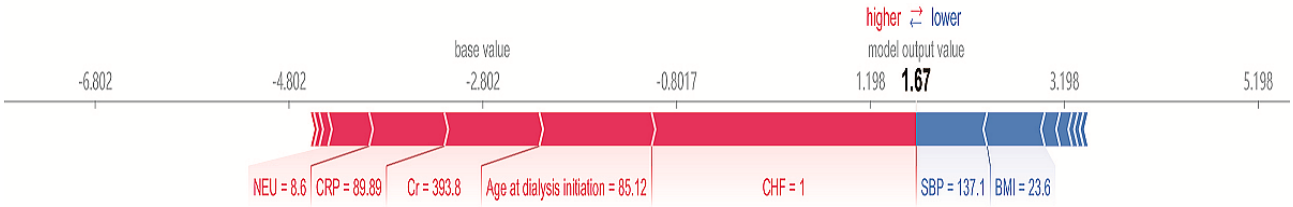


Figure 5. The SHAP value for a single data sample. BMI: body mass index, CHF: congestive heart failure, CRP: C-reactive protein, Cr: creatinine, NEU: neutrophil, SBP: systolic blood pressure.



Discussion

Principal Findings

In this study, by implementing advanced machine learning techniques, we developed and validated 2 clinical risk prediction models for first-year mortality in incident hemodialysis patients. The 2 extreme gradient boosting models were established based on the data available at dialysis initiation and data from 0-3 months after dialysis initiation. The performance of model 1 (AUC 0.83) was similar to that of model 2 (AUC 0.85), suggesting that we can predict first-year mortality in patients undergoing hemodialysis at dialysis initiation.

Mortality for patients undergoing hemodialysis during the first year of dialysis initiation is high [4]. Therefore, early and precise individualized risk estimates are required for clinical decision making. Traditional strategies for building prediction models have contributed to quality improvement and decision support. Nevertheless, these models have some limitations that may lead to missing important predictors and relationships. Our prediction models (model 1: AUC 0.83, model 2: AUC 0.85), compared with previous models (AUC 0.710-0.752) [12-17], were more accurate in stratifying the risk of first-year mortality for patients undergoing hemodialysis. Our prediction models had several unique and important characteristics. First, many clinical features have been reported for the prediction of first-year mortality in incident hemodialysis patients; some of these features are interact with each other. Traditional prediction models do not account for interactions between input features. By using extreme gradient boosting, we selected the 15 most important features from 42 candidate features, and then combined them nonlinearly. Second, missing data and data noise are inevitable in clinical data collected from the real world, which is a complex problem for traditional strategies. Machine learning techniques can deal with missing data and data noise automatically to improve model performance. Third, relationships between data may change over time because of improvements in treatment and changing populations. For

example, the rates of diabetic nephropathy and cardiovascular disease have been increasing yearly [1,2]. Traditional prediction models are always nonrenewable. Machine learning allows for continual updating of the model to incorporate new data and capture changes in the relationships between features. Finally, compared with traditional predictive models, machine learning models are more complex and harder to interpret; it is not easy to determine how these models make decisions. Therefore, we used SHAP values to interpret the models in this study. SHAP values for a single patient can help physicians evaluate prognosis and make individualized treatment regimens.

Previous studies [8,15,29] have used data from distinct time periods. Floege et al [15], by using 90- to 180-day baseline and 0- to 90-day baseline data for the prediction of first-year mortality, revealed that 2 Cox regression models had similar performances. Some studies [8,29] used data obtained at dialysis initiation to predict the 3- to 6- month mortality of patients undergoing hemodialysis. Akbilgic et al [17] developed a random forest model based on 49 predialysis patient features (AUC 0.75, 95% CI 0.74-0.76); however, it may be not feasible for all users because too many features are needed. Our models were based on 15 features that are easily available for clinicians. The performance of model 1 was satisfactory, suggesting that model 1 can be used to classify high-risk patients at the early stage of dialysis. The first-year mortality risk of dialysis patients may be reduced by personalized and targeted preventive therapies.

Limitations and Future Work

Despite the promising prospects demonstrated by our study, it had some limitations. First, our training data were based on retrospective data generated from a single center. Therefore, a possible center effect cannot be excluded. Second, although no restriction was placed on ethnicity, all patients included were Chinese. The primary disease of end-stage renal disease and cardiovascular conditions of patients undergoing hemodialysis in China differ from those of patients undergoing hemodialysis in other regions [2,30]. Thus, the applicability of our models to other ethnic groups and regions needs to be confirmed. Third, we only assessed 1-year mortality, whereas long-term mortality is also important [31]. Therefore, we plan to establish a model to predict 2-year and 5-year mortality in future studies. Finally, therapeutic intervention data, such as dialysis dose and frequency, were not used in this study because therapeutic interventions were not always fixed until 1-2 months after dialysis initiation, and therapeutic interventions in patients varied. We also plan to display the prediction models on the website of the Zhejiang Dialysis Quality Control Center and as a mobile app for better application.

Conclusions

To accurately predict first-year mortality in incident hemodialysis patients, we developed and validated 2 machine learning models based on data available at dialysis initiation and data 0-3 months after dialysis initiation. The overall diagnostic performances of the 2 models were similar. We hope our models may assist clinicians in stratifying the risk of mortality at the early stages of dialysis. Our models need to be evaluated in data sets of patients undergoing hemodialysis from other ethnic groups and regions before implementation in clinical practice. For future research, long-term mortality predictions for patients undergoing incident dialysis will be addressed.

Acknowledgments

This work was supported by National Key Research and Development Projects of China (2018YFC1314003). Study sponsors had no role in study design; collection, analysis, and interpretation of data; writing the report; and the decision to submit the report for publication.

Authors' Contributions

KS and JC conceptualized the study; JC acquired funding; KS, XY, JL, and YH collected data; KS developed methodology, analyzed the data, and wrote the first draft; and PZ reviewed and edited.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Importance rankings of 42 features based on data at dialysis initiation.
[DOC File , 224 KB - medinform_v8i10e20578_app1.doc]

Multimedia Appendix 2
Importance ranking of 42 features based on data 0-3 months after dialysis initiation.
[DOC File , 224 KB - medinform_v8i10e20578_app2.doc]

References

1. Zhang L, Wang F, Wang L, Wang W, Liu B, Liu J, et al. Prevalence of chronic kidney disease in China: a cross-sectional survey. The Lancet 2012 Mar;379(9818):815-822. [doi: 10.1016/s0140-6736(12)60033-6]

2. Saran RR, Abbott KC. US Renal Data System 2018 Annual Data Report: Epidemiology of Kidney Disease in the United States American journal of kidney diseases : the official journal of the National Kidney Foundation 2019, 73(3S1). 2019. URL: https://www.usrds.org/annual-data-report/previous-adrs/ [accessed 2020-09-01]
3. Collaboration GBDCKD. Global, regional, and national burden of chronic kidney disease, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2020 Feb 29;395(10225):709-733 [FREE Full text] [doi: 10.1016/S0140-6736(20)30045-3] [Medline: 32061315]
4. Robinson BM, Zhang J, Morgenstern H, Bradbury BD, Ng LJ, McCullough KP, et al. Worldwide, mortality risk is high soon after initiation of hemodialysis. Kidney Int 2014 Jan;85(1):158-165 [FREE Full text] [doi: 10.1038/ki.2013.252] [Medline: 23802192]
5. Foley RN, Chen S, Solid CA, Gilbertson DT, Collins AJ. Early mortality in patients starting dialysis appears to go unregistered. Kidney Int 2014 Aug;86(2):392-398 [FREE Full text] [doi: 10.1038/ki.2014.15] [Medline: 24522495]
6. Kovesdy CP, Naseer A, Sumida K, Molnar MZ, Potukuchi PK, Thomas F, et al. Abrupt Decline in Kidney Function Precipitating Initiation of Chronic Renal Replacement Therapy. Kidney Int Rep 2018 May;3(3):602-609 [FREE Full text] [doi: 10.1016/j.ekir.2017.12.007] [Medline: 29854967]
7. Jassal SV, Karaboyas A, Comment LA, Bieber BA, Morgenstern H, Sen A, et al. Functional Dependence and Mortality in the International Dialysis Outcomes and Practice Patterns Study (DOPPS). Am J Kidney Dis 2016 Feb;67(2):283-292 [FREE Full text] [doi: 10.1053/j.ajkd.2015.09.024] [Medline: 26612280]
8. Wick JP, Turin TC, Faris PD, MacRae JM, Weaver RG, Tonelli M, et al. A Clinical Risk Prediction Tool for 6-Month Mortality After Dialysis Initiation Among Older Adults. Am J Kidney Dis 2017 May;69(5):568-575. [doi: 10.1053/j.ajkd.2016.08.035] [Medline: 27856091]
9. Saleh T, Sumida K, Molnar MZ, Potukuchi PK, Thomas F, Lu JL, et al. Effect of Age on the Association of Vascular Access Type with Mortality in a Cohort of Incident End-Stage Renal Disease Patients. Nephron 2017 May 18;137(1):57-63 [FREE Full text] [doi: 10.1159/000477271] [Medline: 28514785]
10. Karaboyas A, Morgenstern H, Li Y, Bieber BA, Hakim R, Hasegawa T, et al. Estimating the Fraction of First-Year Hemodialysis Deaths Attributable to Potentially Modifiable Risk Factors: Results from the DOPPS. CLEP 2020 Jan;12:51-60. [doi: 10.2147/clep.s233197]
11. Karaboyas A, Morgenstern H, Waechter S. Low hemoglobin at hemodialysis initiation: an international study of anemia management and mortality in the early dialysis period. Clin Kidney J 2020;13(3):425-433. [doi: 10.1093/ckj/sfz065]
12. Mauri JM, Clèries M, Vela E, Catalan Renal Registry. Design and validation of a model to predict early mortality in haemodialysis patients. Nephrol Dial Transplant 2008 May 26;23(5):1690-1696. [doi: 10.1093/ndt/gfm728] [Medline: 18272779]
13. Chua H, Lau T, Luo N, Ma V, Teo B, Haroon S, et al. Predicting first-year mortality in incident dialysis patients with end-stage renal disease - the UREA5 study. Blood Purif 2014 Feb 26;37(2):85-92 [FREE Full text] [doi: 10.1159/000357640] [Medline: 24589505]
14. Doi T, Yamamoto S, Morinaga T, Sada KE, Kurita N, Onishi Y. Risk Score to Predict 1-Year Mortality after Haemodialysis Initiation in Patients with Stage 5 Chronic Kidney Disease under Predialysis Nephrology Care. PloS one 2015;10(6):e0129180. [doi: 10.3390/ijerph120303002] [Medline: 25768239]
15. Floege J, Gillespie IA, Kronenberg F, Anker SD, Gioni I, Richards S, et al. Development and validation of a predictive mortality risk score from a European hemodialysis cohort. Kidney Int 2015 May;87(5):996-1008 [FREE Full text] [doi: 10.1038/ki.2014.419] [Medline: 25651366]
16. Quinn RR, Laupacis A, Hux JE, Oliver MJ, Austin PC. Predicting the Risk of 1-Year Mortality in Incident Dialysis Patients. Medical Care 2011;49(3):257-266. [doi: 10.1097/mlr.0b013e318202aa0b]
17. Ramspek C, Voskamp P, van Ittersum F, Krediet R, Dekker F, van Diepen M. Prediction models for the mortality risk in chronic dialysis patients: a systematic review and independent external validation study. CLEP 2017 Sep;9:451-464. [doi: 10.2147/clep.s139748]
18. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nat Med 2019 Jan 7;25(1):70-74. [doi: 10.1038/s41591-018-0240-2]
19. Chen T, Li X, Li Y, Xia E, Qin Y, Liang S, et al. Prediction and Risk Stratification of Kidney Outcomes in IgA Nephropathy. Am J Kidney Dis 2019 Sep;74(3):300-309. [doi: 10.1053/j.ajkd.2019.02.016] [Medline: 31031086]
20. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nature medicine 2019 Jan;25(1):30-36. [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
21. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan 7;25(1):44-56. [doi: 10.1038/s41591-018-0300-7]
22. Akbilgic O, Obi Y, Potukuchi PK, Karabayir I, Nguyen DV, Soohoo M, et al. Machine Learning to Identify Dialysis Patients at High Death Risk. Kidney Int Rep 2019 Sep;4(9):1219-1229 [FREE Full text] [doi: 10.1016/j.ekir.2019.06.009] [Medline: 31517141]

23. Levey AS, Stevens LA, Schmid CH, Zhang Y, Castro AF, Feldman HI, CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration). A new equation to estimate glomerular filtration rate. Ann Intern Med 2009 May 05;150(9):604-612 [FREE Full text] [doi: 10.7326/0003-4819-150-9-200905050-00006] [Medline: 19414839]
24. Asar Ö, Ilk O, Dag O. Estimating Box-Cox power transformation parameter via goodness-of-fit tests. Communications in Statistics - Simulation and Computation 2014 Dec 12;46(1):91-105. [doi: 10.1080/03610918.2014.957839]
25. Chen TG. Xgboost: A scalable tree boosting system. 2016 Presented at: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016; San Francisco URL: https://dl.acm.org/doi/pdf/10.1145/2939672.2939785 [doi: 10.1145/2939672.2939785]
26. Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees. Advances in Neural Information Processing Systems. 2013. URL: http://papers.nips.cc/paper/4928-understanding-variable-importances-in-forests-of-randomized [accessed 2020-09-01]
27. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics 2005 Aug 01;21(15):3301-3307. [doi: 10.1093/bioinformatics/bti499] [Medline: 15905277]
28. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017. URL: https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf [accessed 2020-09-01]
29. Couchoud CG, Beuscart JR, Aldigier J, Brunet PJ, Moranne OP, REIN registry. Development of a risk stratification algorithm to improve patient-centered care and decision making for incident elderly patients with end-stage renal disease. Kidney Int 2015 Nov;88(5):1178-1186 [FREE Full text] [doi: 10.1038/ki.2015.245] [Medline: 26331408]
30. Wang F, Yang C, Long J. Executive summary for the 2015 Annual Data Report of the China Kidney Disease Network. Kidney Int 2019 Aug;96(2):501-505. [doi: 10.1016/j.kint.2019.05.004] [Medline: 31331481]
31. Arase H, Yamada S, Hiyamuta H, Taniguchi M, Tokumoto M, Tsuruya K, et al. Modified creatinine index and risk for long-term infection-related mortality in hemodialysis patients: ten-year outcomes of the Q-Cohort Study. Sci Rep 2020 Jan 27;10(1):1241 [FREE Full text] [doi: 10.1038/s41598-020-58181-6] [Medline: 31988325]

Abbreviations

AUC: area under the curve
BMI: body mass index
SHAP: Shapley additive explanation

Edited by G Eysenbach; submitted 23.05.20; peer-reviewed by L Cilar, M Sokolova; comments to author 01.07.20; revised version received 15.08.20; accepted 16.08.20; published 29.10.20.

Please cite as:
Sheng K, Zhang P, Yao X, Li J, He Y, Chen J
Prognostic Machine Learning Models for First-Year Mortality in Incident Hemodialysis Patients: Development and Validation Study
JMIR Med Inform 2020;8(10):e20578
URL: http://medinform.jmir.org/2020/10/e20578/
doi:10.2196/20578
PMID:33118948

Original Paper

Predictive Models for Neonatal Follow-Up Serum Bilirubin: Model Development and Validation

Joseph H Chou[1], MD, PhD

Massachusetts General Hospital, Boston, MA, United States

Corresponding Author:
Joseph H Chou, MD, PhD
Massachusetts General Hospital
55 Fruit Street, Founders 526E
Boston, MA, 02114-2696
United States
Phone: 1 617 724 9040
Email: jchou2@mgh.harvard.edu

Abstract

Background: Hyperbilirubinemia affects many newborn infants and, if not treated appropriately, can lead to irreversible brain injury.

Objective: This study aims to develop predictive models of follow-up total serum bilirubin measurement and to compare their accuracy with that of clinician predictions.

Methods: Subjects were patients born between June 2015 and June 2019 at 4 hospitals in Massachusetts. The prediction target was a follow-up total serum bilirubin measurement obtained <72 hours after a previous measurement. Birth before versus after February 2019 was used to generate a training set (27,428 target measurements) and a held-out test set (3320 measurements), respectively. Multiple supervised learning models were trained. To further assess model performance, predictions on the held-out test set were also compared with corresponding predictions from clinicians.

Results: The best predictive accuracy on the held-out test set was obtained with the multilayer perceptron (ie, neural network, mean absolute error [MAE] 1.05 mg/dL) and Xgboost (MAE 1.04 mg/dL) models. A limited number of predictors were sufficient for constructing models with the best performance and avoiding overfitting: current bilirubin measurement, last rate of rise, proportion of time under phototherapy, time to next measurement, gestational age at birth, current age, and fractional weight change from birth. Clinicians made a total of 210 prospective predictions. The neural network model accuracy on this subset of predictions had an MAE of 1.06 mg/dL compared with clinician predictions with an MAE of 1.38 mg/dL (P<.0001). In babies born at 35 weeks of gestation or later, this approach was also applied to predict the binary outcome of subsequently exceeding consensus guidelines for phototherapy initiation and achieved an area under the receiver operator characteristic curve of 0.94 (95% CI 0.91 to 0.97).

Conclusions: This study developed predictive models for neonatal follow-up total serum bilirubin measurements that outperform clinicians. This may be the first report of models that predict specific bilirubin values, are not limited to near-term patients without risk factors, and take into account the effect of phototherapy.

(JMIR Med Inform 2020;8(10):e21222) doi:10.2196/21222

KEYWORDS

infant, newborn; neonatology; jaundice, neonatal; hyperbilirubinemia, neonatal; machine learning; supervised machine learning; data science; medical informatics; decision support techniques; models, statistical; predictive models

Introduction

Neonatal Jaundice: Bilirubin Production and Clearance

Management of jaundice is one of the most common, yet vexing, problems in newborn medicine and requires consideration of the myriad contributors to the production and clearance of bilirubin [1]. If not recognized and managed appropriately, hyperbilirubinemia can result in permanent harm. A large proportion of neonatal readmissions is related to jaundice [2]. Bilirubin arises from the catabolism of iron protoporphyrin (heme) from hemoglobin in red blood cells. Unconjugated bilirubin is poorly water soluble and largely bound to albumin

but is conjugated in the liver into a more water-soluble form more readily excreted in the bile and urine.

A number of physiological mechanisms put newborn infants at particular risk of developing jaundice in the first few days after birth, including increased red blood cell volume, higher red blood cell turnover, decreased hepatic uptake and conjugation of bilirubin, and increased enterohepatic circulation (intestinal hydrolysis of conjugated bilirubin resulting in reabsorption of unconjugated bilirubin). This initial imbalance of increased bilirubin production and decreased conjugation and clearance results in >80% of newborn infants born near or at term developing visible jaundice in the first week after birth. Preterm neonates may have further decreased ability to conjugate and clear bilirubin [3]. The imbalance between production and clearance typically stabilizes by around 4 days after birth [4]. However, other factors manifesting in the newborn period can further affect bilirubin production and clearance, for example, isoimmune hemolytic jaundice from maternal blood type mismatch and transplacental transmission of maternal immunoglobulins or inadequate enteral intake resulting in dehydration, decreased bile clearance, and increased enterohepatic circulation.

Bilirubin-Induced Morbidity

Although lower levels of hyperbilirubinemia are generally well tolerated by newborn infants, at sufficiently high concentrations, unconjugated bilirubin, presumably unbound to albumin, can cross the blood-brain barrier with potentially devastating consequences [5]. The manifestations of bilirubin-induced neurological dysfunction range from sleepiness, lethargy, discoordinated suck reflex, and high-pitched cry to abnormal muscle tone, athetosis, oculomotor paralysis, and opisthotonos, with associated sensorineural hearing loss and intellectual deficits. Extremely severe cases may result in seizures, coma, and death. Kernicterus originally referred to the pathologic finding of yellow bilirubin staining of the deep nuclei of the brain but is now also used to describe the syndrome of severe bilirubin encephalopathy.

Phototherapy

Phototherapy is an effective treatment to prevent bilirubin-associated morbidity [6]. Absorption of light through the dermis and subcutaneous tissue induces photochemical changes in bilirubin to produce more hydrophilic isomers and derivatives that can be excreted in bile and urine without the need for conjugation. A visible spectrum of blue light from 460 nm to 490 nm in wavelength appears to have maximal efficacy in both penetrating tissue and formation of bilirubin photoproducts. Although phototherapy is not known to affect the rates of bilirubin production, effective administration is often able to increase bilirubin clearance to a rate greater than the rate of ongoing production, thereby lowering the total serum bilirubin concentration.

Consensus Clinical Guidelines

Clinical guidelines have been developed to assist in the management of neonatal hyperbilirubinemia, including specifying thresholds at which phototherapy or other therapies should be provided [1]. The availability of effective treatments and the potentially devastating consequences of not initiating therapy have made it difficult to develop evidence-based guidelines, for example, via randomized controlled clinical trials or systematic observational studies. Therefore, currently available guidelines are largely consensus based.

There is not a single universally accepted guideline. An informal international survey conducted during the development of the Norwegian guidelines [7] for the treatment of neonatal jaundice reported that 18 of the 28 countries surveyed had national consensus treatment guidelines, including the United States [8], South Africa [9], Canada [10], Israel [11], the United Kingdom [12,13], and Norway [7]. They found that these guidelines differed considerably in the recommended total serum bilirubin level at which phototherapy should be initiated, indications for exchange transfusion, addressing the preterm population, use of transcutaneous bilirubinometry, when phototherapy should be discontinued, and recommended follow-up at or after discharge.

Of the identified national guidelines for the management of neonatal hyperbilirubinemia, 14 of 16 included recommendations for late preterm infants (typically born at 35 weeks of gestation or later) and 10 of 16 for early preterm infants. Although less in number, guidelines have also been developed specifically for the preterm population, again consensus based [14-16]. There are also unpublished locally developed treatment practices for preterm infants [17]. For example, in several of the Boston area teaching hospitals, an informal and unpublished rule of thumb for preterm infants is to divide the birth weight in grams by 200 as the phototherapy threshold in mg/dL (eg, 1500 g birth weight yielding a phototherapy threshold of 7.5 mg/dL) and twice that value as an exchange transfusion threshold.

Rebound Hyperbilirubinemia

With the implementation of universal bilirubin screening of newborn infants during birth hospitalization, clinical practice guidelines advise whether to initiate phototherapy (although strict adherence to guidelines varies [18]), but less often provide direction on when to discontinue phototherapy and whether reinitiation of treatment may be required because of rebound hyperbilirubinemia.

Rebound bilirubin, in general, refers to an increase in the bilirubin level after discontinuation of phototherapy, likely related to the removal of the additional bilirubin clearance provided by phototherapy and the resultant return to net balance of greater bilirubin production than clearance. However, the specific definitions of rebound bilirubin vary considerably. Some definitions include the change in bilirubin level on the first follow-up serum bilirubin at any time up to 30 hours after discontinuation of phototherapy [19], between 4 hours and 48 hours after discontinuation [20], within 12 hours [21], or after approximately 6 hours [22]. Over time, the definition began to incorporate the concept of rebound to significant hyperbilirubinemia. The choice of significance could be an arbitrarily chosen constant threshold [23], a measurement between 18 hours and 30 hours after discontinuation that prompted reinstitution of phototherapy [24], or an increase at

any time that resulted in exceeding the age-specific threshold of a specified clinical guideline to initiate phototherapy [25].

Predictive Models

Defining rebound hyperbilirubinemia as exceeding the phototherapy initiation threshold of a practice guideline raises the possibility of developing predictive models to provide clinical decision support.

Predictive models can be generated by a class of statistical approaches referred to as supervised machine learning [26]. With supervised learning, a model is trained using a data set containing predictive features and their known target outcomes, with the aim that the trained model can later be used on a new set of the same predictive features to predict unknown outcomes. The goal might be a classification task—for example, predicting the likelihood of survival, readmission, or need to initiate phototherapy—or a regression task to calculate a continuous numeric outcome, such as a laboratory value. Some examples of machine learning models are as familiar as linear regression (ordinary least squares), which performs a regression prediction, and logistic regression, which performs a classification task despite its historic name. Different machine learning models differ in their approach and the flexibility with which they can predict outcomes. For example, both linear and logistic regression are in the family of generalized linear models and are relatively inflexible as a unit change in the value of each predictor produces a constant linear change in the output. More flexible models may be able to better fit the training data and perform better with new predictions but risk overfitting the training data, resulting in poorer performance on new, previously unseen data, that is, poor model generalization. Examples of strategies to limit overfitting include choosing less-flexible models or applying an approach called regularization that applies a penalty for larger model coefficients. Inappropriate use of too many predictors can also contribute to overfitting as high model flexibility can allow learning what is effectively noise and not signal in the predictive features of the training set. Owing to these risks, in general, it is best to evaluate a predictive model's performance on data that was not previously used for training. Approaches to achieve this include using a completely separate training set and held-out test set or using K-fold cross-validation to partition the data and then training and evaluating models on each partition.

Chang et al [27,28] developed and subsequently simplified a logistic regression model to predict the need to resume phototherapy after an initial treatment episode with decision thresholds defined by the American Academy of Pediatrics (AAP) consensus treatment guidelines [8]. The choice of this specific guideline restricts its applicability to newborn infants born at ≥35 weeks of gestation. As clinical guidelines can vary significantly, the ability to generalize the published model to different guidelines may be limited. Another potential issue is assuming the validity of applying an age-specific treatment guideline, which was developed from a nomogram derived from a cohort of normal newborn infants without any previous phototherapy treatment, on infants who may have received varying duration and intensity of phototherapy. In these published models, there is no prediction distinction between a newborn infant who had phototherapy initiated very early because of the rapid development of jaundice (perhaps related to hemolysis) and another infant who had phototherapy initiated several days after birth as long as their bilirubin levels for a given age were subsequently the same after phototherapy. Moreover, the model can only be applied after an initial episode of phototherapy; it cannot be used to predict the need to initiate a first episode of phototherapy or account for multiple previous episodes of phototherapy.

Aims of This Study

A more general approach that predicts actual bilirubin values, rather than exceeding thresholds defined within a particular treatment guideline, and not limited by gestational age or by restrictions on phototherapy utilization, might be helpful. By predicting actual bilirubin values, the approach could provide clinical decision support related to any given clinical guideline, including those developed for preterm infants. Training more flexible models than generalized linear models might improve prediction performance. This study aims to (1) develop and compare multiple predictive models of follow-up total serum bilirubin measurements that could be utilized regardless of gestational age or previous treatment with phototherapy; (2) to compare accuracy with clinician predictions; and (3) to demonstrate an example application to one specific clinical guideline.

Methods

Patient Cohort

The subjects of this retrospective study were newborn infants born at any gestation between June 2015 and June 2019 at 4 birthing hospitals in Massachusetts within the Partners HealthCare system. The hospitals provided a range of levels of neonatal care [29], with 2 hospitals providing up to level 2 care, 1 hospital providing up to level 3, and 1 hospital providing up to level 4. As the prediction target was a follow-up total serum bilirubin measurement obtained <72 hours after a previous measurement, the inclusion criteria were 2 bilirubin measurements <72 hours apart within the first 10 days after birth. There were no other exclusion criteria.

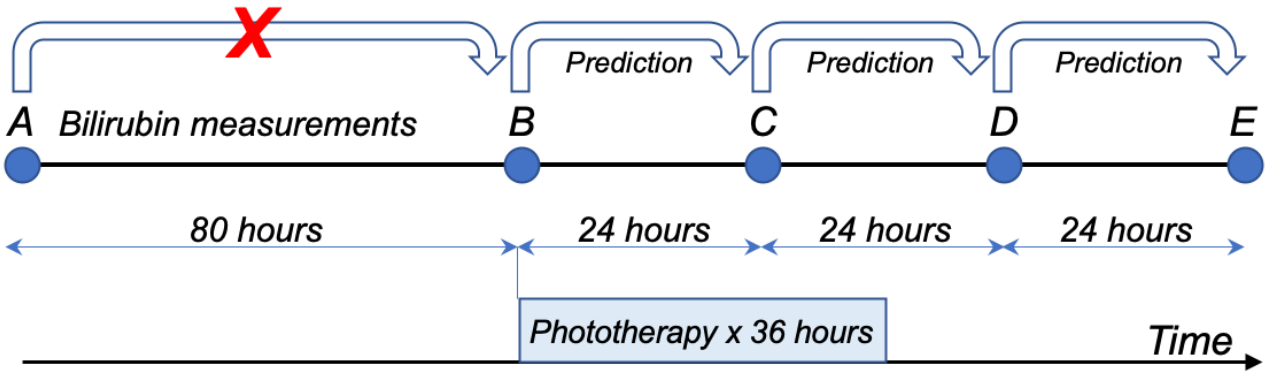
Features of the Predictive Model

Data from inpatient encounters were abstracted from the electronic health record (EHR) by database query and included gestational age at birth, birth weight, gender, maternal age, gravida, para, race and ethnicity, route of delivery and whether the delivery was vacuum assisted or forceps assisted, 1-min and 5-min Apgar scores, maternal and baby blood type and Rh, baby direct Coombs, and initial baby hematocrit. Data from the first 10 days after birth included total serum bilirubin measurements, inpatient phototherapy start and stop times based on physician orders, weights, enteral feeds, urine output, and stools. Feature engineering included encoding nonnumeric (categorical) predictors to binary features of whether known to be present and included maternal race (White, Black, Hispanic, or Asian); ABO incompatibility as maternal blood type O and baby blood type A, B, or AB; Rh mismatch as maternal Rh negative and baby Rh positive; baby direct Coombs positive; cesarean

delivery; forceps-assisted delivery; and vacuum-assisted delivery. If categorical data were unavailable, the feature was set to not known to be present. Median imputation was used for numeric features with missing data. Birth weight Z-score was calculated as described previously [30,31].

Individuals frequently had >2 bilirubin measurements, permitting multiple prediction targets. The goal of prediction was to only use information available at the time of a given bilirubin measurement (the *current* measurement) to predict the subsequent measurement. Available information included age in hours, current measurement, previous bilirubin rate of rise, and proportion of time under phototherapy between the previous and the current measurement. For the first bilirubin measurement, there would be no previous measurement; in this case, the time zero measurement was imputed as 2.0 mg/dL based on previous reports of umbilical cord bilirubin level and extrapolation from postnatal nomograms [4,32-35]. If 2 serum bilirubin measurements were recorded <2 hours apart, the earlier measurement was discarded as this generally reflected an erroneous first measurement. Additional features generated from the available data included fraction weight change from birth and counts of stools, urine output, and feeds on the previous calendar day. To make predictions, the only data permitted from after the current measurement were factors under clinician control, that is, number of hours until the target measurement and the fraction of that time that would be under phototherapy (between 0 for no phototherapy before the next measurement and 1 for continuous phototherapy until the next measurement). Figure 1 shows a schematic of the data inclusion mechanism and illustrates how the predictive feature of the fraction of time under phototherapy was calculated.

Figure 1. Schematic of a data inclusion mechanism for a hypothetical individual with 5 bilirubin measurements, A through E. The blue box represents the time period under phototherapy. Model training used features at the time of a bilirubin measurement to predict the value of a subsequent measurement ≤72 hours later. The predictive feature of fraction of time under phototherapy was 100% between B and C, 50% between C and D, and 0% between D and E. Data from bilirubin measurement A to predict B were not included for model training because the subsequent measurement was ≥72 hours later.



Predictive Model Training

All data for patients born on or after February 1, 2019, were set aside as a held-out test set and not accessed before predictive model testing. The remaining data were used for model training.

Multiple supervised learning models were trained including linear models, linear models with interaction terms regularized via ridge regression or least absolute shrinkage and selection operator (LASSO), random forest, multilayer perceptron (a simple neural network with 2 densely connected hidden layers using the rectified linear unit nonlinear activation function), long short-term memory (LSTM) neural network, and Xgboost. Feature selection was explored by best subset selection for the linear model without interaction terms and variable importance for random forest and Xgboost. To improve neural network convergence, numeric predictors were centered and scaled by subtracting the mean and dividing by the SD of the predictors in the training set; both the validation and test sets were centered and scaled using the training set. The training set for the LSTM neural network was generated by creating a moving window of up to 4 time steps (zero-padded for the first 3 time points), allowing a memory of previous predictors. Analysis was performed using the R statistical programming language (R Core Team, 2018) [36]. Multimedia Appendix 1 includes the R code used for model training and references to the packages used. For data visualizations, smoothed conditional mean curves with 95% CIs were generated using the ggplot2 package [37].

Comparison With Clinician Accuracy

From February 2019, a convenience sample of clinician predictions of follow-up bilirubin measurements was obtained by identifying currently admitted newborn infants at 1 hospital who had a recent bilirubin measurement and a provider clinical order for a follow-up bilirubin level to be obtained within the next 72 hours. Clinicians actively providing care for that neonate were approached and asked to provide predictions. Participation was voluntary and no information identifying the clinician was recorded other than the role group. Role groups included attending board-certified neonatologists, advanced practitioners (including neonatal nurse practitioners, neonatal-perinatal medicine fellows, and pediatric hospitalists with primary roles in the newborn intensive care unit [NICU]), pediatric residents (either interns or seniors during their NICU rotation), and bedside nurses (neonatal nurses, all in the level 2 and level 3 nurseries). Clinicians were asked to use all available information, including data not documented in the EHR, for example, team discussions during bedside rounds, conversations with parents and lactation consultants, etc.

Statistical Analysis

Comparisons were performed using the *t* test (either paired or unpaired), analysis of variance, Wilcoxon rank-sum, Kruskal-Wallis, or Pearson chi-square test, as appropriate. When multiple pairwise comparisons of paired *t* tests were performed, multiple testing adjustment was performed using the Holm method. The absolute value of prediction errors is nonnegative, which results in a right-skewed distribution; therefore, in general, medians and IQRs are reported below. However, in pairwise comparisons of models, the differences in absolute errors were distributed more normally (data not shown). The confidence interval for the area under the receiver operator characteristic (AUROC) curve was obtained using the method of Hanley [38].

Human Subjects' Research

This study was approved by the Partners Human Research Committee institutional review board.

Results

Patient Cohort Characteristics

A total of 52,149 babies born between June 2015 and June 2019 were identified, of whom 46,361 were born before February 1, 2019. The 5788 babies born after February 2019 were set aside as a held-out test set and not accessed until predictive model evaluation.

Of the patients born before February 2019, 9723 babies had at least 2 total serum bilirubin measurements <72 hours apart within the first 10 days after birth and were included in the training set, whereas the remaining 36,638 babies were excluded, as detailed in Multimedia Appendix 2. The patients included in the training set tended to be of lower gestational age, lower birth weight, lower birth weight Z-score, and lower 1-min and 5-min Apgar scores; male; C-sectioned; forceps assisted; vacuum assisted; ABO-mismatched; absence of Rh mismatch; baby direct Coombs positive; and of maternal race Asian, Black, or not White. Of the patients in the training set, the median number of serum bilirubin measurements was 3 (IQR 2-5) and 34.34% (3339/9723) received phototherapy. There were significant missing data (>10%) in both the included and excluded patients for maternal and baby blood type and Rh, baby direct Coombs, and baby hematocrit. If the maternal blood type was O, the baby's blood type was less likely to be missing (1017/18,930, 5.37%). Similarly, if the mother was Rh negative, the baby's Rh status was unlikely to be missing (57/4919, 1.16%).

Predictive Model Training

Of the 9723 babies in the training set, there were a total of 37,151 total serum bilirubin measurements resulting in 27,428 training examples. After feature engineering, 34 candidate predictors were available for model training, including 22 that did not vary with time (gestational age at birth; birth weight; birth weight Z-score; gender; 1-min and 5-min Apgar scores; cesarean versus vaginal delivery; forceps assistance; vacuum assistance; maternal age; gravida; para; maternal race Asian, Black, Hispanic, or White; ABO blood type mismatch; Rh mismatch; baby direct Coombs status; baby initial hematocrit; age; and value of first total serum bilirubin measurement) and 12 predictors that varied with time (current age; current bilirubin level; fractional weight change; count of breast milk, formula and donor human milk feeds, urine output and stools; last rate of rise; last proportion of time under phototherapy; and time to next measurement and fraction of that time under phototherapy). For the linear models, the quadratic age-squared term was added to account for the nonlinearity of bilirubin trajectories with age [4].

During the initial model exploration, it quickly became apparent that a limited number of predictive features would be sufficient for near-optimal model performance. For the simple linear model, the best subset and stepwise forward feature selection chose the same features until the 13th predictive feature was added, but showed limited improvement after the seventh feature (minimal R^2 statistic improvement from 0.783 to 0.785). The features selected, in order of importance, included current result, proportion phototherapy before target measurement, current age, previous proportion of phototherapy, current age squared, time to target measurement, count of breast milk feeds, and first bilirubin measurement.

Random forest and Xgboost models are able to report predictive feature importance contributing to model accuracy. Providing all 34 predictive features to the random forest and Xgboost models and inspection of the variable importance plots also suggested that a limited number of features would provide near-maximal predictive accuracy. For the Xgboost model, the top 8 features included current result, last rate of rise, proportion phototherapy, time to target measurement, birth weight, gestational age, first bilirubin measurement, and current age; each of the remaining 26 features contributed <1% to Xgboost variable importance (data not shown). For the random forest model, the top 8 features included time to target measurement, proportion phototherapy, previous rate of rise, current result, current age, birth weight Z-score, previous proportion of phototherapy, and count of formula feeds.

The 8 features selected for final predictive model training were current result, last rate of rise, proportion of time under phototherapy between the current and the future target measurement, time to target measurement, gestational age, current age, previous proportion of time under phototherapy, and fractional weight change from birth. All models used these features except the age-squared term that was included for the linear models (to allow for nonlinear response with age). The last rate of rise and previous proportion of time under phototherapy were excluded from the LSTM model as those features were available via the preceding time step. Seven predictive models and 1 negative control were generated with the training set (Textbox 1; further detailed in Multimedia Appendix 1).

Textbox 1. Predictive models and descriptions.

- *current*: Negative control, predicting the current bilirubin level as the subsequent level
- *lm*: Linear model with no interaction terms; includes quadratic age-squared term
- *ridge*: Linear model with all combinations of predictors as interaction terms and ridge regression regularization (L2 norm) selected by 10-fold cross-validation for coefficient shrinkage
- *lasso*: Similar to *ridge* but using least absolute shrinkage and selection operator (LASSO) regularization (L1 norm) for coefficient shrinkage and implicit feature selection
- *nn*: Multilayer perceptron (a simple neural network) with 2 fully connected hidden layers
- *lstm*: Long short-term memory recurrent neural network with 4 time steps feeding into a single hidden layer
- *rf*: Decision tree-based random forest ensemble with 500 trees
- *xgboost*: Decision tree-based XGBoost ensemble model with 500 boosting iterations

Predictive Model Assessment

Predictive model performance was assessed using the held-out test set. Of the 5788 babies born after February 2019, 1224 had at least 2 total serum bilirubin measurements <72 hours apart within the first 10 days after birth, with a total of 4544 total serum bilirubin measurements resulting in a test set of 3320 examples.

For each prediction, the error is defined by the predicted value minus the actual value, with positive and negative values reflecting predictions that are too high or too low, respectively. Prediction models often have an overall mean prediction error of 0; simplistically, if the prediction is equally likely to be too high (positive error) or too low (negative error), the mean error may be near 0. Therefore, to assess the model performance, the absolute value of the prediction errors, which can be considered the magnitude of the error, was calculated.

Table 1 summarizes the predictive performance of all 8 models, which included a negative control and 7 models trained by supervised learning. The second and third columns show the mean and median absolute value of prediction errors for the 3320 test set examples for each model. The Xgboost model had the lowest mean (1.04 mg/dL, SD 0.99) and median (0.78 mg/dL) absolute values of prediction error, that is, for the Xgboost model, 50% of the test set predictions were within 0.78 mg/dL of the actual value.

Table 1. Pairwise comparison of the predictive models.

Model[a]	MAE[b] (SD)	Median (IQR)[c]	*P* value[d]						
			current	lm	ridge	lstm	lasso	nn	rf
current	2.105 (1.674)	1.800 (0.900-2.900)	N/A[e]						
lm	1.325 (1.208)	0.997 (0.476-1.808)	<.0001	N/A					
ridge	1.175 (1.095)	0.893 (0.420-1.609)	<.0001	<.0001	N/A				
lstm	1.121 (1.142)	0.809 (0.363-1.493)	<.0001	<.0001	.0056	N/A			
lasso	1.075 (1.036)	0.802 (0.365-1.456)	<.0001	<.0001	<.0001	.0067	N/A		
nn	1.053 (1.007)	0.791 (0.362-1.407)	<.0001	<.0001	<.0001	<.0001	.056	N/A	
rf	1.050 (1.003)	0.782 (0.355-1.438)	<.0001	<.0001	<.0001	<.0001	.090	.74	N/A
xgboost	1.038 (0.989)	0.776 (0.355-1.427)	<.0001	<.0001	<.0001	<.0001	.0045	.29	.29

[a]Models are as described in Textbox 1.

[b]MAE: mean absolute error of bilirubin level predictions with SD (mg/dL) on the held-out test set (n=3320).

[c]Median absolute error of bilirubin level predictions and IQR (mg/dL).

[d]*P* values for pairwise model comparisons by paired *t* test with Holm adjustment for multiple testing.

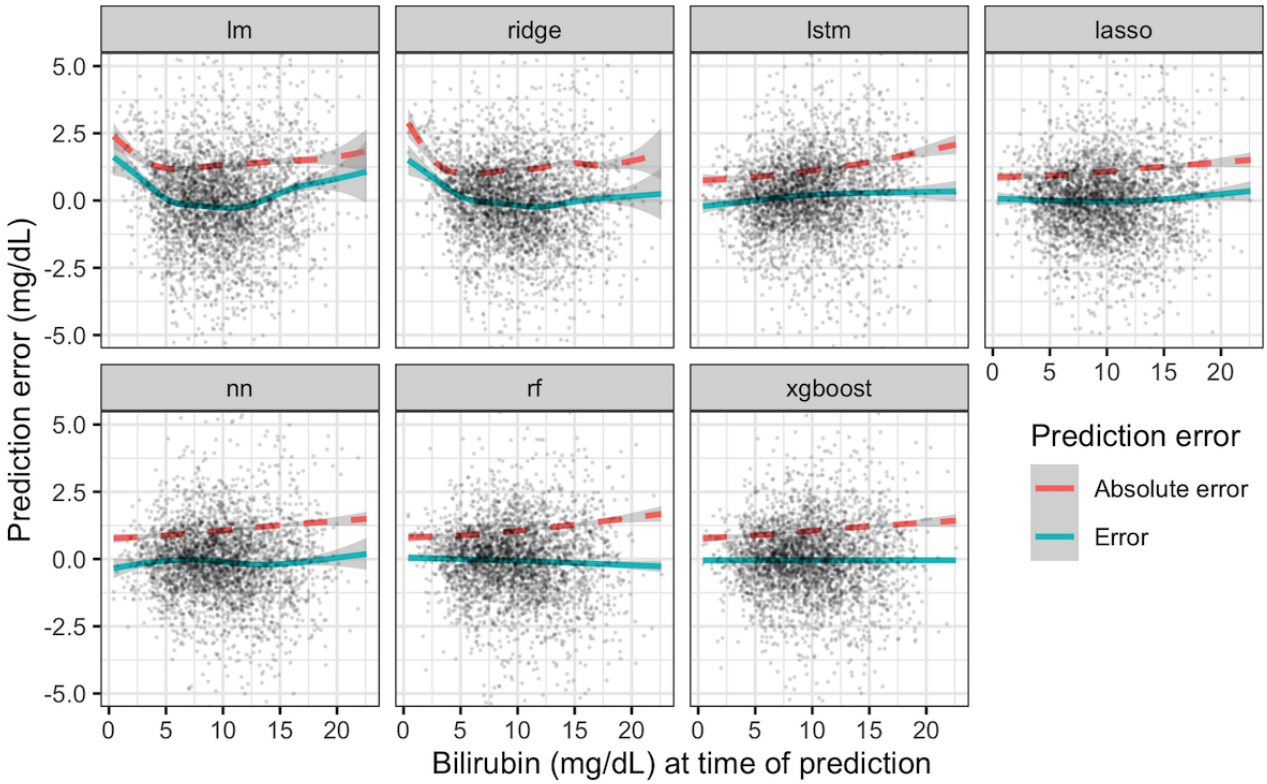
[e]N/A: not applicable.

To assess the performance of each model with respect to each other model, a total of 28 pairwise comparisons of the 8 models' predictions on the same 3320 test set examples, including the negative control, were analyzed by using a paired *t* test with Holm adjustment for multiple testing (Table 1, right-most 7 columns). Xgboost performance (Table 1, last row) was statistically significantly better than the negative control, simple linear model, ridge regression, LSTM neural network, and LASSO models (*P* values from <.0001 to .0045), but was not statistically significantly superior to the simple neural network (*P*=.29) or random forest (*P*=.29) models.

Although Table 1 summarizes model performance across the entire test set, the greatest clinical concern is high bilirubin levels. To visualize whether performance was impacted by bilirubin level, prediction error was visualized with respect to the bilirubin value at the time of prediction for each of the

models (Figure 2). Each point represents the error of a single prediction. For all models, the absolute value of the prediction error with respect to bilirubin level at the time of the prediction (Figure 2, red dashed lines) tended to increase at higher starting bilirubin levels, increasing from approximately 0.8 to 1.6 mg/dL as the starting bilirubin varied from 0 to 20 mg/dL for the neural network, random forest, and Xgboost models. However, the simple linear and ridge regression models also demonstrated a larger error magnitude at the low range of current bilirubin levels. The blue solid line represents the mean error versus the bilirubin level at the time of prediction. The Xgboost model demonstrates a mean prediction error of near 0 across all bilirubin values at the time of prediction (xgboost panel, blue line). In contrast, the simple linear and ridge regression models tend to predict values that are too high when bilirubin values are low at the time of prediction (blue line >0 at low starting bilirubin levels).

Figure 2. Model prediction errors versus bilirubin level at time of prediction. Each panel depicts the performance of a single predictive model, as described in Textbox 1. Each point represents the error of a single prediction in the test set (n=3320, over 98% visible within ±5 mg/dL error). The curves show the smoothed mean error (blue solid) and mean absolute value of error (red dashed); the gray band is the 95% CI of the mean.



Comparison With Clinician Accuracy

Model performance was also assessed by comparing predictions made by the models with prospective predictions made by clinicians participating in the clinical care of newborn infants. A convenience sample of 210 predictions made by clinicians at 1 hospital was compared with model predictions, all from the held-out test set. The clinicians included attending neonatologists, advanced practitioners (neonatal-perinatal medicine fellows, neonatal nurse practitioners, and pediatric hospitalists with primary responsibilities in the NICU), pediatric residents (interns and seniors), and bedside nurses (in the level 2 and level 3 nurseries). All predictive models other than the negative control had a lower absolute error than the clinician's predictions (Table 2).

Table 2. Absolute errors of clinician and model predictions.

Model[a]	Mean (SD)[b]	Median (IQR)[c]	Clinician error difference[d]	*P* value
clinicians	1.38 (1.31)	1.10 (0.60-1.80)	N/A[e]	N/A
current	1.86 (1.55)	1.50 (0.80-2.58)	–0.49 (–0.68 to –0.29)	<.0001
lm	1.19 (1.03)	0.94 (0.50-1.67)	0.19 (0.04 to 0.34)	.0109
ridge	1.14 (1.01)	0.97 (0.48-1.47)	0.23 (0.10 to 0.36)	.0005
lstm	1.08 (1.01)	0.91 (0.37-1.55)	0.29 (0.14 to 0.44)	.0002
lasso	1.08 (0.95)	0.94 (0.51-1.38)	0.30 (0.17 to 0.43)	<.0001
nn	1.06 (1.02)	0.87 (0.34-1.36)	0.32 (0.18 to 0.45)	<.0001
rf	1.04 (0.91)	0.76 (0.34-1.48)	0.34 (0.20 to 0.48)	<.0001
xgboost	1.01 (0.90)	0.88 (0.37-1.41)	0.37 (0.22 to 0.52)	<.0001

[a]Models are as described in Textbox 1, with 210 predictions made by each model. Clinician predictions were from all role groups (attendings, advanced practitioners, residents, and nurses).

[b]Prediction mean absolute error and SD (mg/dL).

[c]Prediction median absolute error and IQR (mg/dL).

[d]Mean error differences (mg/dL, clinician absolute error minus model absolute error) with 95% confidence range and comparisons by paired *t* test. Positive values reflect higher prediction errors by clinicians.

[e]N/A: not applicable.

Clinician accuracy may differ by role (Table 3), but because predictions were made on different subsets of patients, accuracy by role group could not be directly compared with paired testing. Although advanced practitioners and attendings made predictions with lower mean absolute error (MAE), these predictions were made on measurements for which the simple neural network also had the lowest MAE, that is, this subset may have made it easier to make accurate predictions. When comparing predictions made by clinicians in each role with predictions made by the neural network, clinicians had statistically significant higher errors for all except the nursing group, which had the lowest number of predictions (n=31), potentially limiting statistical power.

Table 3. Clinician prediction accuracy by role and comparison with the neural network predictive model.

Role	Clinician MAE (SD)[a]	Clinician median error (IQR)[b]	Model MAE (SD)	Model median error (IQR)[b]	Mean error difference[c]	*P* value
All clinicians (n=210)	1.38 (1.31)	1.10 (0.60-1.80)	1.06 (1.02)	0.87 (0.34-1.36)	0.32 (0.18 to 0.45)	<.0001
Advanced practitioner (n=74)	1.17 (1.09)	0.90 (0.50-1.40)	0.93 (0.76)	0.83 (0.34-1.31)	0.24 (0.04 to 0.44)	.017
Attending (n=60)	1.36 (1.31)	1.20 (0.57-1.70)	0.99 (1.08)	0.76 (0.25-1.33)	0.37 (0.13 to 0.61)	.003
Resident (n=45)	1.54 (1.20)	1.10 (0.70-1.90)	1.12 (0.87)	1.09 (0.50-1.57)	0.43 (0.12 to 0.73)	.0071
Nurse (n=31)	1.65 (1.82)	1.20 (0.50-1.80)	1.40 (1.49)	1.07 (0.37-1.52)	0.25 (−0.25 to 0.75)	.32

[a]Clinician and neural network model prediction mean absolute error (MAE, mg/dL) and SD.

[b]Clinician and neural network model prediction median absolute error (mg/dL) and IQR.

[c]Mean clinician absolute error minus neural network absolute error (mg/dL), with 95% confidence range and comparisons by paired *t* test. Positive values reflect higher prediction errors by clinicians.

Predicting Exceeding the Phototherapy Threshold

Although many treatment guidelines exist for the management of neonatal hyperbilirubinemia [7], the novelty of the approach described here is that the ability to predict actual bilirubin values allows the model to be adapted for different guidelines.

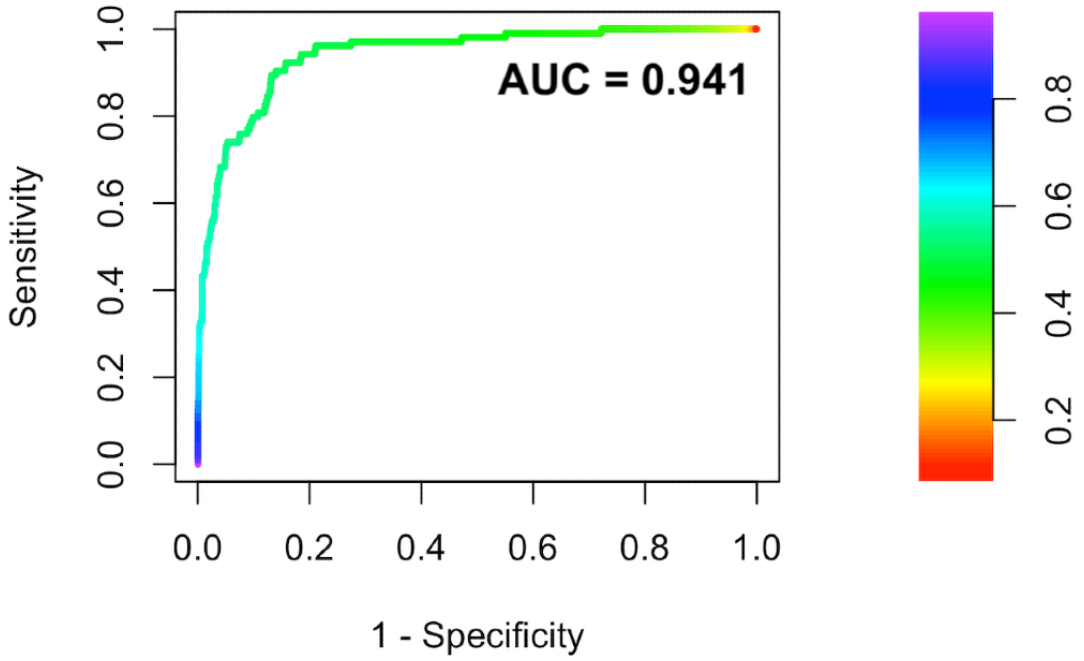
Two previously published models [27,28] predicted the need to resume phototherapy as recommended by consensus treatment guidelines [8], after an initial episode of phototherapy. The guidelines apply only to newborn infants at ≥35 completed weeks of gestation. Both models reported an AUROC curve of 0.88.

As a concrete example of adapting the general approach described in this study to a specific consensus-based guideline, the simple neural network was retrained to make a similar prediction—whether or not the next bilirubin measurement would exceed the phototherapy threshold—using the same 8 predictors described above by (1) limiting the data set to only those babies born at or after 35 weeks and (2) changing the prediction target to the dichotomous outcome exceeding the AAP-recommended phototherapy threshold [8]. The risk

category was determined by gestational age and the presence of potential isoimmune hemolytic disease as reflected by a baby's Coombs-positive result.

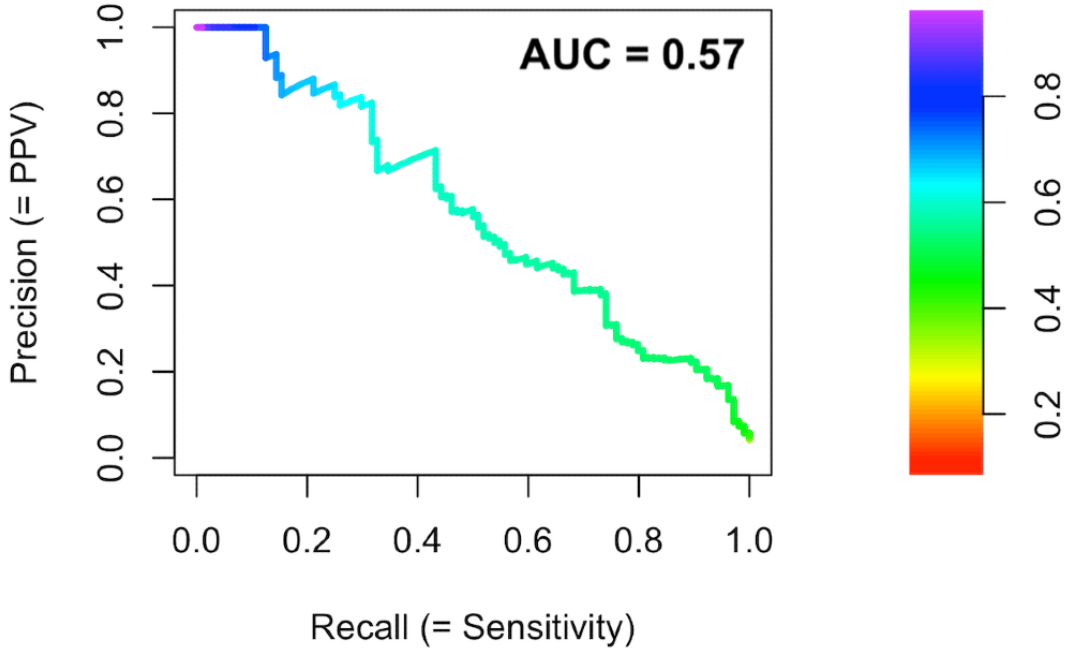
Limiting the data set to only those born after 35 weeks yielded a training set of 19,242 bilirubin prediction targets, of which 910 (910/19,242, 4.73%) exceeded the phototherapy threshold, and a held-out test set of 2449 prediction targets, of which 104 (104/2449, 4.25%) exceeded the phototherapy threshold. After training to make the binomial prediction, this neural network model performed well on the held-out test set with an AUROC curve of 0.941 (95% CI 0.910 to 0.973; Figure 3).

Figure 3. Receiver operator characteristic curve for neural network prediction of exceeding the American Academy of Pediatrics–recommended phototherapy initiation threshold on the subsequent bilirubin measurement in newborn infants ≥35 weeks of gestation. The area under the receiver operator characteristic curve was 0.941 (95% CI 0.910 to 0.973); a no-skill classifier would have an area under the receiver operator characteristic curve of 0.5. The color represents the decision threshold value corresponding to the sensitivity and specificity on the receiver operator characteristic curve. AUROC: area under the receiver operator characteristic curve; ROC: Receiver operator characteristic.



The binary outcome in this data set is significantly imbalanced, with only 4.25% (104/2449) of the test set with a subsequent bilirubin measurement exceeding the phototherapy threshold. Although frequently used to report performance on binary classifiers, the AUROC curve can be misleading in imbalanced data sets for which the area under the precision recall (PR) curve (AUPRC) may be more informative [39,40]. The PR plot displays the relationship between precision (positive predictive value) and recall (sensitivity). Unlike the AUROC, for which a no-skill classifier would have an AUROC curve of 0.5, the no-skill baseline AUPRC varies depending on the class distribution. In this case, the baseline AUPRC for a no-skill classifier would be 0.0425. Figure 4 displays the PR curve performance of the neural network model, again with a good performance on the held-out test set with an AUPRC curve of 0.573.

Figure 4. Precision recall curve for neural network prediction of exceeding the American Academy of Pediatrics–recommended phototherapy initiation threshold on the subsequent bilirubin measurement in newborn infants ≥35 weeks of gestation. The area under the precision recall curve was 0.573; for this data set, a no-skill classifier would have an area under the precision recall curve of 0.043. The color represents the decision threshold value corresponding to the precision and recall on the curve. AUPRC: area under the precision recall curve; PPV: positive predictive value.



Discussion

Principal Findings

Using a cohort of 10,947 babies from 4 hospitals born between 22 and 43 completed weeks of gestation and with 41,695 total serum bilirubin measurements, this study reports the generation and validation of machine learning models to predict follow-up bilirubin levels within 72 hours of a previous measurement during the first 10 days after birth that outperform clinician predictions. A set of 8 predictive features was sufficient for optimal model performance. This may be the first report of predicting specific bilirubin levels in newborn infants at any gestational age when taking into account the effect of phototherapy. This approach was also applied to predict subsequently exceeding the AAP-recommended phototherapy threshold for neonates born at ≥35 weeks of gestation, with good performance.

Potential Applications

Prediction of Exceeding the Phototherapy Threshold of a Clinical Guideline

Previously published models [27,28] predicted the need to resume phototherapy as recommended by consensus treatment guidelines [8] after an initial episode of phototherapy and used logistic regression with only 2 or 3 predictors: gestational age, age at phototherapy initiation, and bilirubin level when phototherapy was discontinued relative to the phototherapy initiation threshold. These models achieved an AUROC curve of 0.88.

The models in this study differ in several ways: prediction of actual bilirubin values rather than a binary outcome, no restriction on gestational age, and taking into account previous episodes of phototherapy or phototherapy before the following measurement. This generalization allows application to any of the many clinical practice guidelines available for the management of neonatal hyperbilirubinemia.

In this study, the AAP 2004 guideline was chosen as a specific example of this approach. Retraining on a data set limited to babies born at ≥35 weeks of gestation and changing the prediction target to the binary outcome resulted in a neural network model with an AUROC curve of 0.941 (95% CI 0.910 to 0.973) on the held-out test set, which compares favorably with the previously reported logistic regression models (AUROC curve of 0.88). The improved performance might be related to taking into account the risk factor for isoimmune hemolytic disease (ie, the presence of baby's Coombs-positive status) that alters the consensus guideline phototherapy initiation thresholds but is not directly included as a predictor in the previously published models.

Predicting subsequently requiring phototherapy when the baseline prevalence is only approximately 4% and with an imperfect predictive model, as demonstrated in the receiver operator characteristic and PR curves (Figures 3 and 4), is challenging and requires a tradeoff between sensitivity, specificity, and positive and negative predictive values.

The choice of a decision threshold depends greatly on the goal of the prediction. For example, if the goal is the relatively low-cost determination of which newborn infants should have a follow-up appointment with their pediatrician sooner rather than later after discharge, a lower decision threshold could be chosen that tolerates a higher false-positive rate. In the predictive model reported here, a decision threshold of 0.3 could be chosen, yielding a lower positive predictive value of 46%, but with a higher sensitivity of 58% and a negative predictive value of 98.1%.

In contrast, if the goal was to determine whether an infant's discharge from the hospital should be delayed to initiate phototherapy, thereby increasing costs related to longer length of stay, a higher threshold might be chosen. For example, choosing a decision threshold of 0.6 for the model presented here would yield a positive predictive value that is increased to 87% but with sensitivity decreased to 25% and a negative predictive value of 96.8%. This choice would be to attempt to avoid prolonging the length of stay by keeping infants who are less likely to actually need phototherapy, but allowing a lower sensitivity and instead relying on outpatient follow-up to identify those infants who would need to be readmitted for phototherapy.

Application to Guidelines to Account for Previous Phototherapy

A more fundamental question is whether the AAP consensus treatment recommendations should be used after phototherapy has already been provided. This usage is not directly addressed or recommended in the 2004 guidelines [8] and there might be issues in the face validity of this practice. For example, early initiation of intensive phototherapy may effectively limit the initial increase in serum bilirubin but this may also result in a potentially falsely-reassuring low (subphototherapy threshold) age-specific bilirubin level, which may be followed by a resumed rapid rate of increase after discontinuation of phototherapy.

The predictive models reported in this study may be useful for developing treatment clinical decision support that predicts the risk of subsequently exceeding consensus-developed thresholds. The present treatment recommendations describe 3 phototherapy initiation curves for different risk categories that plateau at 15, 18, and 21 mg/dL for higher, medium, and lower risk, respectively. The models described here could be used to predict whether a chosen threshold might be exceeded in the future, when taking into account previous phototherapy as well as other clinical features (age, gestational age, empirically observed bilirubin rate of rise, etc).

Predictive Models for Neonatal Readmission

Readmissions of apparently healthy newborn infants are often associated with jaundice. In a retrospective study of 296,114 neonates discharged from 21 well-baby nurseries in the Intermountain Healthcare system, feeding problems (41%) and jaundice (35%) were frequently present in the 5308 early readmissions of apparently healthy neonates [2]. It is possible that the predictions made by the models reported here could be combined with other clinical features to develop a risk calculator for neonatal readmission. This risk assessment might identify higher-risk neonates for closer follow-up with primary care providers, visiting nurses, or lactation consultants. Previous unpublished work in assessing the risk of readmission of apparently healthy newborn infants discharged from a well-baby nursery used gestational age, age at time of discharge, weight loss, size for gestational age (eg, small for gestational age), maternal parity, and maternal race to yield a logistic regression predictive model with fair performance for predicting readmission (AUROC curve of 0.76; Joseph H Chou, unpublished work). An interesting future direction would be to determine whether the addition of present or predicted follow-up bilirubin measurements might improve performance. Ideally, this risk assessment would be performed automatically within the EHR, not requiring clinician input, and made available closer to the time of discharge.

Assessment of Adjunctive Treatment Efficacy

Intravenous immunoglobulin (IVIG) remains a recommended treatment modality for neonatal isoimmune hemolytic disease if total serum bilirubin continues to rise despite intensive phototherapy [8]. However, although some reports suggested a reduction in the need for exchange transfusion after IVIG administration, the practice remains controversial because most clinical trials were not blinded and a recent systematic meta-analysis suggested overall poor quality of evidence and an unknown benefit effect estimate [41]. In nonneonatal populations, IVIG has been rarely associated with worsening of hemolysis [42]; if this phenomenon is present in the neonatal population, there is the possibility of actually worsening jaundice from IVIG therapy.

In the models for bilirubin prediction reported in this study, administration of IVIG was not included as a predictive feature as it was a very rare occurrence (administered in 96 of the 52,149 [0.18%] babies in the starting population). In future work, it would be interesting to determine whether administration of IVIG affected the accuracy of predictions. For example, if IVIG administration was temporally associated with subsequent bilirubin predictions that were consistently too high, this could be interpreted as indirect evidence of IVIG resulting in a lower bilirubin rate of rise. Unlike previous unblinded studies that used avoidance of exchange transfusion as an outcome (albeit a clinically significant outcome), this proposed approach might be less susceptible to bias.

Limitations

Implementation for Clinical Use

The models described in this study are not intended to be used directly by clinicians manually entering predictors, which would likely be too cumbersome for integration into care delivery workflows. Rather, the goal was to generate the best possible performing predictive models using only features easily accessible within the EHR for future integration into automated clinical decision support. Some EHR software providers are beginning to integrate analytics and artificial intelligence modules into their platforms, for example, the Cogito enterprise analytics module by Epic (Epic Systems Corporation) includes business intelligence and machine learning capabilities that is either embedded at the point of care or deployed via a cloud-based platform. A future goal would be to seamlessly provide advanced clinical decision support from within the EHR platform available during care delivery, without clinical provider intervention.

Data Quality and Completeness

The data were limited to those available from routine clinical care, thus predictions outside the norms of clinical care might be less accurate. However, the available data likely reflect the scenarios of highest interest to clinicians. Another concern is the potential for sampling bias. Follow-up bilirubin

measurements may not have been initially planned but were instead obtained after visual recognition of unexpected jaundice, resulting in a bias toward higher bilirubin levels. A similar concern was raised in the AAP guidelines that the Bhutani nomograms should not be considered as describing the natural history of the neonatal bilirubin trend [4,8].

This study intentionally included all study subjects regardless of missing data on the basis that clinicians also often need to make decisions in the face of missing information. The goal of this study was not to generate predictions only if all desired information was available but rather to provide the best predictions possible using the available, potentially incomplete information. For model training, missing data were handled simplistically—median imputation and casting categorical predictors as whether or not known to be present. More sophisticated imputation techniques might yield better prediction performance.

The accuracy of the data extracted from the EHR was another concern. The duration and timing of phototherapy was determined by the timestamps of the clinician orders, which may not reflect actual start and stop times. It was not possible to differentiate between type or intensity of phototherapy or how frequently a baby was permitted to be removed from phototherapy. Of note, for the predictions provided by clinicians, providers actively providing care to the neonates were instructed to take all information into account, even if unavailable for predictive model training.

Some data were not available or were not extracted from the EHR and were therefore not available for model training. For example, glucose-6-phosphate dehydrogenase (G6PD) deficiency can result in jaundice secondary to hemolysis. However, G6PD status was not included as a predictive feature because the results from testing are typically not available in the neonatal time frame and the goal of this study was to generate models usable at the time of neonatal admission. As discussed earlier, IVIG therapy was not included in the model training because of its rarity (96/52,149, 0.18%). Exchange transfusion is another therapy for isoimmune hemolytic jaundice, sometimes utilized after failure of intensive phototherapy and IVIG administration. Exchange transfusion information was not readily extracted from the EHR, but its utilization is likely even less frequent than IVIG administration. As this information was not made available for model training, predictions are unlikely to be accurate in the setting of G6PD deficiency, IVIG administration, or exchange transfusion. However, although more accurate or more complete data for model training might improve prediction accuracy, it is notable that despite potential limitations to data quality, predictive model accuracy still surpassed that of clinicians.

Model Training

In this study, the data were explicitly split into a training set used for data exploration and model parameter choice and a held-out test set used for final model evaluation. A limitation of this approach is that each of the final models was trained once on the training set and evaluated once on the test set, limiting the ability to assess performance variance for each model.

An alternative approach would be to perform, for example, 10-fold cross-validation by combining the data into one large data set, creating 10 overlapping partitions of training and test sets and performing model training on each of the 10 partitions, each resulting in model evaluation on a different test set, allowing a better sense of model performance variance.

However, a major concern with this approach was the potential for data leakage and overestimation of performance. By the time clinician predictions were being collected, the previous data from June 2015 through February 2019 were explored for the initial steps of feature selection, hyperparameter choice (eg, regularization strength), and model architecture (eg, tree ensemble settings, multilayer perceptron structure). If the same data were used to evaluate performance via cross-validated model training, it may result in overly optimistic evaluation metrics.

Instead, the approach was to prevent any possibility of data leakage by completely separating out the post-February 2019 held-out test set and not accessing it until after the models were fully trained on the training set and then reporting model performance on the held-out test set. Another reason for this approach was the limited number of clinician predictions available, all after February 2019.

Future work could use newly acquired data to retrain the predictive models with the previously identified model parameters using K-fold cross-validation to allow a better estimate of model performance and variance.

Model Interpretability

Machine learning model interpretability is a significant issue [43-45]. The risk of trusting uninterpretable predictive models is the potential for failing to recognize when incorrect guidance is being provided. The models trained in this study range from those that are relatively interpretable (linear model with no interaction terms) to those whose functioning is obfuscated (neural network). As is typical in machine learning, a tradeoff between model simplicity and predictive performance was observed. Future work would aim to provide a means to understand model functioning when maintaining prediction accuracy. Another important direction for future work would be to provide confidence ranges for individual predictions.

Limitations of Laboratory Measurement

Nonsystematic laboratory variation of bilirubin analysis limits the achievable prediction accuracy. In a survey of instruments used for neonatal bilirubin measurement, the coefficient of variation (a measure of dispersion used to describe precision) ranged from 2% to 6% [46,47]. In this study, laboratory measurement precision was not evaluated; however, the test set median target bilirubin level was 10.6 mg/dL with a neural network model MAE of 1.05 mg/dL, suggesting that limitations in instrument precision (2% to 6% of 10.6 mg/dL is 0.21 mg/dL to 0.64 mg/dL) might account for a significant proportion of prediction error.

Transcutaneous Bilirubinometry

Transcutaneous bilirubin (TcB) measurement provides a convenient and noninvasive method for estimating serum

bilirubin levels [48]. Nomograms have been developed for normal newborns born at ≥35 completed weeks of gestation [49-51], and a systematic review suggests that TcB measurement is reasonably accurate in the preterm population (born before 37 weeks of gestation) [52].

In this study, TcB data were not included as a predictive feature in model generation as they were noncontributory. In the 4 hospitals included in this study, the clinical practice was to routinely obtain a TcB measurement only in newborn infants born at or after 35 weeks of gestation. The TcB measurement was used mainly as a screening test; if concerning, serum bilirubin was immediately sent and all subsequent management was guided by serum bilirubin measurement.

As the goal of this study was to predict subsequent bilirubin measurements and to include infants born at <35 weeks of gestation and because any concerning TcB measurement was immediately followed by a serum bilirubin measurement, transcutaneous bilirubinometry, as utilized at the 4 hospitals in this study, did not provide additional information useful for model training. However, at other institutions with different practices, TcB measurement is likely to be useful for predictive modeling.

Generalizability

The prediction models were not externally validated to assess generalizability. However, it may be preferable for hospitals to train their own predictive models that account for local equipment and practices. For example, a hospital that routinely uses double overhead fluorescent tube banks of phototherapy as well as a bilirubin blanket under the baby will likely have different phototherapy efficacy compared with using only a single overhead fluorescent tube bank of phototherapy. The increasing accessibility of EHR data and relative ease of machine learning model training may make hospital-specific predictive models possible. Although personalized medicine has typically referred to practice tailored to individual patient variation, personalization should also be applied to hospital systems.

Conclusions

Models were developed to predict follow-up total serum bilirubin levels in newborn infants <10 days old, which outperform clinicians. This may be the first report of models that predict actual bilirubin levels, are not limited to term and late preterm patients, and take into account the effect of phototherapy. The predictive features are readily accessible in EHRs, making integrated clinical decision support potentially feasible. Important directions for future work include improving model interpretability while maintaining prediction accuracy and providing confidence ranges of predictions.

Acknowledgments

This study was not externally funded. The author is grateful to Dr Paul Lerou and Dr Caitlin Li for their support and discussion and to the many clinicians who provided bilirubin predictions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
R code for feature selection and model training.
[DOC File , 42 KB - medinform_v8i10e21222_app1.doc]

Multimedia Appendix 2
Patient cohorts, excluded versus included in the training set.
[DOCX File , 19 KB - medinform_v8i10e21222_app2.docx]

References

1. Maisels MJ. Managing the jaundiced newborn: a persistent challenge. Can Med Assoc J 2015 Mar 17;187(5):335-343. [doi: 10.1503/cmaj.122117] [Medline: 25384650]
2. Young PC, Korgenski K, Buchi KF. Early readmission of newborns in a large health care system. Pediatrics 2013 May;131(5):e1538-e1544. [doi: 10.1542/peds.2012-2634] [Medline: 23569092]
3. Bhutani VK, Wong RJ, Stevenson DK. Hyperbilirubinemia in preterm neonates. Clin Perinatol 2016 Jun;43(2):215-232. [doi: 10.1016/j.clp.2016.01.001] [Medline: 27235203]
4. Bhutani VK, Johnson L, Sivieri EM. Predictive ability of a predischarge hour-specific serum bilirubin for subsequent significant hyperbilirubinemia in healthy term and near-term newborns. Pediatrics 1999 Jan;103(1):6-14. [doi: 10.1542/peds.103.1.6] [Medline: 9917432]
5. Wallenstein MB, Bhutani VK. Jaundice and kernicterus in the moderately preterm infant. Clin Perinatol 2013 Dec;40(4):679-688. [doi: 10.1016/j.clp.2013.07.007] [Medline: 24182955]

6. Maisels MJ, McDonagh AF. Phototherapy for neonatal jaundice. N Engl J Med 2008 Feb 28;358(9):920-928. [doi: 10.1056/NEJMct0708376] [Medline: 18305267]
7. Bratlid D, Nakstad B, Hansen TW. National guidelines for treatment of jaundice in the newborn. Acta Paediatr 2011 Apr;100(4):499-505. [doi: 10.1111/j.1651-2227.2010.02104.x] [Medline: 21114525]
8. American Academy of Pediatrics Subcommittee on Hyperbilirubinemia. Management of hyperbilirubinemia in the newborn infant 35 or more weeks of gestation. Pediatrics 2004 Jul;114(1):297-316. [doi: 10.1542/peds.114.1.297] [Medline: 15231951]
9. Horn AR, Kirsten GF, Kroon SM, Henning PA, Möller G, Pieper C, et al. Phototherapy and exchange transfusion for neonatal hyperbilirubinaemia: neonatal academic hospitals' consensus guidelines for South African hospitals and primary care facilities. S Afr Med J 2006 Sep;96(9):819-824. [Medline: 17068653]
10. Guidelines for detection, management and prevention of hyperbilirubinemia in term and late preterm newborn infants (35 or more weeks' gestation) - Summary. Paediatr Child Health 2007 May;12(5):401-418 [FREE Full text] [doi: 10.1093/pch/12.5.401] [Medline: 19030400]
11. Kaplan M, Merlob P, Regev R. Israel guidelines for the management of neonatal hyperbilirubinemia and prevention of kernicterus. J Perinatol 2008 Jun;28(6):389-397 [FREE Full text] [doi: 10.1038/jp.2008.20] [Medline: 18322551]
12. Rennie J, Burman-Roy S, Murphy MS, Guideline Development Group. Neonatal jaundice: summary of NICE guidance. Br Med J 2010 May 19;340:c2409. [doi: 10.1136/bmj.c2409] [Medline: 20484363]
13. National Institute for Health and Care Excellence (UK). A summary of selected new evidence relevant to NICE clinical guideline. Neonatal jaundice: Evidence Update 2012. [Medline: 31886966]
14. van Imhoff DE, Dijk PH, Hulzebos CV, BARTrial study group, Netherlands Neonatal Research Network. Uniform treatment thresholds for hyperbilirubinemia in preterm infants: background and synopsis of a national guideline. Early Hum Dev 2011 Aug;87(8):521-525. [doi: 10.1016/j.earlhumdev.2011.04.004] [Medline: 21621933]
15. Maisels MJ, Watchko JF, Bhutani VK, Stevenson DK. An approach to the management of hyperbilirubinemia in the preterm infant less than 35 weeks of gestation. J Perinatol 2012 Sep;32(9):660-664 [FREE Full text] [doi: 10.1038/jp.2012.71] [Medline: 22678141]
16. Pillai A, Pandita A, Osiovich H, Manhas D. Pathogenesis and management of indirect hyperbilirubinemia in preterm neonates less than 35 weeks: moving toward a standardized approach. Neoreviews 2020 May;21(5):e298-e307. [doi: 10.1542/neo.21-5-e298] [Medline: 32358143]
17. Dani C, Poggi C, Barp J, Romagnoli C, Buonocore G. Current Italian practices regarding the management of hyperbilirubinaemia in preterm infants. Acta Paediatr 2011 May;100(5):666-669. [doi: 10.1111/j.1651-2227.2011.02172.x] [Medline: 21314845]
18. Wickremasinghe AC, Kuzniewicz MW, McCulloch CE, Newman TB. Efficacy of subthreshold newborn phototherapy during the birth hospitalization in preventing readmission for phototherapy. JAMA Pediatr 2018 Apr 1;172(4):378-385 [FREE Full text] [doi: 10.1001/jamapediatrics.2017.5630] [Medline: 29482208]
19. Yetman RJ, Parks DK, Huseby V, Mistry K, Garcia J. Rebound bilirubin levels in infants receiving phototherapy. J Pediatr 1998 Nov;133(5):705-707. [doi: 10.1016/s0022-3476(98)70117-9] [Medline: 9821435]
20. Maisels MJ, Kring E. Rebound in serum bilirubin level following intensive phototherapy. Arch Pediatr Adolesc Med 2002 Jul;156(7):669-672. [doi: 10.1001/archpedi.156.7.669] [Medline: 12090833]
21. Erdeve O, Tiras U, Dallar Y. Rebound bilirubin measurement is not required for hyperbilirubinemia regardless of the background attributes of newborns. J Trop Pediatr 2004 Oct;50(5):309. [doi: 10.1093/tropej/50.5.309] [Medline: 15510765]
22. Berkwitt A, Osborn R, Grossman M. The utility of inpatient rebound bilirubin levels in infants readmitted after birth hospitalization for hyperbilirubinemia. Hosp Pediatr 2015 Feb;5(2):74-78. [doi: 10.1542/hpeds.2014-0074] [Medline: 25646199]
23. Kaplan M, Kaplan E, Hammerman C, Algur N, Bromiker R, Schimmel MS, et al. Post-phototherapy neonatal bilirubin rebound: a potential cause of significant hyperbilirubinaemia. Arch Dis Child 2006 Jan;91(1):31-34 [FREE Full text] [doi: 10.1136/adc.2005.081224] [Medline: 16223746]
24. Bansal A, Jain S, Parmar VR, Chawla D. Bilirubin rebound after intensive phototherapy for neonatal jaundice. Indian Pediatr 2010 Jul;47(7):607-609 [FREE Full text] [doi: 10.1007/s13312-010-0133-z] [Medline: 20019393]
25. Erdeve O. Rebound bilirubin: on what should the decision to recommence phototherapy be based? Arch Dis Child 2006 Jul;91(7):623; author reply 16223746 [FREE Full text] [Medline: 16790727]
26. James G, Witten D, Hastie T, Tibshirani R, editors. An Introduction to Statistical Learning: With Applications in R. New York, USA: Springer; 2013.
27. Chang PW, Kuzniewicz MW, McCulloch CE, Newman TB. A clinical prediction rule for rebound hyperbilirubinemia following inpatient phototherapy. Pediatrics 2017 Mar;139(3). [doi: 10.1542/peds.2016-2896] [Medline: 28196932]
28. Chang P, Newman T. A simpler prediction rule for rebound hyperbilirubinemia. Pediatrics 2019 Jul;144(1). [doi: 10.1542/peds.2018-3712] [Medline: 31196939]
29. American Academy of Pediatrics Committee on Fetus And Newborn. Levels of neonatal care. Pediatrics 2012 Sep;130(3):587-597. [doi: 10.1542/peds.2012-1999] [Medline: 22926177]
30. Fenton TR, Kim JH. A systematic review and meta-analysis to revise the Fenton growth chart for preterm infants. BMC Pediatr 2013 Apr 20;13:59 [FREE Full text] [doi: 10.1186/1471-2431-13-59] [Medline: 23601190]

31. Chou JH, Roumiantsev S, Singh R. Peditools electronic growth chart calculators: applications in clinical care, research, and quality improvement. J Med Internet Res 2020 Jan 30;22(1):e16204 [FREE Full text] [doi: 10.2196/16204] [Medline: 32012066]
32. Castillo A, Grogan TR, Wegrzyn GH, Ly KV, Walker VP, Calkins KL. Umbilical cord blood bilirubins, gestational age, and maternal race predict neonatal hyperbilirubinemia. PLoS One 2018;13(6):e0197888 [FREE Full text] [doi: 10.1371/journal.pone.0197888] [Medline: 29856776]
33. Bandi C, Vanaki R, Badakali AV, Pol RR, Yelamali B. Predictive value of total serum bilirubin within 6 hour of birth for the development of hyperbilirubinemia after 72 hours of birth. J Clin Diagn Res 2016 Sep;10(9):SC01-SC04 [FREE Full text] [doi: 10.7860/JCDR/2016/16314.8460] [Medline: 27790538]
34. Calkins K, Roy D, Molchan L, Bradley L, Grogan T, Elashoff D, et al. Predictive value of cord blood bilirubin for hyperbilirubinemia in neonates at risk for maternal-fetal blood group incompatibility and hemolytic disease of the newborn. J Neonatal Perinatal Med 2015;8(3):243-250 [FREE Full text] [doi: 10.3233/NPM-15814111] [Medline: 26518407]
35. Sarici SU, Yurdakök M, Serdar MA, Oran O, Erdem G, Tekinalp G, et al. An early (sixth-hour) serum bilirubin measurement is useful in predicting the development of significant hyperbilirubinemia and severe ABO hemolytic disease in a selective high-risk population of newborns with ABO incompatibility. Pediatrics 2002 Apr;109(4):e53. [doi: 10.1542/peds.109.4.e53] [Medline: 11927726]
36. R: A Language and Environment for Statistical Computing. R Core Team. 2020. URL: https://www.r-project.org/ [accessed 2020-10-13]
37. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York, USA: Springer-Verlag; 2009.
38. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982 Apr;143(1):29-36. [doi: 10.1148/radiology.143.1.7063747] [Medline: 7063747]
39. 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]
40. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. 2006 Presented at: CML'06; June 25-29, 2006; Pittsburgh, Pennsylvania URL: https://doi.org/10.1145/1143844.1143874
41. Zwiers C, Scheffer-Rath ME, Lopriore E, de Haas M, Liley HG. Immunoglobulin for alloimmune hemolytic disease in neonates. Cochrane Database Syst Rev 2018 Mar 18;3:CD003313 [FREE Full text] [doi: 10.1002/14651858.CD003313.pub2] [Medline: 29551014]
42. Padmore R. Possible mechanisms for intravenous immunoglobulin-associated hemolysis: clues obtained from review of clinical case reports. Transfusion 2015 Jul;55(Suppl 2):S59-S64. [doi: 10.1111/trf.13090] [Medline: 26174898]
43. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018 Nov 27;19(6):1236-1246 [FREE Full text] [doi: 10.1093/bib/bbx044] [Medline: 28481991]
44. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019 Apr 4;380(14):1347-1358. [doi: 10.1056/NEJMra1814259] [Medline: 30943338]
45. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc 2018 Oct 1;25(10):1419-1428 [FREE Full text] [doi: 10.1093/jamia/ocy068] [Medline: 29893864]
46. Lo SF, Jendrzejczak B, Doumas BT, College of American Pathologists. Laboratory performance in neonatal bilirubin testing using commutable specimens: a progress report on a College of American Pathologists study. Arch Pathol Lab Med 2008 Nov;132(11):1781-1785 [FREE Full text] [doi: 10.1043/1543-2165-132.11.1781] [Medline: 18976015]
47. Lo SF, Doumas BT. The status of bilirubin measurements in U.S. laboratories: why is accuracy elusive? Semin Perinatol 2011 Jun;35(3):141-147. [doi: 10.1053/j.semperi.2011.02.008] [Medline: 21641487]
48. Maisels M. Historical perspectives: transcutaneous bilirubinometry. NeoReviews 2006 May 1;7(5):e217-e225 [FREE Full text] [doi: 10.1542/neo.7-5-e217]
49. Maisels MJ, Kring E. Transcutaneous bilirubin levels in the first 96 hours in a normal newborn population of > or = 35 weeks' gestation. Pediatrics 2006 Apr;117(4):1169-1173. [doi: 10.1542/peds.2005-0744] [Medline: 16585312]
50. de Luca D, Jackson GL, Tridente A, Carnielli VP, Engle WD. Transcutaneous bilirubin nomograms: a systematic review of population differences and analysis of bilirubin kinetics. Arch Pediatr Adolesc Med 2009 Nov;163(11):1054-1059. [doi: 10.1001/archpediatrics.2009.187] [Medline: 19884597]
51. Han S, Yu Z, Liu L, Wang J, Wei Q, Jiang C, Chinese Multicenter Study Coordination Group for Neonatal Hyperbilirubinemia. A model for predicting significant hyperbilirubinemia in neonates from China. Pediatrics 2015 Oct;136(4):e896-e905. [doi: 10.1542/peds.2014-4058] [Medline: 26391945]
52. Nagar G, Vandermeer B, Campbell S, Kumar M. Reliability of transcutaneous bilirubin devices in preterm infants: a systematic review. Pediatrics 2013 Nov;132(5):871-881. [doi: 10.1542/peds.2013-1713] [Medline: 24127472]

Abbreviations

AAP: American Academy of Pediatrics
AUPRC: area under the precision recall curve
AUROC: area under the receiver operator characteristic
G6PD: glucose-6-phosphate dehydrogenase
EHR: electronic health record
IVIG: intravenous immunoglobulin
LASSO: least absolute shrinkage and selection operator
LSTM: long short-term memory
MAE: mean absolute error
NICU: neonatal intensive care unit
PR: precision recall
TcB: transcutaneous bilirubinometry

Edited by C Lovis; submitted 08.06.20; peer-reviewed by R Singh, P Washington; comments to author 10.08.20; revised version received 03.09.20; accepted 27.09.20; published 29.10.20.

Please cite as:
Chou JH
Predictive Models for Neonatal Follow-Up Serum Bilirubin: Model Development and Validation
JMIR Med Inform 2020;8(10):e21222
URL: http://medinform.jmir.org/2020/10/e21222/
doi:10.2196/21222
PMID:33118947

Original Paper

Machine-Learning Monitoring System for Predicting Mortality Among Patients With Noncancer End-Stage Liver Disease: Retrospective Study

Yu-Jiun Lin[1,2], MD; Ray-Jade Chen[3,4], MD, MSc; Jui-Hsiang Tang[5], MD; Cheng-Sheng Yu[1,2*], PhD; Jenny L Wu[1,2], BSc; Li-Chuan Chen[6,7], MSN; Shy-Shin Chang[1,2,6*], MD, PhD

[1]Department of Family Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[2]Department of Family Medicine, Taipei Medical University Hospital, Taipei, Taiwan
[3]Department of Surgery, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[4]Division of General Surgery, Department of Surgery, Taipei Medical University Hospital, Taipei, Taiwan
[5]Division of Gastroenterology and Hepatology, Department of Internal Medicine, Taipei Medical University Hospital, Taipei, Taiwan
[6]Department of Community and Preventive Medicine, Taipei Medical University Hospital, Taipei, Taiwan
[7]School of Gerontology Health Management, College of Nursing, Taipei Medical University, Taipei, Taiwan
[*]these authors contributed equally

Corresponding Author:
Shy-Shin Chang, MD, PhD
Department of Family Medicine
School of Medicine, College of Medicine
Taipei Medical University
250 Wuxing Street
Taipei, 11031
Taiwan
Phone: 886 2 23565926
Email: sschang0529@gmail.com

Abstract

Background: Patients with end-stage liver disease (ESLD) have limited treatment options and have a deteriorated quality of life with an uncertain prognosis. Early identification of ESLD patients with a poor prognosis is valuable, especially for palliative care. However, it is difficult to predict ESLD patients that require either acute care or palliative care.

Objective: We sought to create a machine-learning monitoring system that can predict mortality or classify ESLD patients. Several machine-learning models with visualized graphs, decision trees, ensemble learning, and clustering were assessed.

Methods: A retrospective cohort study was conducted using electronic medical records of patients from Wan Fang Hospital and Taipei Medical University Hospital. A total of 1214 patients from Wan Fang Hospital were used to establish a dataset for training and 689 patients from Taipei Medical University Hospital were used as a validation set.

Results: The overall mortality rate of patients in the training set and validation set was 28.3% (257/907) and 22.6% (145/643), respectively. In traditional clinical scoring models, prothrombin time-international normalized ratio, which was significant in the Cox regression (P<.001, hazard ratio 1.288), had a prominent influence on predicting mortality, and the area under the receiver operating characteristic (ROC) curve reached approximately 0.75. In supervised machine-learning models, the concordance statistic of ROC curves reached 0.852 for the random forest model and reached 0.833 for the adaptive boosting model. Blood urea nitrogen, bilirubin, and sodium were regarded as critical factors for predicting mortality. Creatinine, hemoglobin, and albumin were also significant mortality predictors. In unsupervised learning models, hierarchical clustering analysis could accurately group acute death patients and palliative care patients into different clusters from patients in the survival group.

Conclusions: Medical artificial intelligence has become a cutting-edge tool in clinical medicine, as it has been found to have predictive ability in several diseases. The machine-learning monitoring system developed in this study involves multifaceted analyses, which include various aspects for evaluation and diagnosis. This strength makes the clinical results more objective and reliable. Moreover, the visualized interface in this system offers more intelligible outcomes. Therefore, this machine-learning monitoring system provides a comprehensive approach for assessing patient condition, and may help to classify acute death

patients and palliative care patients. Upon further validation and improvement, the system may be used to help physicians in the management of ESLD patients.

(JMIR Med Inform 2020;8(10):e24305) doi:10.2196/24305

KEYWORDS

visualized clustering heatmap; machine learning; ensemble learning; noncancer-related end-stage liver disease; data analysis; medical information system

Introduction

End-stage liver disease (ESLD) is a major public health problem. It is estimated that 1 million patients died from ESLD globally in 2010, accounting for approximately 2% of all deaths [1-6]. Despite improvements in health care, mortality due to ESLD increased by 65% from 1999 to 2016 [7]. Patients with ESLD have limited treatment options and have a deteriorated quality of life with an uncertain prognosis [8]. Early identification of patients with ESLD who have a poor prognosis is fundamental for palliative care.

Several ESLD risk prediction models have been developed using traditional statistical modeling, including the Child-Pugh score [9], model for end-stage liver disease (MELD) [9,10], adjusted MELD scores (eg, MELD-Na score and integrated MELD score) [11-13], albumin-bilirubin score [14], Chronic Liver Failure Consortium (CLIF) Acute Decompensation Score [15], CLIF Sequential Organ Failure Score [16], CLIF Consortium Acute-on-Chronic Liver Failure Score [17], and a novel score recently developed by our group [18]. Unfortunately, these prediction scores were all found to have poor discrimination between survival and death [19-22]. In addition, these traditional risk scores cannot differentiate patients that need acute care or palliative care.

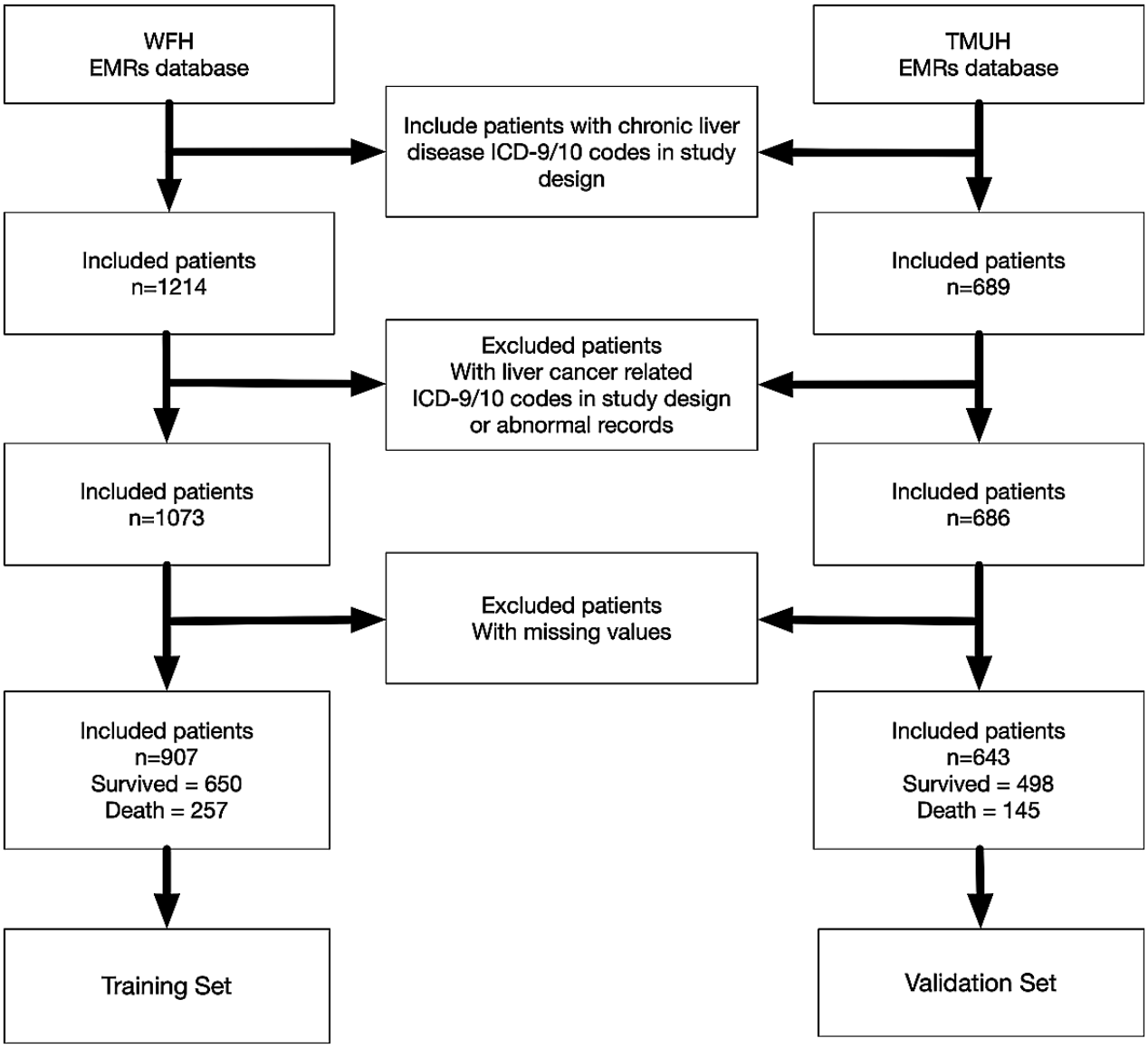
Machine learning, which is the use of computer algorithms that improve automatically through experience, has recently been utilized in disease diagnosis and prediction. In fact, several studies found that machine-learning models have either better or similar performances as traditional statistical modeling approaches [23-26]. Supervised machine-learning models can predict binary disease outcomes, but the prediction accuracy drops when the disease outcome involves several stages. Unsupervised machine-learning models have been successfully utilized to classify diseases that have several stages, such as chronic kidney diseases [27,28]. ESLD is a progressive disease that requires either acute or palliative care. Therefore, the goal of this study was to utilize both supervised and unsupervised machine learning to improve the care of ESLD patients. Specifically, we aimed to create a machine-learning monitoring system that combines several machine-learning models with visualized graphs, including decision trees, ensemble learning methods, and clustering, to predict the mortality of ESLD patients.

Methods

Study Participants and Data Collection

We conducted a retrospective cohort study using the electronic medical records (EMRs) of patients from Wan Fang Hospital and Taipei Medical University (TMU) Hospital (Figure 1). The training dataset comprised patients from Wan Fang Hospital only, whereas the validation set comprised patients from TMU Hospital. By validating our results in different settings, we tried to ensure that the models developed remained valid and robust in different hospitals.

Figure 1. Study flowchart depicting the series of procedures from enrollment to outcome for data collection from patients with noncancer-related end-stage liver disease. WFH: Wan Fang Hospital; TMUH: Taipei Medical University Hospital; EMR: electronic medical record; ICD: International Classification of Diseases.



The study included all adults (aged>18 years) who were diagnosed as having chronic liver diseases with or without related complications of spontaneous bacterial peritonitis, hepatic coma, and esophageal varices (Table 1). In addition, included patients needed to have laboratory EMR data available within 24 hours of admission. Exclusion criteria included pregnancy, cancer, or had a liver transplantation.

Wan Fang Hospital and TMU Hospital are both managed by TMU. The clinical database of TMU includes the EMRs of the two hospitals. The study was approved by the TMU Institutional Review Board (approval number: N202002023) and was conducted in accordance with the Helsinki Declaration.

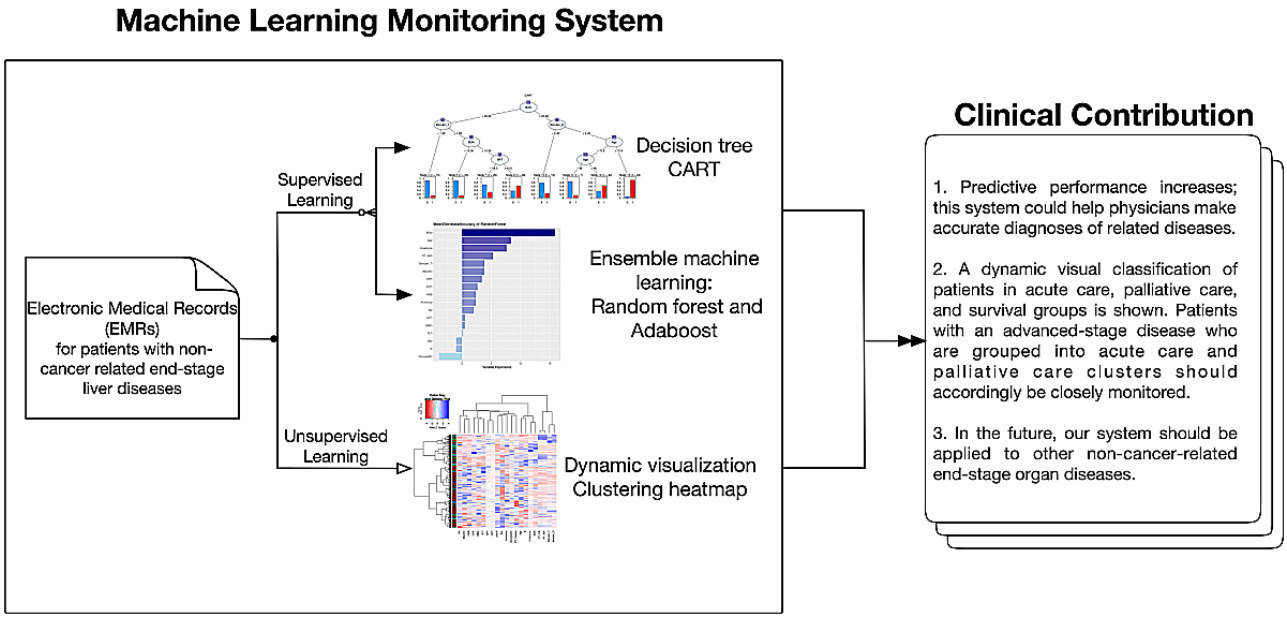
Table 1. International Classification of Diseases (ICD)-9-Clinical Modification (CM) and ICD-10-CM codes for noncancer end-stage liver disease (ESLD).

Diseases	Included in noncancer ESLD	ICD-9-CM code	ICD-10-CM
Cirrhosis	Yes	571.xx	K74.xx
Hepatic coma	Yes	070.xx; 572.xx	K70.xx
Spontaneous bacterial peritonitis	Yes	567.xx	K65.2
Esophageal varices	Yes	456.xx	I85.xx
Malignant neoplasm of the liver	No	155.xx	C22.xx; Z51.12
Liver transplant	No	996.82	Z94.4;T86.4x

Study Overview and Design

The aim of this study was to develop noncancerous liver disease survival prediction models using both traditional statistical modeling approaches and machine-learning approaches (Figure 2). Both supervised and unsupervised machine-learning models were investigated in parallel. For supervised machine learning, the main output was to identify the model with the best survival prediction performance via comparison of the concordance statistic (c-statistic). For unsupervised learning, the main output was the dynamic visualization of ESLD patients to aid in the palliative care of patients. Therefore, ESLD patients were classified into acute death, palliative care, and survived. Acute death was defined as death within 30 days and palliative care was defined as death within 1-9 months from the date of first admission. Mortality was defined using EMR codes related to patient death or critical illness and discharge against medical advice.

Figure 2. Flowchart depicting the structure of the machine learning medical system. CART: classification and regression tree; Adaboost: adaptive boosting.



Data input was based on the literature and the physician's clinical judgment. For example, the following biochemical parameters associated with chronic liver disease were recorded: ammonia, albumin, blood urea nitrogen (BUN), complete blood count, C-reactive protein (CRP), creatinine, glutamic-pyruvic transaminase, prothrombin time (PT) and international normalized ratio (INR), glutamic-oxaloacetic transaminase, serum sodium, serum potassium, and total bilirubin.

Statistical Analysis

Continuous variables were compared by the nonparametric Wilcoxon rank-sum test and categorical variables were compared by the chi-square test.

An initial bivariate analysis was performed to identify significant associations between mortality and all variables available in the study. Significant variables (P<.10) were subsequently tested in a stepwise multivariate logistic regression and stepwise Cox proportional hazards regression to identify independent predictors of mortality (P<.05). The final model for the stepwise regressions was selected as that with the lowest Akaike information criterion.

The validation dataset was used to compare the performances among all models. Performance was assessed according to comparison of receiver operating characteristic (ROC) curves for the different machine-learning models, including random forest with the MELD score, MELD-Na score, and our novel score [18].

All statistical analyses were performed using R (version 3.6.1) and SAS Enterprise Guide (version 7.1) software. For all analyses, P<.05 represented statistical significance.

Machine-Learning Techniques

Machine learning is a statistical-based model that computer systems use to perform a task without using explicit instructions or inferences [29]. In general, machine-learning algorithms can be subdivided into either supervised or unsupervised learning algorithms. Supervised learning involves building a mathematical model of a dataset, termed training data, that contains the inputs and desired outputs known as a supervisory signal. The model is then tested using a validation set. Supervised learning algorithms involve classification and regression. The supervised machine-learning tools utilized in this study included linear discriminant analysis (LDA), support vector machine (SVM), naive Bayes classifier, decision tree, random forest, and adaptive boosting. By contrast, for unsupervised learning, a dataset is taken that contains only inputs and the structure is identified in the data, such as through grouping and clustering.

LDA

LDA is commonly used in multivariate statistical analysis, as it can find a linear combination of features that separates two groups of objects. Hence, LDA is usually used in classification and dimensionality reduction. In this study, LDA was applied

to predict the mortality of patients with chronic liver diseases using the "MASS" package in R [30].

SVM

SVM constructs a hyperplane in a high-dimensional space for classification and regression. The ideal hyperplane will have the largest distance of margins that separates the two groups of objects. SVM is a nonprobabilistic binary classifier, as it can divide two groups of subjects and can assign new events to one group or the other [31].

Naive Bayes Classifiers

Naïve Bayes classifier is based on the Bayes' theorem, with an independence assumption among these features as probabilistic classifiers. Naïve Bayes can be considered a conditional probability model, which assigns a class label according to the maximum a posteriori decision rule [32].

Decision Tree

A decision tree model is a nonparametric and effective machine-learning model. Classification and regression tree (CART) is a typical tree-based model that can predict either a continuous (regression tree) or categorical (classification tree) outcome, and visualizes the decision rule [33]. In decision tree, the Gini index (Equation 1) is used to decide the nodes on a decision branch where p_i represents the relative frequency of the class that is being observed in the dataset and c represents the number of classes. The process of the CART algorithm at each node for classification is as follows: (1) construct a split condition, (2) select a split condition, (3) calculate the impurity by the Gini index (Equation 1), (4) execute steps 2 to 4 until the minimum impurity is selection, and (5) construct the classification in the node.

The Gini index is calculated as:

where p_i is the probability of an object being classified to a particular class i.

In this study, the tree depth of CART was controlled at 4 (ie, maxdepth=4) in the R package to avoid overfitting based on a previous study [26].

Ensemble Learning

Ensemble learning uses multiple learning algorithms to improve machine-learning results, and has generally been found to have better predictive performance than a single model. This is achieved by combining several decision classification and regression tree models [34]. Two types of ensemble learning (random forest and adaptive boosting) were used in this study.

Random Decision Tree

Random forest, a random decision tree model, can extract the most relevant variables by performing classification, regression, or other applications based on a decision tree structure. Parallel methods were used to exploit the independence between the base learners because the error can be minimized by averaging. By creating multiple decision trees and combining the output generated by each tree, the model increases predictive power and reduces bias.

The basic single tree model in random forest is a CART using the Gini index as the selection criterion, and the random forest algorithm applies the bagging technique to implement the teamwork of numerous decision tree models, thereby improving the performance of a single model. The bagging procedure is as follows:

(1) Given a training set $X = x_1, x_2, \ldots, x_n$, with response $Y = y_1, y_2, \ldots, y_n$;

(2) For b = 1, 2, …, B, as the repeated bagging time;

(3) Bagging select a random sample X_b, Y_b with replacement of the training set;

(4) Generate a classification tree from X_b, Y_b;

(5) Prediction for unseen or testing samples z by taking the majority vote from all of the individual classification trees.

The variable importance is determined by the decrease in node impurity, which is weighted by the probability of reaching the node. We determined the node probability by the number of samples that reached the node divided by the total number of samples. Thus, the variable becomes more significant as the value gets higher. The feature importance was implemented by Scikit-learn according to Equations (2) and (3). Assuming a binary tree, Scikit-learn calculates a node's importance using the Gini index.

$$importance\ (n_i) = w_i G_i - w_{left(i)} G_{left(i)} - w_{right(i)} G_{right\ (i)} \quad \textbf{(2)}$$

Where $importance\ (n_i)$ is the importance of node i, w_i is the weighted number of samples reaching node i, G_i is the impurity value of node j, $left(i)$ is the left child node from node i, $right(i)$ is the right child node from node i, and fi_j is the importance of feature j.

The final feature importance at the random forest is the average over all CART tree models after normalization. That is, the sum of the feature's importance values on each tree is divided by the total number of trees [35]. We used the R package randomForest in this study [36].

Adaptive Boosting

Adaptive boosting is an ensemble learning method in which base learners are generated sequentially. It is also used in conjunction with many weak learners (ie, those with poor-performance classifiers) to improve performance. Improving weak learners and creating an aggregated model to improve model accuracy is crucial for boosting algorithm performance. The output of weak learners is combined into a weighted sum that represents the final output of the boosted classifier. Adaptive boosting is adaptive because the motivation for using sequential methods is exploiting the dependence between the base learners. In addition, the predictive ability can be boosted by weighing previously mislabeled examples with

a higher weight. In addition, bagging, a method that combines bootstrapping and aggregating, was used. Because the bootstrap estimate of the data distribution parameters is more accurate and robust, after combining them, a similar method can be used to obtain a classifier with superior properties [37,38]. This study used the "adabag" package for implementing adaptive boosting in R.

ROC

We used ROC curves to compare the mortality predictive performances based on the c-statistic, which is equivalent to the area under the curve (AUC) value. The false positive rate (related to specificity) and the true positive rate (also called sensitivity or recall) were calculated for comparison.

Heatmap and Clustering

A heatmap was used to visualize the pattern of the clinical variables. The clinical and laboratory data of patients are represented as grids of colors with hierarchical clustering analysis applied for both rows and columns [39]. Patients were separated by Euclidean distance (Equation 4) and clustered using the Ward hierarchical clustering algorithm (Equation 5). Clustering can be upgraded using different similarity measures and clustering algorithms [40]. The heatmap was constructed using the "ggplot" package in R. The Euclidean distance between points p and q is the length in multidimensional n-space calculated as:

We followed the general agglomerative hierarchical clustering procedure suggested by the Ward method. The criterion for choosing a pair of clusters to merge at each step is based on the Ward minimum variance method, which can be defined and implemented recursively by a Lance–Williams algorithm [41]. The recursive formula gives the updated cluster distances following the pending merge of clusters. We used the following formula to compute the updated cluster distance:

$$d(C_i \cup C_j, C_k) = a_i d(C_i, C_k) + a_j d(C_j, C_k) + \beta d(C_i, C_j) + \gamma \mid d(C_i, C_k) - d(C_j, C_k) \mid \quad \textbf{(5)}$$

where $d(C_i, C_j)$ is the distance defined between cluster i and cluster j; thus, for each of the metrics we can compute the parameters α_i, α_j, β, and γ.

The Ward minimum variance method can be implemented by the Lance–Williams formula as follows:

$$d(C_i \cup C_j, C_k) = \frac{n_i+n_k}{n_i+n_j+n_k} d(C_i, C_k) + \frac{n_j+n_k}{n_i+n_j+n_k} d(C_j, C_k) - \frac{n_k}{n_i+n_j+n_k} d(C_i, C_j) \quad \textbf{(6)},$$

where n_i, n_j, and n_k is the size of each cluster, a_i is $n_i+n_k/n_i+n_j+n_k$, a_j is $n_j+n_k/n_i+n_j+n_k$, β is $-n_k/n_i+n_j+n_k$, and γ is 0.

The "ggplot" package provides the function to apply heatmap and hierarchical clustering in R. In the function, "scale" was subject to normalization, and "RowSideColors" were set according to the death outcomes.

Results

Figure 1 shows an overview of the study participants and Figure 2 gives an overview of the study. Initially, a total of 1214 patients from Wan Fang Hospital were used to establish a dataset for training and 689 patients from TMU Hospital were used for validation. After data preprocessing (ie, excluding cases with abnormal records and liver cancer cases), the overall mortality rate of patients in the training set at Wan Fang Hospital was 28.3% (257/907) and that at TMU Hospital was 22.6% (145/643). Table 2 and Table 3 summarize the baseline characteristics of all patients according to survival status and separated by the training and validation datasets, respectively.

Table 2. Demographic and laboratory characteristics of patients with noncancerous liver diseases according to survival status.

Demographic variables[a]	Died (n=257)	Survived (n=650)	*P* value
Sex, n (%)			.25
Male	155 (60.3)	419 (64.5)	
Female	102 (39.7)	231 (35.5)	
Age (years)	69 (56-82)	60 (50-72)	<.001
Albumin (g/dL)	2.8 (2.5-3.1)	3.1 (2.8-3.6)	<.001
Ammonia (μg/dL)	55 (34-82)	43 (31-68)	.003
Blood urea nitrogen (mg/dL)	29 (18-52)	16 (12-26)	<.001
Total bilirubin (mg/dL)	1.7 (0.9-4.3)	1.3 (0.7-2.5)	<.001
Direct bilirubin (mg/dL)	1 (0.4-2.9)	0.5 (0.2-1.2)	<.001
Creatinine (mg/dL)	1.2 (0.8-2.2)	0.9 (0.7-1.3)	<.001
C-reactive protein (mg/dL)	5.2 (2.8-8.9)	4.1 (1.2-6.3)	<.001
eGFR[b] (mL/min/1.73 m^2)	54.6 (33.4-60.5)	62 (57-80)	<.001
Glucose ante cibum (mg/dL)	111 (96-139)	107 (94-139)	.31
Serum GOT[c] (U/L)	54 (32-94)	40 (26-72)	<.001
Serum GPT[d] (U/L)	35 (22-59)	33 (21-58)	.42
Hemoglobin (g/dL)	10 (9-11)	12 (10-13)	<.001
Potassium (mEq/L)	4 (3.7-4.4)	3.9 (3.7-4.2)	.001
Sodium (mEq/L)	138 (134-141)	138 (136-139)	.43
Platelets (10^3/μL)	130 (86-177)	162 (110-217)	<.001
PT[e] Control (seconds)	10.8 (10.8-12.5)	11.8 (11.8-12.6)	<.001
PT fibrinogen (seconds)	14.8 (12.7-17.1)	13.5 (12.2-15.1)	<.001
PT international normalized ratio	1.3 (1.13-1.54)	1.15 (1.04-1.29)	<.001
Leukocyte count (10^3/μL)	8.10 (5.99-10.82)	7.02 (5.43-9.28)	<.001

[a]Continuous variables are presented as median (IQR).

[b]eGFR: estimated glomerular filtration rate.

[c]GOT: glutamic-oxaloacetic transaminase.

[d]GPT: glutamic-pyruvic transaminase.

[e]PT: prothrombin time.

Table 3. Demographic and laboratory characteristics of patients with noncancerous liver diseases in the training and validation datasets.

Demographic variables[a]	Training (n=907)	Validation (n=643)	*P* value
Sex, n (%)			.51
Male	574 (63.3)	420 (65.3)	
Female	333 (36.7)	223 (34.7)	
Age (years)	62 (52-75)	61 (51-73)	.11
Albumin (g/dL)	3 (2.7-3.5)	3.3 (3.1-3.7)	<.001
Ammonia (μg/dL)	48 (31-75)	83 (49-116)	<.001
Blood urea nitrogen (mg/dL)	18 (13-33)	16 (12-27)	.003
Total bilirubin (mg/dL)	1.4 (0.8-2.8)	1.5 (0.7-3.1)	.77
Direct bilirubin (mg/dL)	0.6 (0.2-1.7)	1.1 (0.5-2.7)	<.001
Creatinine (mg/dL)	0.9 (0.7-1.5)	0.9 (0.7-1.2)	<.001
C-reactive protein (mg/dL)	4.4 (1.6-7)	3.3 (1.3-4.9)	<.001
eGFR[b] (mL/min/1.73 m^2)	60.5 (47.6-73.5)	94.1 (65.1-123.6)	<.001
Glucose ante cibum (mg/dL)	108 (94-139)	121(104-151)	<.001
Serum GOT[c] (U/L)	43 (27-80)	53 (34-91)	<.001
Serum GPT[d] (U/L)	34 (21-59)	39 (25-64)	.001
Hemoglobin (g/dL)	11 (10-13)	11 (10-13)	.50
Potassium (mEq/L)	4 (3.7-4.2)	3.9 (3.6-4.2)	.003
Sodium (mEq/L)	138 (135-140)	137 (135-139)	.049
Platelets (10^3/μL)	154 (102-209)	138 (87-197)	<.001
PT[e] control (seconds)	11.7 (10.8-12.6)	13.3 (13.2-13.4)	<.001
PT fibrinogen (seconds)	13.8 (12.3-15.6)	15 (13.7-17.4)	<0.001
PT international normalized ratio	1.19 (1.05-1.37)	1.23 (1.08-1.48)	<.001
Leukocyte count (10^3/μL)	7.38 (5.56-9.71)	6.8 (5.28-8.82)	<.001

[a]Continuous variables are presented as median (IQR).

[b]eGFR: estimated glomerular filtration rate.

[c]GOT: glutamic-oxaloacetic transaminase.

[d]GPT: glutamic-pyruvic transaminase.

[e]PT: prothrombin time.

Table 4 shows the risk factors of mortality-based stepwise multivariate logistic and Cox regression analyses for the training dataset. PT-INR, which was significant in the Cox regression, had a prominent influence on predicting mortality. Moreover, BUN and CRP had significant effects on mortality.

Table 4. Significant factors in stepwise multivariate logistic and Cox regression analyses.

Factors	*P* value	Odds ratio/hazard ratio[a] (95% CI)
Stepwise multivariate logistic regression		
Age	<.001	1.029 (1.017-1.042)
Albumin	.002	0.590 (0.421-0.827)
Blood urea nitrogen	.04	1.009 (1.000-1.018)
C-reactive protein	<.001	1.101 (1.056-1.147)
Hemoglobin	<.001	0.795 (0.717-0.882)
Sodium	.02	1.053 (1.007-1.101)
Platelets	<.001	0.995 (0.992-0.997)
Total bilirubin	<.001	1.149 (1.087-1.216)
Leukocyte count	.01	1.075 (1.016-1.137)
Stepwise Cox regression		
Age	.03	1.005 (1.001-1.010)
Blood urea nitrogen	<.001	1.013 (1.009-1.018)
Creatinine	.002	0.920 (0.873-0.969)
C-reactive protein	<.001	1.027 (1.013-1.042)
PT[b] international normalized ratio	<.001	1.288 (1.131-1.468)
Total bilirubin	<.001	1.036 (1.018-1.053)

[a]Odds ratios are reported for logistic regression and hazard ratios are reported for Cox regression.

[b]PT: prothrombin time.

Similar results were obtained using machine-learning methods. Figure 3 shows the variable of importance for random forest and adaptive boosting, which had better performances among all of the supervised machine-learning methods tested (Table 5, Figure 4). BUN was regarded as the primary factor for predicting mortality by both random forest and adaptive boosting models. Creatinine, PT-INR, and bilirubin also emerged as remarkable factors in prediction.

Figure 3. Variable importance ordered by the accuracy of mean decrease in random forest, adaptive boosting (AdaBoost), and AdaBoost + bootstrap. The order of variables is followed by the rank of leading variables in the random forest.

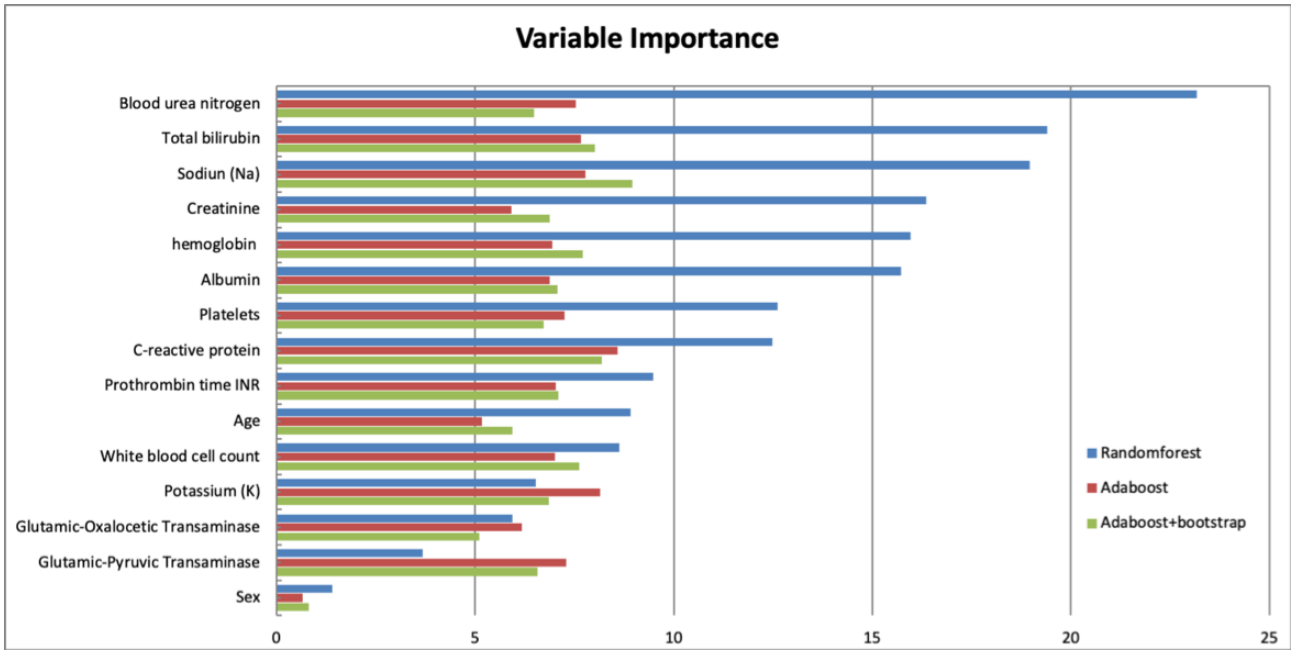


Table 5. Performance of different machine-learning models on predicting mortality of patients with noncancer end-stage liver diseases using the validation dataset.

Model	Accuracy	Sensitivity	Specificity	c-statistic[a]
LDA[b]	0.823	0.701	0.839	0.829
SVM[c]	0.818	0.310	0.966	0.817
Naïve Bayes	0.784	0.290	0.928	0.824
CART[d]	0.790	0.379	0.910	0.744
Random Forest	0.824	0.372	0.956	0.852
Adaboost[e]	0.813	0.455	0.918	0.833

[a]c-statistic: concordance statistic of the receiver operating characteristic curve.

[b]LDA: linear discriminant analysis.

[c]SVM: support vector machine.

[d]CART: classification and regression tree.

[e]AdaBoost: adaptive boosting.

Figure 4. Receiver operating characteristic (ROC) curves with area under the curve (AUC) statistics of classification and regression tree (CART), supervised machine learning (SVM), linear discriminant analysis (LDA), random forest, naive Bayes, and adaptive boosting (Adaboost).

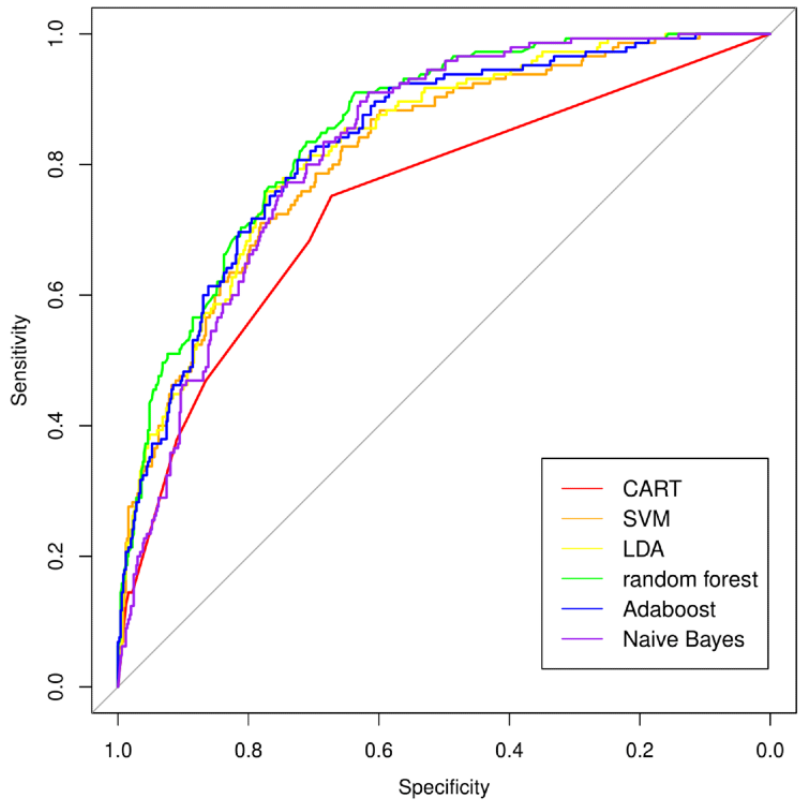


Figure 5 compares the ROC curves for mortality prediction between random forest, as the top-performing machine-learning model, with traditional risk scores. It is clear that random forest (blue curve) had better predictability than all traditional risk scores. However, there were overlaps among traditional risk scores, and it is difficult to differentiate the predictive ability of the MELD score (red, AUC=0.76), MELD-Na (orange, AUC=0.79), and novel score (green, AUC=0.75). Figure 6 shows the calibration plots for the different machine-learning models. The calibration plot is divided into 5 risk strata to match the MELD score. In general, most of the points are close to the diagonal, and the random forest model was found to be better calibrated than other machine-learning techniques. Therefore, the majority of machine-learning models showed better performance (according to the c-statistic in Table 5) than the traditional scoring models. The specificity of each machine-learning model was also above 0.80.

Figure 5. Receiver operator characteristic (ROC) curves with area under the curve (AUC) statistics of random forest, model for end-stage liver disease (MELD) score, MELD-NA score, and novel score.

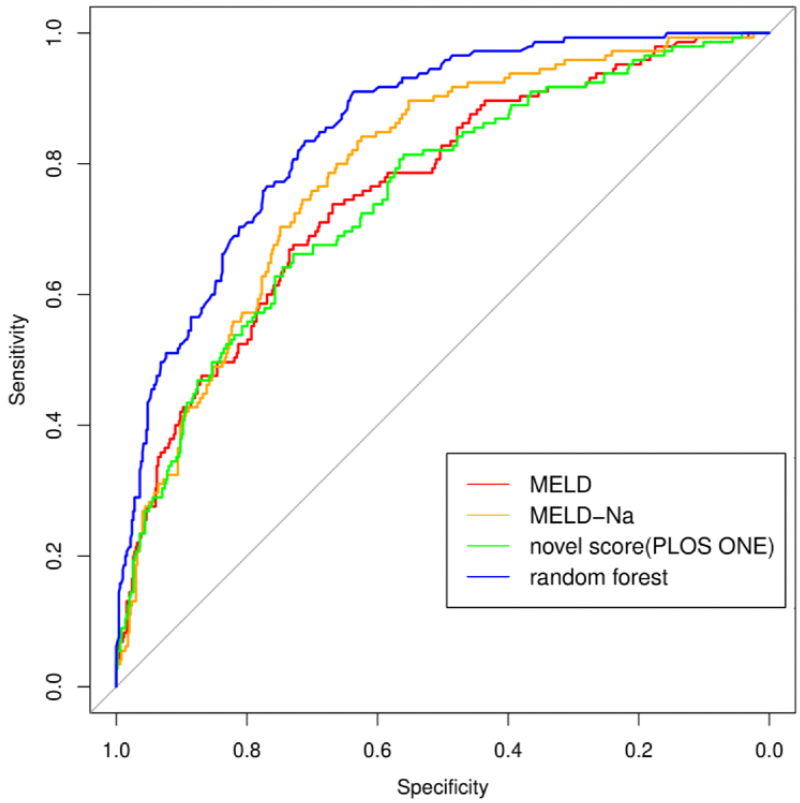
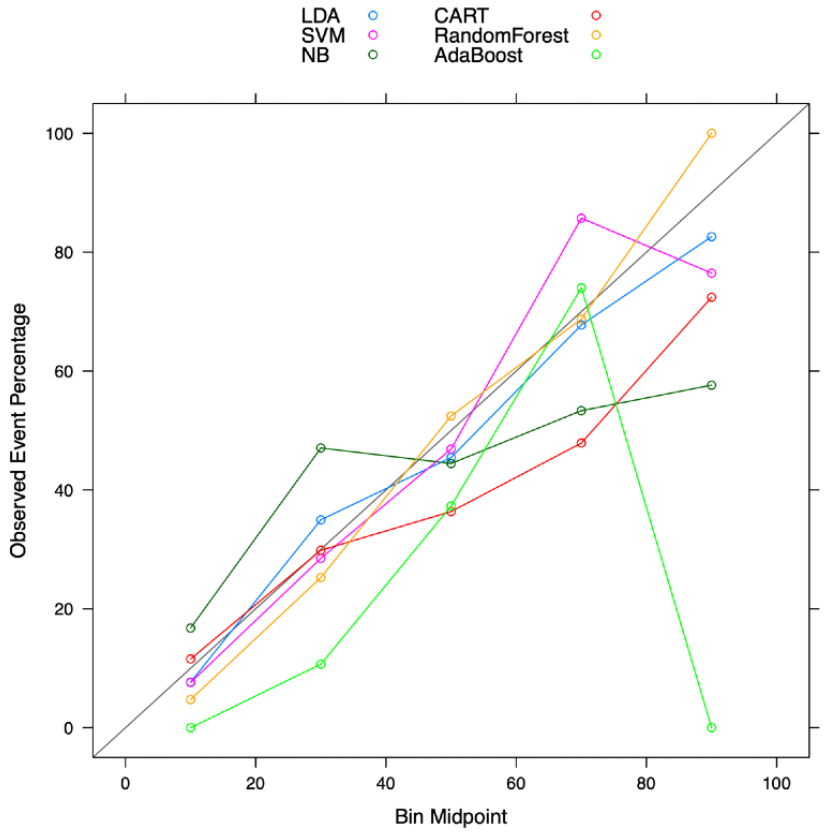
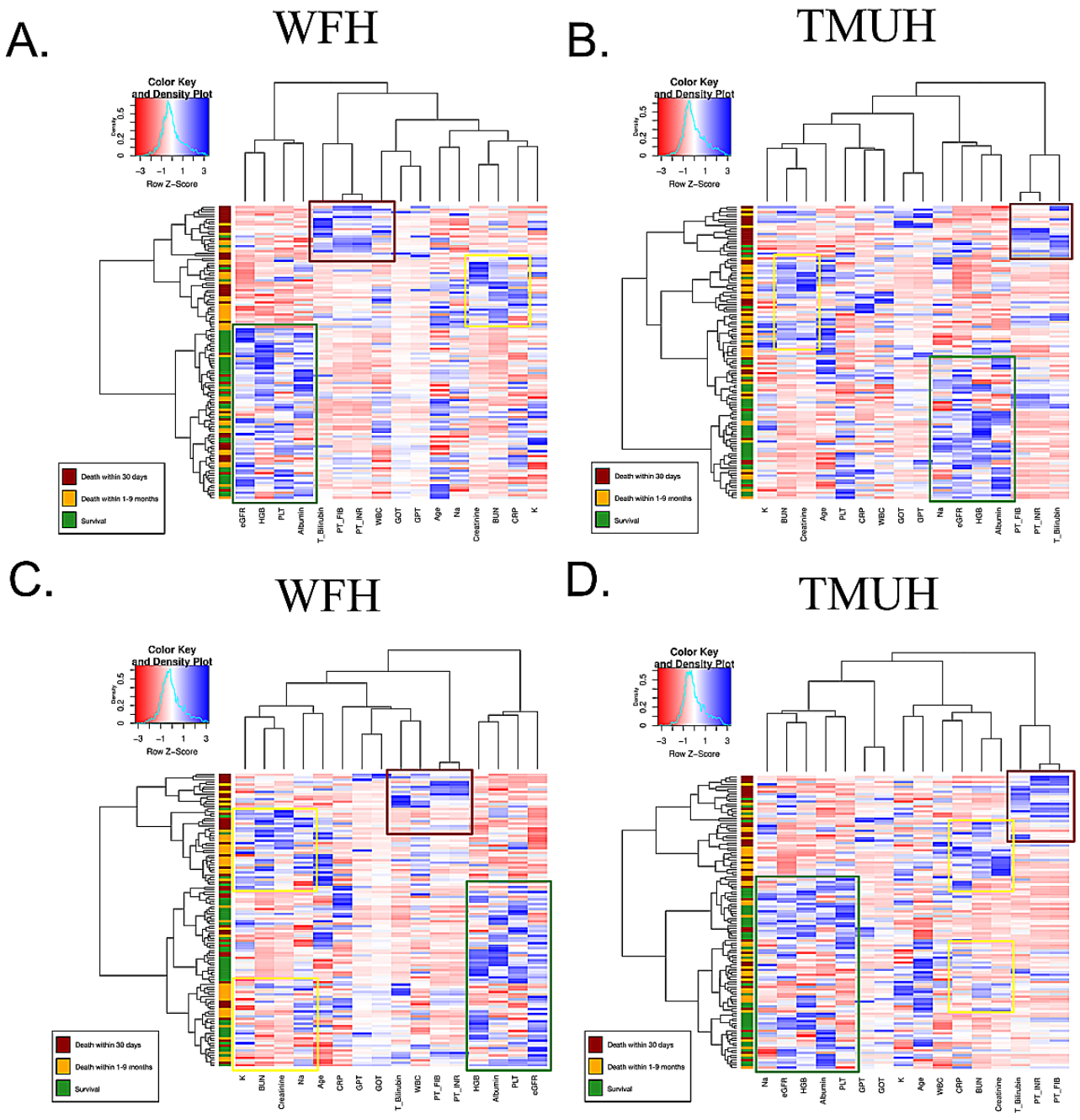


Figure 6. Calibration plots of classification and regression tree (CART), supervised machine learning (SVM), linear discriminant analysis (LDA), random forest, naive Bayes (NB), and adaptive boosting (Adaboost).



In unsupervised machine learning using the heatmap, patients were grouped into death within 30 days (red), death within 1-9 months (yellow), and survival (green) (Figure 7). We found that different clusters had specific color patterns related to laboratory outcomes.

Figure 7. Heatmap showing the classification of acute care (death within 30 days), palliative care (death within 1-9 months), and survival groups in Wan Fang Hospital (WFH) cohort (A and C) and Taipei Medical University Hospital (TMUH) cohort (B and D). BUN: blood urea nitrogen; Bilirubin_T: total bilirubin; CRP: C-reactive protein; eGFR: estimated glomerular filtration rate; GlucoseAC: glucose ante cibum; GOT: serum glutamic-oxaloacetic transaminase; GPT: serum glutamic-pyruvic transaminase; HGB: hemoglobin; K: potassium; Na: sodium; PLT: platelets; PT: prothrombin time; INR: international normalized ratio; WBC: leukocyte count.



Discussion

Principal Findings

A major limitation in traditional statistical modeling is poor predictive ability, especially in nonhomogeneous patients representing several different disease stages. Supervised and unsupervised machine-learning methods are data-driven techniques that have been shown to have either better or similar performances as traditional statistical modeling approaches. In this study, we found that supervised ensemble learning models have better predictive performance than traditional statistical modeling. The AUC of traditional statistical modeling techniques was around 0.75, whereas that of machine-learning techniques was around 0.80. The AUC of the machine-learning technique with the best performance (random forest) was 0.85. In unsupervised learning analysis using hierarchical clustering,

ESLD patients were separated into three clusters: acute death, palliative care, and survived.

Traditional regression analysis showed that PT-INR had the highest odds ratio among all of the significant variables in predicting mortality. This is likely because critically ill patients develop hemostatic abnormalities, and PT-INR has been associated with early death among patients with sepsis-associated coagulation disorders [42]. Similar to previous studies, we also found that BUN and CRP can predict mortality in critically ill patients and for those receiving palliative care [43,44]. A prior study also found that total bilirubin is an excellent predictor of short-term (1-week) mortality in patients with chronic liver failure [45]. High bilirubin levels combined with low albumin levels may be used to predict the severity and progression of liver injury [46,47]. Hyperkalemia (high potassium) and hyponatremia (low sodium) have also been found to increase the mortality risk of ESLD patients [48,49].

In the variable of importance analysis using supervised machine-learning models, BUN was regarded as the primary factor for predicting mortality. This result is in line with a recent study showing that a high BUN concentration is robustly associated with adverse outcomes in critically ill patients, and the results remained robust after correction for renal failure [43]. Interestingly, our variable of importance analysis suggested that BUN might be a more crucial parameter for risk stratification than creatinine level in critically ill patients. We hypothesize that BUN could be an independent risk factor for renal failure, which might indicate neurohumoral activation and disturbed protein metabolism.

In the unsupervised learning analysis, ESLD patients were successfully separated into three clusters. We found that leukocyte count, PT, and bilirubin had specific and similar patterns in the acute death cluster when compared with the palliative care and survival clusters. This is likely related to the fact that these parameters are excellent predictors of short-term mortality and were therefore classified with the acute patient group [42,45]. Acute - on - chronic liver failure (ACLF) is one of the main causes of mortality of ESLD patients. One of the marked pathophysiological features of ACLF is excessive systemic inflammation, which is mainly manifested by a significant increase in the levels of plasma proinflammatory factors, leukocyte count, and CRP [50,51], as observed in our study.

ESLD patients with hepatorenal syndrome typically have the worst prognosis. There are two types of hepatorenal syndrome: type 1 progresses quickly to renal failure, whereas type 2 evolves slowly. Type 2 hepatorenal syndrome is typically associated with refractory ascites and the 3-month survival is 70% [52]. Although BUN, creatinine, sodium, and potassium are indicators of renal function, considering the progression of hepatorenal syndrome, the clustering heatmap classified these parameters in the palliative care group. Thus, visualization of the monitoring system using machine-learning techniques may furnish health care personnel with sufficient relevant information to manage the treatment of patients with chronic liver diseases.

Strengths and Limitations

Medical artificial intelligence has become a cutting-edge tool in clinical medicine, as it has been found to have predictive ability in several diseases. The machine-learning monitoring system developed in this study involves multifaceted analyses, which provide various aspects for evaluation and diagnosis. This strength makes the clinical results more objective and reliable. Moreover, the visualized interface in this system offers more intelligible outcomes.

However, this study has several limitations. First, although this study enrolled thousands of ESLD patients, the numbers of ESLD patients who received palliative care or who experienced acute death were small relative to the number of ESLD patients that have survived. Including data from a larger sample of ESLD patients who received palliative care or who died from acute disease will further improve the accuracy of the machine-learning model in differentiating these three types of ESLD patients. Second, this study enrolled only patients in the Taiwanese population, and the external validity of this study with a cohort of different ethnicity remains to be tested. Third, this was a retrospective study, and a cohort study with prospectively enrolled patients is required to determine the usefulness of our system in clinical practice.

Conclusions and Implications

Our machine-learning monitoring system provides a comprehensive approach for evaluating the condition of patients with ESLD. We found that supervised machine-learning models have better predictive performance than traditional statistical modeling, and the random forest model had the best performance of all models investigated. In addition, our unsupervised machine-learning model may help to differentiate patients that require either acute or palliative care, and may help physicians in their decision in patient treatment. In the future, it will be beneficial to apply our model to several other end-stage organ diseases without the involvement of cancer.

Acknowledgments

This study was supported by the Ministry of Science and Technology Grant (MOST108-2314-B-038-073 and MOST109-2314-B-038-080) and Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan (DP2-109-21121-01-A-10). The funding bodies did not have any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

None declared.

References

1. Bhala N, Aithal G, Ferguson J. How to tackle rising rates of liver disease in the UK. BMJ 2013 Feb 08;346:f807. [doi: 10.1136/bmj.f807] [Medline: 23396387]
2. Cox-North P, Doorenbos A, Shannon SE, Scott J, Curtis JR. The Transition to End-of-Life Care in End-Stage Liver Disease. J Hosp Palliat Nurs 2013;15(4):209-215. [doi: 10.1097/njh.0b013e318289f4b0]
3. da Rocha MC, Marinho RT, Rodrigues T. Mortality Associated with Hepatobiliary Disease in Portugal between 2006 and 2012. GE Port J Gastroenterol 2018 Apr 6;25(3):123-131 [FREE Full text] [doi: 10.1159/000484868] [Medline: 29761148]
4. Mokdad A, Lopez A, Shahraz S, Lozano R, Mokdad A, Stanaway J, et al. Liver cirrhosis mortality in 187 countries between 1980 and 2010: a systematic analysis. BMC Med 2014 Sep 18;12:145 [FREE Full text] [doi: 10.1186/s12916-014-0145-y] [Medline: 25242656]
5. Liver cirrhosis, age-standardized death rates (15+), per 100,000 population. The Global Health Observatory.: World Health Organization URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/liver-cirrhosis-age-standardized-death-rates-(15-)-per-100-000-population [accessed 2020-07-17]
6. Eurostat (European Commission). European social statistics 2013 edition. Luxembourg: Publications Office of the European Union; 2013. URL: https://ec.europa.eu/eurostat/documents/3930297/5968986/KS-FP-13-001-EN.PDF/6952d836-7125-4ff5-a153-6ab1778bd4da
7. Tapper EB, Parikh ND. Mortality due to cirrhosis and liver cancer in the United States, 1999-2016: observational study. BMJ 2018 Jul 18;362:k2817. [doi: 10.1136/bmj.k2817] [Medline: 30021785]
8. Verma M, Tapper EB, Singal AG, Navarro V. Nonhospice Palliative Care Within the Treatment of End-Stage Liver Disease. Hepatology 2020 Jun 30;71(6):2149-2159. [doi: 10.1002/hep.31226] [Medline: 32167615]
9. Peng Y, Qi X, Guo X. Child-Pugh Versus MELD Score for the Assessment of Prognosis in Liver Cirrhosis: A Systematic Review and Meta-Analysis of Observational Studies. Medicine 2016 Feb;95(8):e2877. [doi: 10.1097/MD.0000000000002877] [Medline: 26937922]
10. Kamath PS, Wiesner RH, Malinchoc M, Kremers W, Therneau TM, Kosberg CL, et al. A model to predict survival in patients with end-stage liver disease. Hepatology 2001 Feb;33(2):464-470. [doi: 10.1053/jhep.2001.22172] [Medline: 11172350]
11. Kim WR, Biggins SW, Kremers WK, Wiesner RH, Kamath PS, Benson JT, et al. Hyponatremia and mortality among patients on the liver-transplant waiting list. N Engl J Med 2008 Sep 04;359(10):1018-1026 [FREE Full text] [doi: 10.1056/NEJMoa0801209] [Medline: 18768945]
12. Lai JC, Covinsky KE, Dodge JL, Boscardin WJ, Segev DL, Roberts JP, et al. Development of a novel frailty index to predict mortality in patients with end-stage liver disease. Hepatology 2017 Aug;66(2):564-574 [FREE Full text] [doi: 10.1002/hep.29219] [Medline: 28422306]
13. Luca A, Angermayr B, Bertolini G, Koenig F, Vizzini G, Ploner M, et al. An integrated MELD model including serum sodium and age improves the prediction of early mortality in patients with cirrhosis. Liver Transpl 2007 Aug;13(8):1174-1180. [doi: 10.1002/lt.21197] [Medline: 17663415]
14. Chen RC, Cai YJ, Wu JM, Wang XD, Song M, Wang YQ, et al. Usefulness of albumin-bilirubin grade for evaluation of long-term prognosis for hepatitis B-related cirrhosis. J Viral Hepat 2017 Mar;24(3):238-245. [doi: 10.1111/jvh.12638] [Medline: 27862671]
15. Moreau R, Jalan R, Gines P, Pavesi M, Angeli P, Cordoba J, CANONIC Study Investigators of the EASL–CLIF Consortium. Acute-on-chronic liver failure is a distinct syndrome that develops in patients with acute decompensation of cirrhosis. Gastroenterology 2013 Jun;144(7):1426-1437. [doi: 10.1053/j.gastro.2013.02.042] [Medline: 23474284]
16. Jalan R, Pavesi M, Saliba F, Amorós A, Fernandez J, Holland-Fischer P, CANONIC Study Investigators; EASL-CLIF Consortium. The CLIF Consortium Acute Decompensation score (CLIF-C ADs) for prognosis of hospitalised cirrhotic patients without acute-on-chronic liver failure. J Hepatol 2015 Apr;62(4):831-840. [doi: 10.1016/j.jhep.2014.11.012] [Medline: 25463539]
17. Jalan R, Saliba F, Pavesi M, Amoros A, Moreau R, Ginès P, CANONIC study investigators of the EASL-CLIF Consortium. Development and validation of a prognostic score to predict mortality in patients with acute-on-chronic liver failure. J Hepatol 2014 Nov;61(5):1038-1047. [doi: 10.1016/j.jhep.2014.06.012] [Medline: 24950482]
18. Tsai YW, Tzeng IS, Chen YC, Hsieh TH, Chang SS. Survival prediction among patients with non-cancer-related end-stage liver disease. PLoS One 2018;13(9):e0202692 [FREE Full text] [doi: 10.1371/journal.pone.0202692] [Medline: 30240398]
19. Kim HJ, Lee HW. Important predictor of mortality in patients with end-stage liver disease. Clin Mol Hepatol 2013 Jun;19(2):105-115 [FREE Full text] [doi: 10.3350/cmh.2013.19.2.105] [Medline: 23837134]
20. Said A, Williams J, Holden J, Remington P, Gangnon R, Musat A, et al. Model for end stage liver disease score predicts mortality across a broad spectrum of liver disease. J Hepatol 2004 Jun;40(6):897-903. [doi: 10.1016/j.jhep.2004.02.010] [Medline: 15158328]
21. Moraes ACOD, Oliveira PCD, Fonseca-Neto OCLD. The Impact of the Meld Score on Liver Transplant Allocation and Results: an Integrative Review. Arq Bras Cir Dig 2017;30(1):65-68. [doi: 10.1590/0102-6720201700010018] [Medline: 28489174]

22. Tsai YW, Chan YL, Chen YC, Cheng YH, Chang SS. Association of elevated blood serum high-sensitivity C-reactive protein levels and body composition with chronic kidney disease: A population-based study in Taiwan. Medicine 2018 Sep;97(36):e11896. [doi: 10.1097/MD.0000000000011896] [Medline: 30200074]
23. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA 2018 Apr 03;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
24. Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. N Engl J Med 2017 Jun 29;376(26):2507-2509 [FREE Full text] [doi: 10.1056/NEJMp1702071] [Medline: 28657867]
25. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 2017 Jun 14;38(23):1805-1814 [FREE Full text] [doi: 10.1093/eurheartj/ehw302] [Medline: 27436868]
26. Yu CS, Lin YJ, Lin CH, Wang ST, Lin SY, Lin SH, et al. Predicting Metabolic Syndrome With Machine Learning Models Using a Decision Tree Algorithm: Retrospective Cohort Study. JMIR Med Inform 2020 Mar 23;8(3):e17110 [FREE Full text] [doi: 10.2196/17110] [Medline: 32202504]
27. Yu CS, Lin CH, Lin YJ, Lin SY, Wang ST, Wu JL, et al. Clustering Heatmap for Visualizing and Exploring Complex and High-dimensional Data Related to Chronic Kidney Disease. J Clin Med 2020 Feb 02;9(2):403 [FREE Full text] [doi: 10.3390/jcm9020403] [Medline: 32024311]
28. Yu CS, Lin YJ, Lin CH, Lin SY, Wu JL, Chang SS. Development of an Online Health Care Assessment for Preventive Medicine: A Machine Learning Approach. J Med Internet Res 2020 Jun 05;22(6):e18585 [FREE Full text] [doi: 10.2196/18585] [Medline: 32501272]
29. Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.
30. Abdi H. Discriminant correspondence analysis. In: Salkind N, editor. Encyclopedia of Measurement and Statistics. Thousand Oaks: Sage; 2007:270-275.
31. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/BF00994018]
32. Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach. Upper Saddle River: Prentice Hall; 2009.
33. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Boca Raton: CRC Press; 2017.
34. Rokach L. Ensemble-based classifiers. Artif Intell Rev 2009 Nov 19;33(1-2):1-39. [doi: 10.1007/s10462-009-9124-7]
35. Breiman L. Bagging predictors. Mach Learn 1996 Aug;24(2):123-140. [doi: 10.1007/bf00058655]
36. Breiman L. Random forests. Mach Learn 2001;45:5-32. [doi: 10.1023/A:1010933404324]
37. Freund Y, Schapire R. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J Comput Syst Sci 1997 Aug;55(1):119-139. [doi: 10.1006/jcss.1997.1504]
38. Alfaro E, Gámez M, García N. ADABAG: An R package for classification with boosting and bagging. J Stat Soft 2013;54(2). [doi: 10.18637/jss.v054.i02]
39. Perrot A, Bourqui R, Hanusse N, Lalanne F, Auber D. Large interactive visualization of density functions on big data infrastructure. 2015 Presented at: 2015 IEEE 5th Symposium on Large Data Analysis and Visualization (LDAV); October 25-26, 2015; Chicago, IL. [doi: 10.1109/LDAV.2015.7348077]
40. Ward JH. Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc 1963 Mar;58(301):236-244. [doi: 10.1080/01621459.1963.10500845]
41. Gordon AD. Classification, 2nd Edition. Boca Raton: Chapman & Hall/CRC; 2019.
42. Iba T, Arakawa M, Ohchi Y, Arai T, Sato K, Wada H, et al. Prediction of Early Death in Patients With Sepsis-Associated Coagulation Disorder Treated With Antithrombin Supplementation. Clin Appl Thromb Hemost 2018 Dec;24(9_suppl):145S-149S. [doi: 10.1177/1076029618797474] [Medline: 30198317]
43. Arihan O, Wernly B, Lichtenauer M, Franz M, Kabisch B, Muessig J, et al. Blood Urea Nitrogen (BUN) is independently associated with mortality in critically ill patients admitted to ICU. PLoS One 2018;13(1):e0191697 [FREE Full text] [doi: 10.1371/journal.pone.0191697] [Medline: 29370259]
44. Amano K, Maeda I, Morita T, Miura T, Inoue S, Ikenaga M, et al. Clinical Implications of C-Reactive Protein as a Prognostic Marker in Advanced Cancer Patients in Palliative Care Settings. J Pain Symptom Manage 2016 May;51(5):860-867 [FREE Full text] [doi: 10.1016/j.jpainsymman.2015.11.025] [Medline: 26826676]
45. López-Velázquez JA, Chávez-Tapia NC, Ponciano-Rodríguez G, Sánchez-Valle V, Caldwell SH, Uribe M, et al. Bilirubin alone as a biomarker for short-term mortality in acute-on-chronic liver failure: an important prognostic indicator. Ann Hepatol 2013;13(1):98-104 [FREE Full text] [Medline: 24378272]
46. Chen B, Lin S. Albumin-bilirubin (ALBI) score at admission predicts possible outcomes in patients with acute-on-chronic liver failure. Medicine 2017 Jun;96(24):e7142. [doi: 10.1097/MD.0000000000007142] [Medline: 28614241]
47. Ma T, Li QS, Wang Y, Wang B, Wu Z, Lv Y, et al. Value of pretransplant albumin-bilirubin score in predicting outcomes after liver transplantation. World J Gastroenterol 2019 Apr 21;25(15):1879-1889 [FREE Full text] [doi: 10.3748/wjg.v25.i15.1879] [Medline: 31057301]
48. Fernández-Esparrach G, Sánchez-Fueyo A, Ginès P, Uriz J, Quintó L, Ventura PJ, et al. A prognostic model for predicting survival in cirrhosis with ascites. J Hepatol 2001 Jan;34(1):46-52. [doi: 10.1016/s0168-8278(00)00011-8] [Medline: 11211907]

49. Borroni G, Maggi A, Sangiovanni A, Cazzaniga M, Salerno F. Clinical relevance of hyponatraemia for the hospital outcome of cirrhotic patients. Digest Liver Dis 2000 Oct;32(7):605-610 [FREE Full text] [doi: 10.1016/S1590-8658(00)80844-0]
50. Garcia-Martinez R, Caraceni P, Bernardi M, Gines P, Arroyo V, Jalan R. Albumin: pathophysiologic basis of its role in the treatment of cirrhosis and its complications. Hepatology 2013 Nov;58(5):1836-1846. [doi: 10.1002/hep.26338] [Medline: 23423799]
51. Mahmud N, Kaplan DE, Taddei TH, Goldberg DS. Incidence and Mortality of Acute-on-Chronic Liver Failure Using Two Definitions in Patients with Compensated Cirrhosis. Hepatology 2019 May;69(5):2150-2163 [FREE Full text] [doi: 10.1002/hep.30494] [Medline: 30615211]
52. de Mattos Â, de Mattos AA, Méndez-Sánchez N. Hepatorenal syndrome: Current concepts related to diagnosis and management. Ann Hepatol 2016;15(4):474-481 [FREE Full text] [Medline: 27236146]

Abbreviations

ACLF: acute-on-chronic liver failure
AUC: area under the curve
BUN: blood urea nitrogen
CART: classification and regression tree
CLIF: Chronic Liver Failure Consortium
CRP: C-reactive protein
c-statistic: concordance statistic of the receiver operating characteristic curve
EMR: electronic medical record
ESLD: end-stage liver disease
INR: international normalized ratio
LDA: linear discriminant analysis
MELD: model for end-stage liver disease
PT: prothrombin time
ROC: receiver operating characteristic
SVM: support vector machine
TMU: Taipei Medical University

Edited by G Eysenbach; submitted 14.09.20; peer-reviewed by MT Lee, JY Wu; comments to author 21.09.20; revised version received 25.09.20; accepted 30.09.20; published 30.10.20.

Please cite as:
Lin YJ, Chen RJ, Tang JH, Yu CS, Wu JL, Chen LC, Chang SS
Machine-Learning Monitoring System for Predicting Mortality Among Patients With Noncancer End-Stage Liver Disease: Retrospective Study
JMIR Med Inform 2020;8(10):e24305
URL: http://medinform.jmir.org/2020/10/e24305/
doi:10.2196/24305
PMID:33124991

Original Paper

Exploring Eating Disorder Topics on Twitter: Machine Learning Approach

Sicheng Zhou[1], MSc; Yunpeng Zhao[2], MSc; Jiang Bian[2], PhD; Ann F Haynos[3], PhD; Rui Zhang[1,4], PhD

[1]Institute for Health Informatics, University of Minnesota, Minneapolis, MN, United States

[2]Department of Health Outcomes & Biomedical Informatics, University of Florida, Gainsville, FL, United States

[3]Department of Psychiatry, University of Minnesota, Minneapolis, MN, United States

[4]Department of Pharmaceutical Care & Health Systems, University of Minnesota, Minneapolis, MN, United States

Corresponding Author:
Rui Zhang, PhD
Institute for Health Informatics
University of Minnesota
8-100 Phillips-Wangensteen Building
516 Delaware Street SE
Minneapolis, MN, 55455
United States
Phone: 1 612 626 4209
Email: zhan1386@umn.edu

Abstract

Background: Eating disorders (EDs) are a group of mental illnesses that have an adverse effect on both mental and physical health. As social media platforms (eg, Twitter) have become an important data source for public health research, some studies have qualitatively explored the ways in which EDs are discussed on these platforms. Initial results suggest that such research offers a promising method for further understanding this group of diseases. Nevertheless, an efficient computational method is needed to further identify and analyze tweets relevant to EDs on a larger scale.

Objective: This study aims to develop and validate a machine learning–based classifier to identify tweets related to EDs and to explore factors (ie, topics) related to EDs using a topic modeling method.

Methods: We collected potential ED-relevant tweets using keywords from previous studies and annotated these tweets into different groups (ie, ED relevant vs irrelevant and then promotional information vs laypeople discussion). Several supervised machine learning methods, such as convolutional neural network (CNN), long short-term memory (LSTM), support vector machine, and naïve Bayes, were developed and evaluated using annotated data. We used the classifier with the best performance to identify ED-relevant tweets and applied a topic modeling method—Correlation Explanation (CorEx)—to analyze the content of the identified tweets. To validate these machine learning results, we also collected a cohort of ED-relevant tweets on the basis of manually curated rules.

Results: A total of 123,977 tweets were collected during the set period. We randomly annotated 2219 tweets for developing the machine learning classifiers. We developed a CNN-LSTM classifier to identify ED-relevant tweets published by laypeople in 2 steps: first relevant versus irrelevant (F_1 score=0.89) and then promotional versus published by laypeople (F_1 score=0.90). A total of 40,790 ED-relevant tweets were identified using the CNN-LSTM classifier. We also identified another set of tweets (ie, 17,632 ED-relevant and 83,557 ED-irrelevant tweets) posted by laypeople using manually specified rules. Using CorEx on all ED-relevant tweets, the topic model identified 162 topics. Overall, the coherence rate for topic modeling was 77.07% (1264/1640), indicating a high quality of the produced topics. The topics were further reviewed and analyzed by a domain expert.

Conclusions: A developed CNN-LSTM classifier could improve the efficiency of identifying ED-relevant tweets compared with the traditional manual-based method. The CorEx topic model was applied on the tweets identified by the machine learning–based classifier and the traditional manual approach separately. Highly overlapping topics were observed between the 2 cohorts of tweets. The produced topics were further reviewed by a domain expert. Some of the topics identified by the potential ED tweets may provide new avenues for understanding this serious set of disorders.

(JMIR Med Inform 2020;8(10):e18273) doi:10.2196/18273

KEYWORDS

eating disorders; topic modeling; text classification; social media; public health

Introduction

Background of Eating Disorders and Social Media

Eating disorders (EDs) are a prevalent type of mental illness affecting more than 30 million people across different age groups in the United States [1]. These disorders are commonly underdiagnosed and undertreated [2], and even among individuals who receive diagnosis and treatment, recovery takes a long time to achieve and remains elusive to many [3,4]. Unfortunately, there are serious consequences associated with EDs. Affected individuals often experience significant negative psychological, physical, and interpersonal effects of ED symptoms [5]. Although evidence-based interventions are available for EDs, they are not helpful for many, suggesting that they may not be targeting the correct psychological variables for these individuals [6]. Thus, it is important to gather additional information on the thoughts, emotions, and behaviors of individuals with EDs to identify treatment targets to improve or develop effective interventions for these populations [7].

During the past decade, the number of users of social media platforms, such as Twitter and Facebook, has increased sharply. These platforms provide the general public with opportunities to express their thoughts and opinions and share information about their daily lives, including their health information. This practice has yielded a large number of social media messages that may provide valuable information on a variety of health topics. The analysis of these messages could produce knowledge, permitting more sensitive and accurate education and intervention design in different areas of public health [8]. As such, applying data mining techniques to analyze Twitter data has become a popular methodological approach in health care research.

For instance, a study in 2011 used the Ailment Topic Aspect Model, incorporated with previous knowledge, to create structured disease information from tweets that was subsequently used for the surveillance of a series of different ailments [8]. In 2016, Xu et al [9] checked the frequency of discussions on cancer-related topics among Twitter users and found differences among different race and ethnicity groups. In 2019, Musaev et al [10] applied a latent Dirichlet allocation (LDA) topic model to tweets to examine public discussions about cardiovascular disease and found that state health departments play an important role in communicating with the public about cardiovascular health. We have also conducted a series of studies using Twitter data on various health-related topics, from detecting adverse events to studying laypeople's discussion of human papillomavirus vaccination [11-17]. We have also conducted initial work on using Twitter data to identify discussion topics relevant to EDs. These studies indicate that the analysis of tweets on a particular health topic, coupled with information derived from user profiles, may facilitate novel knowledge discovery in these areas of medicine. EDs have become a popular topic on Twitter. Thus, data mining methods and tools that can more effectively and efficiently identify and analyze these tweets and associated user profiles may help advance the research on ED-related content on social media. In addition, these methods can be translated for use in research on social media use relevant to health-related topics.

ED-Related Research and Gaps

Due to the self-protective nature of EDs, many individuals may not be willing to communicate about their experience of the disorder with others, potentially limiting the ability of researchers and clinicians to understand the factors that promote ED symptoms [18,19]. However, many individuals with an ED use social media to engage in a more open discourse about ED content with others with shared experiences [20]. Although data on Twitter are publicly available, few studies have explored public discussions about EDs. One study investigated how some Twitter accounts promote ED symptoms and the associated negative health consequences among Twitter users. They manually collected data from 45 ED-promotion Twitter accounts, including the profile information, the tweets posted by these accounts, and information about their followers. Through content analysis, they identified a list of ED-related keywords in these tweets and found a positive correlation between the percentages of ED-relevant tweets posted by the ED-promotion accounts and their followers [19]. Another study collected and reviewed ED-relevant tweets and manually classified the collected tweets into different subgroups to provide insights on EDs and to inform future web-based interventions for EDs [21]. These studies indicate that analysis of ED-relevant tweets may help to provide insight into factors that motivate ED behaviors, which may further be used to prevent and treat EDs. However, these studies mainly used keyword-searching strategies to collect ED-relevant tweets and analyzed the content of tweets through manual review. This approach is limited because it only permits analysis of a relatively small number of tweets within a limited time frame with compromised efficiency in content analysis. As a result, the obtained information may not be sufficiently comprehensive, and manual analyses may not be scalable. To improve these studies, computational methods are needed to identify and analyze ED-relevant tweets.

Objective of the Study

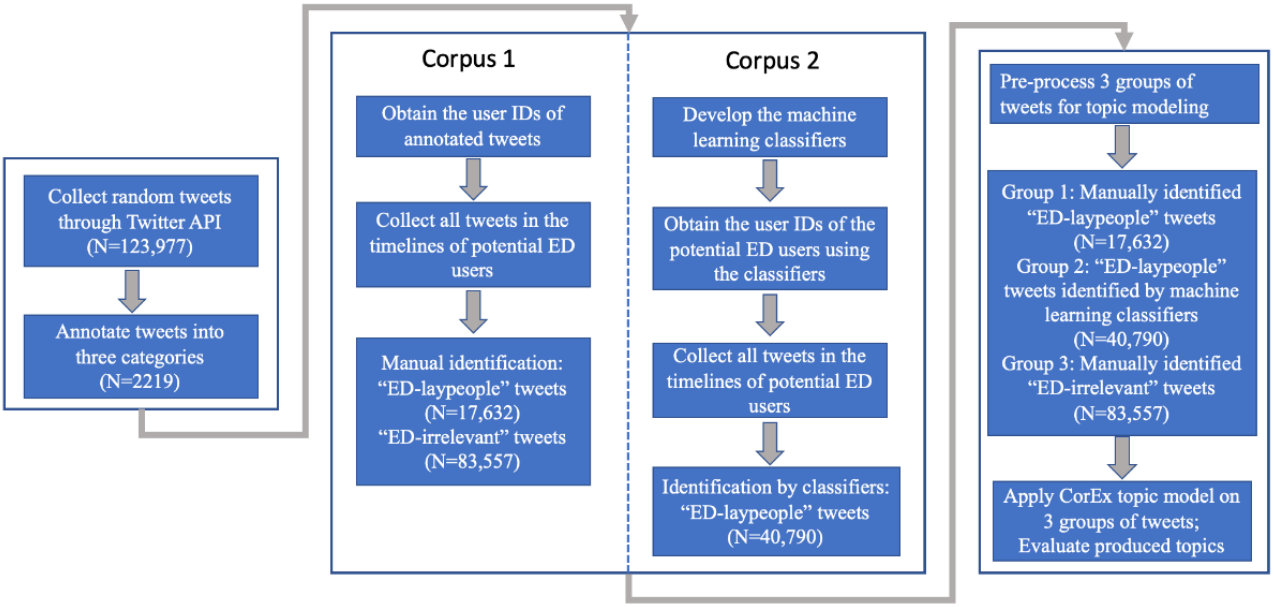
To expand upon previous research on social media engagement among individuals with EDs, the focus of this study is to develop an automatic approach to better understand public perceptions and thoughts about EDs and ED-related behaviors using Twitter data. Specifically, a machine learning approach was developed to identify ED-relevant tweets, and a topic modeling method was implemented to analyze the content of the identified tweets. Potential ED-related factors, such as behaviors, thoughts, and mental status, were summarized through content analysis.

Methods

Overview of Experimental Pipeline

The overall experimental pipeline is shown in Figure 1. We used 2 data sets spanning September 2012 to October 2019. We then randomly selected and annotated 2219 tweets to develop a training data set for the machine learning tasks of identifying ED-relevant tweets. We explored different machine learning–based methods to filter out ED-irrelevant tweets and classified the remaining tweets into either promotional and educational information or laypeople's discussions. We determined if a tweet was published by a *layperson* on the basis of the content of the tweets. For example, if the tweet content was purely about an advertisement, the tweet was considered likely to be published by a *nonlayperson*. Thereafter, we performed topic modeling on 2 corpora. One corpus was built manually on the basis of 2219 annotated tweets and regarded as the gold standard for tweets that contain laypeople's discussions about EDs. Another corpus of ED-relevant tweets was built using the developed machine learning classifiers. We evaluated the topic modeling results to validate whether the corpus built by our developed classifiers could produce similar ED-relevant topics compared with the manually identified gold standard corpus. We also further identified and analyzed topics from the ED-irrelevant tweets posted by the potential ED users.

Figure 1. The workflow of the study. N represents the number of tweets at the corresponding step. CorEx: Correlation Explanation; ED: eating disorder.



Twitter Data Collection

The Twitter data used in this study were from 2 different sources: (1) we used a list of ED keywords we developed from previous studies [19,21] (Textbox 1) and used these keywords to search for potential ED-relevant tweets from a database of random tweets collected from January 1, 2012, to September 30, 2018, using the Twitter streaming application programming interface (API); and (2) we used the same list of keywords to collect more Twitter data by using the Twitter search API from September 26, 2019, to October 30, 2019. This time span was selected because we have been collecting Twitter data since 2012.

Textbox 1. List of eating disorder–relevant keywords refined from previous studies.

Eating disorder–relevant keywords:

- Anorexia
- Anorexic
- Anamia
- ana/mia
- Anahelp
- Anabuddy
- Anaprobs
- Binge
- Bulimia
- bulimic
- #BingeEatingDisorder
- #CompulsiveEating
- ednos
- edlogic
- edprobs
- edproblems
- pro ana
- proana
- pro mia
- promia
- purge

Identification of Target Tweets

Annotation of Tweets

Of the 123,977 total tweets, we randomly selected 2219 tweets on the basis of keyword distribution for gold-standard data set development. The tweets were annotated into 3 categories: *ED-irrelevant*, *ED-promotional and education*, and *ED-laypeople*. *ED-irrelevant* tweets were tweets considered irrelevant to EDs, *ED-promotional and education* tweets were tweets considered to be published by companies and institutes to promote their products or educate the public about EDs, and *ED-laypeople* tweets were considered to be ED-relevant tweets posted by individual users. The tweets labeled with *ED-laypeople* were our target tweets, and the individual users who posted the *ED-laypeople* tweets were defined as potential ED users. Owing to the nature of social media, it is extremely difficult to determine which user does indeed have an ED. However, a large portion of these users were highly engaged in the discussion of ED symptoms or frequently posted their activities and thoughts about EDs and how ED symptoms affected their lives. Two annotators started with 100 tweets to develop an initial annotation guideline to identify the category of each tweet on the basis of its content. Thereafter, we annotated another 100 tweets to refine this guideline. Finally, each of the remaining 2019 tweets were annotated by 2 coders. Agreements were calculated, and conflicts were resolved through group discussions.

Manually Identified ED-Laypeople Tweets and ED-Irrelevant Tweets

Within the 2219 tweets, we extracted the Twitter user accounts that posted tweets in the *ED-laypeople* category. We manually reviewed the usernames and parts of their tweets and removed users whose usernames and tweets indicated that they were not *laypeople*; for example, if a username contained the name of a company or an institute, the tweets from that account would be classified as not belonging to the *ED-laypeople* category. Through this process, 31 accounts were removed for being companies or institutes. The remaining accounts were classified as potential ED users. We then collected all the tweets posted by these potential ED users to construct their Twitter timelines. We checked all of their tweets to see if these tweets contained one of the ED keywords in Textbox 1. If so, the tweets were regarded as *ED-laypeople* tweets, otherwise as *ED-irrelevant* tweets. The *ED-laypeople* tweets and the *ED-irrelevant* tweets were further analyzed using a topic modeling method. Although these *ED-irrelevant* tweets were classified as not directly associated with EDs, they were considered to reflect general topics within the potential ED users' daily lives, which could help to differentiate ED-related experiences from other aspects of these users' lives.

Machine Learning Classifier Development

Supervised text classifiers can learn patterns from annotated input samples and automatically classify tweets into desired categories. To develop the classifiers, we first preprocessed the annotated tweets by replacing (1) hyperlinks (eg, *http://t.co/xxxx*) with <*url*>, (2) mentions (eg, *@username*) with <*user*>, (3) hashtags (eg,*#eatdisorder*) with <*hashtag*> *eatdisorder*, and (4) emojis with <*emojies*>. Then we explored 2 supervised deep learning algorithms (ie, convolutional neural network [CNN] and long short-term memory [LSTM]) and 5 machine learning algorithms (ie, naïve Bayes [NB], linear regression, support vector machine [SVM], random forest [RF], and gradient boosting trees [GB]) to classify our large tweet corpus automatically. We developed our classifiers in a 2-step process using the annotated tweets so that each classifier produced a binary output. In the first task, classifiers were trained to distinguish between ED-relevant and ED-irrelevant tweets (*ED-irrelevant* vs the union of *ED-promotional and education* and *ED-laypeople*) to filter out irrelevant tweets. In the second task, classifiers were trained to distinguish between the *ED-promotional and education* and *ED-laypeople* tweets. We developed a CNN classifier for the first task. The architecture of the CNN model included an embedding layer, a convolutional layer, a global max pooling layer, and a sigmoid output dense layer. We initialized the embedding layer with the Global Vectors for Word Representation pretrained 200-dimension Twitter word embeddings. In the convolutional layer, we set the number of filters to 64, the length of filters to 3, and the dropout rate to 0.2. For the second task, we developed an LSTM model. The architecture of the LSTM model included an embedding layer, an LSTM layer, a global max pooling layer, and a sigmoid output dense layer.

Identifying Topics From Tweets of Potential ED Users

To understand the mental status and everyday life of potential ED users, we applied topic modeling to explore their tweets. Topic modeling has been a popular method for identifying latent patterns of words in a large collection of documents [22]. The most representative method for topic modeling is LDA—a probabilistic generative model [23]. In LDA, each document is assumed to contain a mixture of topics, where each topic is a probability distribution over the words in the document [24]. Some new topic models were developed to solve some of the limitations in LDA, such as the Biterm Topic Model (BTM) and the Correlation Explanation (CorEx) model [24,25]. BTM mainly improves the LDA's problem of sparse word co-occurrence patterns at the document level; thus, it uses the term *co-occurrence patterns* in the entire corpus to learn topics [24]. The CorEx model does not have an assumption about how the underlying data are generated, similar to LDA, which avoids assigning the characteristics of topics ahead of time. The CorEx model identifies the topics that are *maximally informative* about a collection of documents [25]. The BTM and the CorEx model were tested in our preliminary study [17], and the CorEx model was adopted because it produced more meaningful topics using our collected tweets.

We implemented the CorEx model on 3 groups of Twitter data: (1) the *ED-laypeople* tweets identified through manually curated rules posted by potential ED users, (2) the *ED-irrelevant* posted by potential ED users, and (3) *ED-laypeople* tweets identified by the machine learning algorithm, as mentioned earlier. For each group, we tested the CorEx model with different numbers of topics (n=50, 60, 70, 80, 90, 100). Although quantitative metrics are used to infer a reasonable number of topics (eg, perplexity and coherence), they sometimes cannot identify the optimal number of topics. On the basis of our experience in a previous study [17], we manually reviewed the topics produced by the different experiments to determine the optimal number of topics for further topic evaluation.

Topic Evaluation

The results of the topic modeling experiments were further reviewed and analyzed by the domain expert (AH). Three steps were taken to evaluate the topic modeling results. First, the domain expert summarized the theme for each topic on the basis of the topic keywords. Second, on the basis of the top 10 most-relevant tweets for each topic, the expert judged whether each tweet was coherent with the summarized topic theme. The coherence rate of each topic, defined as the percentage of coherent tweets per topic, was calculated. Finally, topic themes with similar meanings were merged into higher-level categories.

Results

Tweet Collection and Annotation

Two coders annotated 2219 tweets into 3 classes, as mentioned earlier. Within 2219 tweets, 669 tweets were annotated as *ED-irrelevant*, 579 tweets were annotated as *ED-promotional and education*, and 971 tweets were annotated as *ED-laypeople*. The interrater agreement score between the 2 annotators was 0.84 on the basis of the first 200 tweets. We used the Cohen kappa test to calculate the score.

Identification of the Target Tweets

Manual Identification of ED-Laypeople and ED-Irrelevant Tweets

As described earlier, we manually identified 17,632 *ED-laypeople* tweets and 89,312 *ED-irrelevant* tweets posted by the potential ED users.

Machine Learning Classifier Development

As mentioned earlier, 7 classifiers were explored (ie, CNN, LSTM, NB, LN, SVM, RF, and GB). We developed our classifiers in a 2-step process (ie, *ED-irrelevant* vs the other 2 labels and then *ED-promotional and education* vs *ED-laypeople*). Overall, 79.99% (1775/2219) of the tweets were used as the training set, and 20.01% (444/2219) tweets were used for evaluation. Table 1 shows the performances of the classifiers.

Table 1. Performances of the developed classifiers.

Classifier	Precision	Recall	F_1 score	*P* value
***ED*[a]*-irrelevant* versus other 2 labels[b]**				
CNN[c]	*0.88*	*0.89*	*0.89*	N/A[d]
LSTM[e]	0.86	0.89	0.88	.15
NB[f]	0.85	0.73	0.75	<.001
LN[g]	0.84	0.78	0.81	<.001
SVM[h]	0.87	0.83	0.85	<.001
RF[i]	0.86	0.85	0.86	.005
GB[j]	0.77	0.75	0.76	<.001
***ED-promotional and education* versus *ED-laypeople*[k]**				
LSTM	*0.90*	*0.89*	*0.90*	N/A
CNN	0.87	0.87	0.87	.006
NB	0.80	0.74	0.76	<.001
LN	0.83	0.80	0.81	<.001
SVM	0.82	0.79	0.80	<.001
RF	0.84	0.82	0.83	<.001
GB	0.84	0.82	0.83	<.001

[a]ED: eating disorder.

[b]ED-irrelevant versus other 2 labels: in this task, the performances of CNN and LSTM have no significant difference; they are both significantly higher than the others (*P*<.01).

[c]CNN: convolutional neural network.

[d]N/A: not applicable.

[e]LSTM: long short-term memory.

[f]NB: naïve Bayes.

[g]LN: linear regression.

[h]SVM: support vector machine.

[i]RF: random forest.

[j]GB: gradient boosting trees.

[k]ED-promotional and education versus ED-laypeople: in this task, the performance LSTM is significantly higher than the others (*P*<.01).

In the first task, the CNN outperformed the other classifiers (F_1 score=0.89). The CNN classifier identified 88,261 tweets that were ED-relevant. In the second task, LSTM obtained the best performance (F_1 score=0.90). Thus, we adopted LSTM for the second task. The LSTM method identified 40,790 *ED-laypeople* tweets posted by 21,600 Twitter users.

CorEx Topic Model Implementation

The CorEx topic model was implemented on 3 groups of Twitter data, as mentioned earlier. After preprocessing, the first group (*ED-laypeople* tweets manually identified) contained 17,632 tweets. The second group (*ED-laypeople* tweets identified by the developed CNN-LSTM classifier) contained 40,790 tweets. The third group (*ED-irrelevant* tweets posted by the potential ED users) contained 83,557 tweets. After the initial review, the optimal number of topics was determined to be 70 for groups 1 and 2. For group 3, the optimal number of topics was determined to be 80. Textbox 2 shows the representative words of selected topics obtained from 3 groups of tweets.

Textbox 2. Representative words of selected topics obtained from 3 groups of tweets.

Weight loss

- Weight, lose, lost, gain, lb

Eating disorder symptoms

- Purge, binge, crave, buffet, bathroom

Food and drink

- Coke, breakfast, sandwich, chicken, yogurt

Body image

- Collar bone, thigh, fat, mirror

Media or advertising or portrayals

- Instagram, twitter, media, social, tumblr

Mental illness

- Mental, ill, breakdown, think, disorder

Negative consequences

- Sleep, hunger, stress, escape, pain

Negative emotions

- Depress, apart, alone, sad, pointless

Education or awareness or treatment

- Therapist, session, save, visit, came

Recovery

- Battle, strength, courage, stronger, inspiring

Topics Evaluation

The CorEx model results of the 3 experimental groups were reviewed and analyzed by a domain expert (AH). For group 1, which used manually identified *ED-laypeople* tweets, 54 of 70 topics were identified as meaningful, and each of them was assigned a topic theme. Similar themes were further grouped into 15 higher-level categories. The top 10 relevant tweets of each topic were reviewed and judged whether they were coherent with the summarized topic theme, and the coherence rate was calculated. Table 2 lists the summary of group 1, including the identified higher-level topic categories, the number of topic themes under each category, some representative topic themes, and the coherence rates for each topic category.

For the second group that used *ED-laypeople* tweets identified by the developed classifier, 63 of 70 topics were identified as meaningful topics and were assigned topic themes. The 63 topics were further merged into 19 categories. Table 3 shows a summary of the topics in the second group, including the identified topic categories, the number of topic themes under each category, example of representative topic themes, and the coherence rates for each category.

For the third group, which used manually identified *ED-irrelevant* tweets posted by the potential ED users, 47 of 80 topics were reviewed as significant topics. The 47 topics were further merged into 19 categories. Table 4 shows a summary of the results from group 3.

Compared with our previous study [17], several new topics were identified, including *Questions or Concerns*, *Reflection or Planning*, *Comorbidity*, *Ambivalence*, *Insults*, and *Diagnostic criteria*.Textbox 3 shows these topic themes and example tweets.

We also explored the ED-irrelevant tweets posted by the potential ED users (ie, group 3). Selected topics and relevant tweets are listed in Textbox 4.

Table 2. A summary of the topics using group 1 tweets (ie, manually identified ED-laypeople tweets).

Topic category	Population, n	Number of topics under each category	Representative topic themes	Coherence rate, n (%)
ED[a] recovery	90	9	Learning from the past; Hope; Moving forward	69 (77)
ED symptoms	70	7	Weight loss and gain; Binge-eating and purging	61 (87)
Education or awareness or treatment	60	6	ED education; ED treatment	51 (85)
Random words	60	6	Love; Big; Life; Rock	48 (80)
Negative consequences	50	5	Health damage; Feeling trapped	36 (72)
Body image	50	5	Collar bones; Thinness	39 (78)
Food and drink	30	3	Food and drink	23 (77)
Pro-ana	30	3	Pro-ana	24 (80)
Negative emotions	30	3	Guilt and shame; Fear; Sadness	23 (77)
Media or advertising or portrayals	20	2	Media and advertising	12 (60)
Comorbidity	10	1	Comorbidity	10 (100)
Reflection or planning	10	1	Reflection or planning	9 (90)
Ambivalence	10	1	Ambivalence	8 (80)
Diagnostic criteria	10	1	Diagnostic criteria	7 (70)
Questions or concerns	10	1	Questions or concerns	10 (100)

[a]ED: eating disorder.

Table 3. A summary of the topics using group 2 tweets (ie, ED-laypeople tweets identified by the developed classifiers).

Topic category	Population, n	Number of topics under each category	Representative topic themes	Coherence rate, n (%)
ED[a] symptoms	100	10	Restriction; Appetite suppression; Binge-eating; Purging	65 (65.0)
Education or awareness or treatment	90	9	ED education; Support group	76 (84.4)
Media or advertising or portrayals	90	9	Media or advertising	65 (72.2)
ED recovery	80	8	Passion; Hope; Love	63 (78.8)
Negative consequences	70	7	Health damage; Social	44 (62.9)
Food and drink	20	2	Food and drink	13 (65.0)
Social media	20	2	Twitter; Social media	11 (55.0)
Pro-ana	20	2	Pro-ana	16 (80.0)
Insults	20	2	Insults	18 (90.0)
Reflection or planning	20	2	Reflection or planning	17 (85.0)
Comorbidity	20	2	Comorbidity	19 (95.0)
Mental illness	10	1	Mental illness	9 (90.0)
Negative emotions	10	1	Negative emotions	8 (80.0)
Body image	10	1	Body image	8 (80.0)
Weight extremes	10	1	Weight extremes	5 (50.0)
Negative social reactions	10	1	Negative social reactions	8 (80.0)
Diagnosis	10	1	Diagnosis	9 (90.0)
Questions or concerns	10	1	Questions or concerns	6 (60.0)
Random words	10	1	Anger	9 (90.0)

[a]ED: eating disorder.

Table 4. A summary of the topics using group 3 tweets (ie, manually identified ED-irrelevant tweets posted by the potential ED users).

Topic category	Population, n	Number of topics under each category	Representative topic themes	Coherence rate, n (%)
Negative emotions or attitude	80	8	Negative emotion; Pressure; Hate	68 (85)
ED[a] behaviors	70	7	Restriction; Purging; Laxatives	55 (79)
Body image	50	5	Hair; Appearance	36 (72)
Exercise	40	4	Exercise	26 (65)
Media	30	3	Entertainment; Music	20 (67)
Self-harm	20	2	Self-harm	17 (85)
Negative consequences	20	2	Mental health damage	19 (95)
Communication	20	2	Seeking communication; Connection	12 (60)
Weight loss	20	2	Weight loss; Concern about weight	18 (90)
Positive emotions	20	2	Positive emotions; Encouraging	14 (70)
Holidays	20	2	Christmas; Halloween	15 (75)
Food and drink	10	1	Food and drink	9 (90)
Social media	10	1	Social media	8 (80)
Suicide	10	1	Desire for suicide	9 (90)
Appreciation	10	1	Appreciation	10 (100)
Sleep deprivation	10	1	Sleep deprivation	10 (100)
Grocery or shopping	10	1	Grocery or shopping	6 (60)
Intimate relationships	10	1	Intimate relationships	7 (70)
School	10	1	Negative social reactions	6 (60)

[a]ED: eating disorder.

Textbox 3. New topics identified from ED-laypeople tweets in groups 1 and 2.

Comorbidity

- "#mentalhealthawareness We're hiring. I myself suffer with 6 mental health illnesses. Anxiety, Depression, OCD, BPD, Anorexia and PTSD"
- "Got my full diagnosis list. I have major depression with psychotic features, GAD, anorexia nervosa, binge eating, purging type, ptsd, and bpd"
- "Literally my everyday life, so many diagnosis, ocd, did, bulimia, anorexia ....anxiety depression until 11 years later "bpd" or eupd as they call it in uk, its hard to tell ppl more should be done to raise awareness so much praise for @xxxxxxx"

Reflection or planning

- "A few months ago I was hiding in my dorm room severely depressed and relapsing from anorexia; tomorrow morning marks one month of me taking recovery seriously; is also my move-in day at a new school & I am so excited to get healthy again and for a new start and I am truly happy"
- "So yesterday was a struggle with Candy it was all around constantly and I caved . :(but I will get back on track today. I will pretend to be sick on thanksgiving and on Christmas Eve that way I don't get tempted by candy .... I'm done with looking at that scale go up. #proana"
- "Live for today, let yesterday go, and keep smiling for tomorrow. #Anorexia #ED #life #strength #dontgiveup"

Insults

- "Net of 0 today. My mom made me eat a lean cuisine :(she's yelling all day that I'm "anorexic". Stfu, leave me alone I'm FAT."

Diagnostic criteria

- "@xxxxxx even at my sickest, I didn't meet the weight criteria for anorexia diagnosis. But I had severe muscle wastage at that point"
- "@xxxxxxx what a sweetheart you are. I was quoting the medically accepted diagnostic criteria for anorexia nervosa... thx for the information."

Ambivalence

- "@xxxxxx @xxxxx Hi I also have an eating disorder, and I often find it comforting. Please don't speak on behalf of every single person with an ED. We are all very different, and you never know what helps. Let's try building each other up."
- "Worst feeling ever is missing my thin anorexia body even though logically I know it was unhealthy and killing me. And now I have to struggle everyday to live in this "fat" body. Just completely unbearable at times."
- "My eating disorder is my worst enemy, yet my closest friend."

Textbox 4. The selected topics and the representative tweets in group 3.

Negative emotions or attitudes

- "There was a certain weight threshold I never wanted to go above. And now I'm above that. I feel like a complete utter failure."
- "Honestly I'm disgusting. How can I even contemplate eating when I'm this morbidly obese. Jesus."
- "Sorry for being bitchy guys. I'm really tired, I need a shower, I'm worried, and I'm hungry. I'm just grumpy."

Eating disorder symptoms

- "Haven't eaten in 25hours, will eat in 30minutes. Salad & Chicken."
- "I haven't properly fasted for such a long time, only restricted or binged. My fast began 3 hours ago and I'm actually excited! #fasting"
- "@xxxxx I'm liquid fasting for at least 24hrs now."

Self-harm

- "My wrist n cuts are getting dry, they disappear so fast"
- "I have so many cuts at the moment. I went a bit mental last week. Everything's just falling apart."
- "All because I'm so emotional, I'm gonna be flaunting some major self harming. The county fair is next week... Everyone will see..."

Negative consequences

- "I feel so empty, I just want to cut my wrist and bleed out #depressed #depressedgirl #nomoreburns #numb"
- "crying. life and mind are falling apart again."
- "The self hatred is so mentally and physically draining."
- "I feel so mentally unstable at the moment. Like I'm on the verge of a complete mental breakdown any minute."

Communication

- "Someone talk to me please. I'm just going mad please!"
- "Someone please talk me out of the Popsicle I want so bad. #HelpMe"
- "I could use someone to talk to. By talk, I mean text lol feelin kinda lonely all alone. :(#miasisters"

Suicide

- "I'm getting recurrent suicidal thoughts. My mental health is just awful at the moment. I feel very self destructive and unstable. I hate it."
- "I want to die so bad but having to commit suicide will change my family's life and other peoples outlook on them"
- "I'm so depressed, I feel like the only escape is suicide. Life is so pointless, you live in mental hell, you live with physical pain, you end up alone hen death."

Sleep deprivation

- "It's 4:40am and I can't fall asleep. I'm yawning cause I'm tired, but I can't find sleep. I have too many things going through my mind."
- "Laying in bed and my stomach won't stop churning, think I'm gonna be having a restless night tonight!"
- "In the past two days I've drank 23 cups of coffee and actually hallucinated. I'm not sure if its from sleep deprivation or caffeine #Whoops"

Laxative usage

- "Uh-oh! Can start to feel the lax cramps now....."
- "I can start to feel my laxatives kicking in. I took 5 so far. Just hate the cramps when I take laxatives. :-/"
- "I've been using a lot of laxatives lately. It doesn't seem to do much except give me cramps."

Discussion

Identification of ED-Relevant Tweets and Topic Modeling

In earlier studies, the identification of ED-relevant tweets was based on a manual search method using ED keywords or hashtags and filtered by manually curated rules [19,21]. This approach was not efficient and may have included numerous irrelevant tweets owing to the ambiguity of the keywords. In this study, we developed deep learning classifiers to automatically identify relevant tweets posted by *ED-laypeople*. We simplified the classification task into a 2-step process and achieved good performance with 2 supervised deep learning classifiers (ie, CNN for step 1 and LSTM for step 2) and

obtained reasonable results with F_1 scores of 0.89 and 0.90, respectively. For content analysis of the identified tweets, we applied topic modeling, an unsupervised method, and manual review, which could summarize comprehensive information from tens of thousands of tweets. This approach is more efficient compared with the completely manual review, which could only cover several hundreds of tweets [21].

Analysis of Topic Modeling Results

The overall coherence rate for the 3 experimental groups was 77.07% (1264/1640), which indicates high quality of the topics produced by the CorEx model. The categories and higher-level themes of the produced topics were summarized by the domain expert (AH). Some of these categories may seem unclear or disparate if not considered within the right context. For example, the *Insults* category referred to using ED symptoms or terms to insult someone (eg, using the word *anorexic* as a derogatory term). Therefore, this content was ED-related, although not pertaining to a traditional ED topic (eg, ED symptoms). This principle was similar for the other categories. For instance, the category of *Questions* referred to a range of ED content (eg, "@instagram This is not a difficult question - how do I escalate a complaint about dangerous, self-harm-encouraging, pro ana content that your algorithm has deemed safe?"), although the organizing principle for this group of Tweets was question-asking (ie, all tweets were posed in the form of a question). We believe that the diversity of topics identified by this algorithm is a strength of this investigation. Producing a range of topics will allow researchers to better understand the social media content on EDs, including content that would not have been expected. This information could ultimately inform future mechanistic and treatment research on EDs. For instance, we would not have anticipated that ED terminology would be used as insults on social media. We hypothesize that this type of language could further stigmatize this set of disorders, which present a target for future prevention efforts. In addition, the *Questions* category identified issues important to potential ED users on social media, such as the need for better social media monitoring and blocking of content that could be detrimental to individuals with, or vulnerable to developing, an ED. This further highlights the importance of generating algorithms for identifying ED-related content on social media. Improvement of these methods could be used by social media platforms such as Twitter to improve filtering practices for harmful content and/or to provide appropriate mental health resources to individuals posting or viewing such content.

The first and second groups focused on the *ED-laypeople* tweets posted by manually identified potential ED users. One aim of comparing these 2 groups was to verify that the *ED-laypeople* tweets identified by manually specified rules and a machine learning classifier could produce similar topics. The 2 groups were found to have 14 overlapping topic categories, such as *Negative consequences*, *ED symptoms*, *Education or awareness or treatment*, and *ED recovery*. This result was in agreement with our previous study [17]. There were slight discrepancies in the identified categories between the 2 groups. *Body image* uniquely appeared among the highest frequency topics in group 1, whereas *Media or advertising or portrayals* uniquely appeared among the highest frequency topics in group 2. The higher prevalence of the *Media or advertising or portrayals* topic category in group 2 reveals a pitfall of the developed machine learning classifier; some *ED-promotional and education* tweets were misclassified as *ED-laypeople* tweets by the classifier, leading to a larger *Media or advertising or portrayals* topic category. When we manually identified the *ED-laypeople* tweets, we could check both the content of the tweets and the user profile of the accounts that posted the tweets. However, the developed classifier only used the content from the tweets themselves and did not incorporate profile information. Including user information such as usernames as features in the classifier would have the potential to mitigate this misclassification problem. Furthermore, group 1 had one unique topic category, *Ambivalence*, occupying 2% (1/54) of total meaningful topics in this group, whereas group 2 had unique categories of *Negative social reactions*,*Insults*,*Mental illness*, and *Social media*, occupying 10% (6/63) of the total meaningful topics. In general, the 2 groups of tweets produced highly similar topics, indicating that our machine learning classifier was mostly as effective as the manual method. This is meaningful because it suggests that this method could be used to identify ED-relevant social media content for future larger scale investigations. In addition, such methods could ultimately aid in identifying at-risk groups for whom prevention efforts could be targeted, which is especially important given the low level of ED detection in typical practice [2].

According to Textbox 4, 6 new topics were identified compared with our previous study [17]. These topics are consistent with earlier literature on EDs and may provide novel insights into the thoughts and experiences of individuals with EDs. Several areas of the content reflected topics that may be relevant to understanding the decisional mechanisms involved in these disorders. The *Reflection or planning* category (eg, "A few months ago I was hiding in my dorm room severely depressed and relapsing from anorexia; tomorrow morning marks one month of me taking recovery seriously; is also my move-in day at a new school.") and *Ambivalence* category (eg, "My eating disorder is my worst enemy, yet my closest friend.") reflect the strong pull of individuals with EDs toward and against ED symptoms [26]. This suggests that for some, the decision about whether to engage in ED behavior involves consideration of the pros and cons of engaging in these behaviors and planning future actions [26]. In future research, it would be informative to determine whether individuals producing tweets reflecting the deliberative processes of weighing pros and cons of ED behaviors vary in a clinically relevant manner from those not producing these tweets, as there is a suggestion that deliberative processes might characterize earlier stages of illness [18]. In addition, the *Comorbidity* and *Diagnostic criteria* categories highlight the importance of considering heterogeneity across a range of severity and symptom profiles in ED research and treatment [27]. Many ED researchers have highlighted the importance of considering certain ED characteristics (eg, whether a person is emotionally dysregulated vs constrained) in planning treatments [28]. In addition, it has long been acknowledged that *Diagnostic criteria* categories fail to capture many individuals who do not present with classic ED symptoms (eg, meet all criteria for anorexia nervosa but are not underweight) [29]. These topics also might characterize content

from clinically unique subgroups of individuals with EDs, warranting further consideration.

The summarized topics for the group 3 tweets are listed in Textbox 3. The high-frequency topics (n>5) identified from the ED-irrelevant tweets included *Negative emotions or attitudes* (n=8), *Eating disorder behaviors* (n=7), and *Body image* (n=5). The content identified in these analyses reveals what other topics are common to potential ED users in their daily lives. One notable feature of these results is that the content of many of the tweets identified as *ED-irrelevant* pertains explicitly to ED cognitive and behavior symptoms (eg, *ED symptoms* and *Body image* categories). One interpretation is that for individuals with ED, the experience with their disorder becomes so pervasive that it infiltrates content that is not explicitly intended as ED-relevant. However, this could also indicate that the algorithm needs to be further refined to capture all the content relevant to EDs. An additional finding from this analysis is that many of the tweets demonstrate that negative emotion and pain predominate the experience of having an ED, as reflected in the categories *Negative emotions or attitude*, *Negative consequences*, *Self-harm*, and *Suicide*. These categories correspond with theories that ED symptoms may result from a surfeit of negative emotions and that the symptoms themselves may function to alleviate emotional pain in a similar fashion to self-harm and suicide planning [30-32].

Limitations

We developed machine learning classifiers that could identify ED-relevant tweets with high performance, but there is still a small percentage of misclassified tweets, especially in the task of differentiating the *ED-promotional and education* versus *ED-laypeople* tweets. This may be partially due to the short length of some tweets, which makes them difficult to classify. Furthermore, these 2 types of tweets are sometimes semantically similar. With the increase in the number of collected tweets, there will be a large number of misclassified tweets, although the misclassification rate is low, which will further influence the results of topic modeling as it will produce some noise topics. Another limitation is that we cannot collect all the tweets in the entire timeline of our target potential ED users owing to the restriction of Twitter API, which may lead to other useful topics being missed.

Conclusions

Our study developed a 2-step process using 2 classifiers (ie, CNN and LSTM) that could automatically identify ED-relevant tweets posted by the potential ED users. The F_1 scores of the 2 classifiers were 0.89 and 0.90, respectively. A CorEx model was applied on the tweets identified by the classifiers and those identified by a traditional manual method separately. Highly overlapping topics were produced. Through a review of these topics by a domain expert, important features of the social media content of potential ED were identified. These findings provided novel insights into the experience of having an ED, which could be expanded upon in future research using the methods derived in this investigation.

Acknowledgments

This study was supported by the National Center for Complementary and Integrative Health of the National Institutes of Health (NIH) under Award Number R01AT009457 (principal investigator [PI]: RZ), the National Institute of Mental Health under Award Number K23 MH112867 (PI: AH), and the National Science Foundation (NSF) under Award Number 1734134 (PI: JB). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the NSF.

Conflicts of Interest

None declared.

References

1. Hudson JI, Hiripi E, Pope HG, Kessler RC. The prevalence and correlates of eating disorders in the National Comorbidity Survey Replication. Biol Psychiatry 2007 Feb 01;61(3):348-358 [FREE Full text] [doi: 10.1016/j.biopsych.2006.03.040] [Medline: 16815322]
2. Kutz AM, Marsh AG, Gunderson CG, Maguen S, Masheb RM. Eating Disorder Screening: a Systematic Review and Meta-analysis of Diagnostic Test Characteristics of the SCOFF. J Gen Intern Med 2020 Mar;35(3):885-893. [doi: 10.1007/s11606-019-05478-6] [Medline: 31705473]
3. Eddy KT, Tabri N, Thomas JJ, Murray HB, Keshaviah A, Hastings E, et al. Recovery From Anorexia Nervosa and Bulimia Nervosa at 22-Year Follow-Up. J Clin Psychiatry 2017 Feb;78(2):184-189. [doi: 10.4088/JCP.15m10393] [Medline: 28002660]
4. Steinhausen H. The outcome of anorexia nervosa in the 20th century. Am J Psychiatry 2002 Aug;159(8):1284-1293. [doi: 10.1176/appi.ajp.159.8.1284] [Medline: 12153817]
5. Bauer S, Kindermann SS, Moessner M. [Prevention of eating disorder: a review]. Z Kinder Jugendpsychiatr Psychother 2017 Sep;45(5):403-413. [doi: 10.1024/1422-4917/a000506] [Medline: 27951744]
6. Berkman ND, Bulik CM, Brownley KA, Lohr KN, Sedway JA, Rooks A, et al. Management of eating disorders. Evid Rep Technol Assess (Full Rep) 2006 Apr(135):1-166. [Medline: 17628126]

7. Kass AE, Kolko RP, Wilfley DE. Psychological treatments for eating disorders. Curr Opin Psychiatry 2013 Nov;26(6):549-555 [FREE Full text] [doi: 10.1097/YCO.0b013e328365a30e] [Medline: 24060917]
8. Paul MJ, Dredze M. You are what you tweet: Analyzing Twitter for public health. 2011 Presented at: Fifth International AAAI Conference on Weblogs and Social Media; 17-21, July, 2011; Barcelona, Spain p. A.
9. Xu S, Markson C, Costello KL, Xing CY, Demissie K, Llanos AA. Leveraging Social Media to Promote Public Health Knowledge: Example of Cancer Awareness via Twitter. JMIR Public Health Surveill 2016;2(1):e17 [FREE Full text] [doi: 10.2196/publichealth.5205] [Medline: 27227152]
10. Musaev A, Britt RK, Hayes J, Britt BC, Maddox J, Sheinidashtegol P. Study of Twitter communications on cardiovascular disease by state health departments. 2019 Presented at: InInternational Conference on Web Services (pp.). Springer, Cham; 2019 Jun 25; Milan, Italy p. 181-189. [doi: 10.1007/978-3-030-23499-7_12]
11. Bian J, Topaloglu U, Yu F. Towards large-scale twitter mining for drug-related adverse events. 2012 Oct Presented at: InProceedings of the 2012 international workshop on Smart health and wellbeing (pp.); 2012 Oct 29; Maui, USA p. 25-32. [doi: 10.1145/2389707.2389713]
12. Zhang H, Wheldon C, Dunn AG, Tao C, Huo J, Zhang R, et al. Mining Twitter to assess the determinants of health behavior toward human papillomavirus vaccination in the United States. J Am Med Inform Assoc 2020 Feb 01;27(2):225-235. [doi: 10.1093/jamia/ocz191] [Medline: 31711186]
13. Modave F, Zhao Y, Krieger J, He Z, Guo Y, Huo J, et al. Understanding Perceptions and Attitudes in Breast Cancer Discussions on Twitter. Stud Health Technol Inform 2019 Aug 21;264:1293-1297 [FREE Full text] [doi: 10.3233/SHTI190435] [Medline: 31438134]
14. Zhao Y, Guo Y, He X, Wu Y, Yang X, Prosperi M, et al. Assessing mental health signals among sexual and gender minorities using Twitter data. Health Informatics J 2020 Jun;26(2):765-786. [doi: 10.1177/1460458219839621] [Medline: 30969146]
15. Wang Y, Zhao Y, Zhang J, Bian J, Zhang R. Detecting associations between dietary supplement intake and sentiments within mental disorder tweets. Health Informatics J 2020 Jun;26(2):803-815. [doi: 10.1177/1460458219867231] [Medline: 31566452]
16. Hicks A, Hogan WR, Rutherford M, Malin B, Xie M, Fellbaum C, et al. Mining Twitter as a First Step toward Assessing the Adequacy of Gender Identification Terms on Intake Forms. AMIA Annu Symp Proc 2015;2015:611-620 [FREE Full text] [Medline: 26958196]
17. Zhou S, Bian J, Zhao Y, Haynos AF, Rizvi R, Zhang R. Analysis of Twitter to Identify Topics Related to Eating Disorder Symptoms. IEEE Int Conf Healthc Inform 2019 Jun;2019 [FREE Full text] [doi: 10.1109/ichi.2019.8904863] [Medline: 32030368]
18. Walsh BT. The enigmatic persistence of anorexia nervosa. Am J Psychiatry 2013 May;170(5):477-484 [FREE Full text] [doi: 10.1176/appi.ajp.2012.12081074] [Medline: 23429750]
19. Arseniev-Koehler A, Lee H, McCormick T, Moreno MA. #Proana: Pro-Eating Disorder Socialization on Twitter. J Adolesc Health 2016 Jun;58(6):659-664. [doi: 10.1016/j.jadohealth.2016.02.012] [Medline: 27080731]
20. Kenny TE, Boyle SL, Lewis SP. #recovery: Understanding recovery from the lens of recovery-focused blogs posted by individuals with lived experience. Int J Eat Disord 2020 Aug;53(8):1234-1243. [doi: 10.1002/eat.23221] [Medline: 31886573]
21. Cavazos-Rehg PA, Krauss MJ, Costello SJ, Kaiser N, Cahn ES, Fitzsimmons-Craft EE, et al. "I just want to be skinny": A content analysis of tweets expressing eating disorder symptoms. PLoS One 2019;14(1):e0207506 [FREE Full text] [doi: 10.1371/journal.pone.0207506] [Medline: 30650072]
22. Alghamdi R, Alfalqi K. A Survey of Topic Modeling in Text Mining. Int. J. Adv. Comput. Sci. Appl 2015;6(1). [doi: 10.14569/IJACSA.2015.060121]
23. Barde BV, Bainwad AM. An overview of topic modeling methods and tools. 2017 Jun Presented at: International Conference on Intelligent Computing and Control Systems (ICICCS); 2017 Jun 15; Madurai, India p. 745-750. [doi: 10.1109/iccons.2017.8250563]
24. Yan X, Guo J, Lan Y, Cheng X. A biterm topic model for short texts. 2013 May Presented at: Proceedings of the 22nd international conference on World Wide Web; 2013 May 13; Rio de Janeiro, Brazil p. 1445-1456. [doi: 10.1145/2488388.2488514]
25. Gallagher RJ, Reing K, Kale D, Ver Steeg G. Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge. Transactions of the Association for Computational Linguistics 2017 Dec;5:529-542. [doi: 10.1162/tacl_a_00078]
26. Serpell L, Treasure J, Teasdale J, Sullivan V. Anorexia nervosa: friend or foe? Int J Eat Disord 1999 Mar;25(2):177-186. [doi: 10.1002/(sici)1098-108x(199903)25:2<177::aid-eat7>3.0.co;2-d] [Medline: 10065395]
27. Wildes JE, Marcus MD. Alternative methods of classifying eating disorders: models incorporating comorbid psychopathology and associated features. Clin Psychol Rev 2013 Apr;33(3):383-394 [FREE Full text] [doi: 10.1016/j.cpr.2013.01.006] [Medline: 23416343]
28. Wildes JE, Marcus MD, Crosby RD, Ringham RM, Dapelo MM, Gaskill JA, et al. The clinical utility of personality subtypes in patients with anorexia nervosa. J Consult Clin Psychol 2011 Oct;79(5):665-674 [FREE Full text] [doi: 10.1037/a0024597] [Medline: 21767000]

29. Dunn EC, Geller J, Brown KE, Bates ME. Addressing the EDNOS issue and improving upon the utility of DSM-IV: classifying eating disorders using symptom profiles. Eur Eat Disord Rev 2010;18(4):271-280. [doi: 10.1002/erv.1005] [Medline: 20552559]
30. Haynos AF, Fruzzetti AE. Anorexia nervosa as a disorder of emotion dysregulation: evidence and treatment implications. Clin Psychol Sci Prac 2011;18(3):183-202. [doi: 10.1111/j.1468-2850.2011.01250.x]
31. Pisetsky EM, Haynos AF, Lavender JM, Crow SJ, Peterson CB. Associations between emotion regulation difficulties, eating disorder symptoms, non-suicidal self-injury, and suicide attempts in a heterogeneous eating disorder sample. Compr Psychiatry 2017 Feb;73:143-150 [FREE Full text] [doi: 10.1016/j.comppsych.2016.11.012] [Medline: 27978502]
32. Wang SB, Borders A. The unique effects of angry and depressive rumination on eating-disorder psychopathology and the mediating role of impulsivity. Eat Behav 2018 Apr;29:41-47. [doi: 10.1016/j.eatbeh.2018.02.004] [Medline: 29477016]

Abbreviations

API: application programming interface
BTM: Biterm Topic Model
CNN: convolutional neural network
CorEx: Correlation Explanation
ED: eating disorder
GB: gradient boosting trees
LDA: latent Dirichlet allocation
LSTM: long short-term memory
NB: naïve Bayes
NIH: National Institutes of Health
NSF: National Science Foundation
PI: principal investigator
RF: random forest
SVM: support vector machine

Edited by C Lovis; submitted 15.02.20; peer-reviewed by L Cui, R Rodgers, K Reuter; comments to author 28.04.20; revised version received 14.07.20; accepted 06.09.20; published 30.10.20.

Please cite as:
Zhou S, Zhao Y, Bian J, Haynos AF, Zhang R
Exploring Eating Disorder Topics on Twitter: Machine Learning Approach
JMIR Med Inform 2020;8(10):e18273
URL: http://medinform.jmir.org/2020/10/e18273/
doi:10.2196/18273
PMID:33124997

Original Paper

A Computer-Interpretable Guideline for COVID-19: Rapid Development and Dissemination

Shan Nan[1,2], PhD; Tianhua Tang[1], BSc; Hongshuo Feng[1], BSc; Yijie Wang[3], BSc; Mengyang Li[1], BSc; Xudong Lu[1,2], PhD; Huilong Duan[1], PhD

[1]College of Biomedical Engineering and Instrumental Science, Zhejiang University, Hangzhou, China

[2]Information Systems Industrial Engineering & Innovation Sciences, Technical University of Eindhoven, Eindhoven, Netherlands

[3]Hangzhou Vico Software Cooperation, Hangzhou, China

Corresponding Author:
Xudong Lu, PhD
College of Biomedical Engineering and Instrumental Science
Zhejiang University
Zhouyiqing Building, 512
38 Zheda Road, Hangzhou
Hangzhou, 310027
China
Phone: 86 13957118891
Email: lvxd@zju.edu.cn

Abstract

Background: COVID-19 is a global pandemic that is affecting more than 200 countries worldwide. Efficient diagnosis and treatment are crucial to combat the disease. Computer-interpretable guidelines (CIGs) can aid the broad global adoption of evidence-based diagnosis and treatment knowledge. However, currently, no internationally shareable CIG exists.

Objective: The aim of this study was to establish a rapid CIG development and dissemination approach and apply it to develop a shareable CIG for COVID-19.

Methods: A 6-step rapid CIG development and dissemination approach was designed and applied. Processes, roles, and deliverable artifacts were specified in this approach to eliminate ambiguities during development of the CIG. The Guideline Definition Language (GDL) was used to capture the clinical rules. A CIG for COVID-19 was developed by translating, interpreting, annotating, extracting, and formalizing the Chinese COVID-19 diagnosis and treatment guideline. A prototype application was implemented to validate the CIG.

Results: We used 27 archetypes for the COVID-19 guideline. We developed 18 GDL rules to cover the diagnosis and treatment suggestion algorithms in the narrative guideline. The CIG was further translated to object data model and Drools rules to facilitate its use by people who do not employ the non-openEHR archetype. The prototype application validated the correctness of the CIG with a public data set. Both the GDL rules and Drools rules have been disseminated on GitHub.

Conclusions: Our rapid CIG development and dissemination approach accelerated the pace of COVID-19 CIG development. A validated COVID-19 CIG is now available to the public.

(JMIR Med Inform 2020;8(10):e21628) doi:10.2196/21628

KEYWORDS

COVID-19; guideline; CDSS; openEHR; Guideline Definition Language; development; dissemination; electronic health record; algorithm

Introduction

COVID-19 is a global pandemic that is affecting over 200 countries and territories worldwide [1]. As of June 2020, 8,690,140 cases of COVID-19 have been diagnosed, and 461,274 deaths from the disease have been reported [2]. Medical resources, especially intensive care resources, have been drained by the COVID-19 pandemic in both developed and developing countries [3,4]. Proper prevention, efficient diagnosis, and effective treatment based on established evidence are crucial to save patients, ease the burden of medical workers, and accelerate eradication of the disease [5]. Unfortunately, the perception and

knowledge of COVID-19 diagnosis and treatment among caregivers are still at low levels, which significantly hinders the pace of managing the disease [6].

Information technology is crucial for combating the COVID-19 pandemic [7,8]. Many efforts have already been contributed to estimate the trend of the pandemic at national or global levels [9-12], predict the prognosis of an individual patient [13,14], visualize and track reported cases of COVID-19 in real time [15], assist diagnosis based on chest computed tomography images [16], provide telemedicine for chronic disease patients [17], improve caregivers' work efficiency [18-20], and survey the public attitude and response towards COVID-19 [21,22]. Some electronic medical record (EMR) system vendors have pushed out updates to their software to help caregivers detect potential patients with COVID-19 [23]. However, to our best knowledge, efforts in this line supporting evidence-based COVID-19 diagnosis and treatment are limited. Particularly, a publicly available computer-interpretable guideline (CIG) for COVID-19 has not been reported. Such a CIG could accelerate the wide and rapid adoption of evidence-based diagnosis and treatment guidelines.

The lack of a CIG is unsurprising if one considers the enormous challenges of developing a shareable CIG for COVID-19 in a limited timeframe. Sharing CIGs among different organizations involves many difficulties. The use of a site-specific data model (known as the "curly braces problem") in a CIG limits it to a specific clinical site [24]. Due to different output formats, it is challenging to integrate a CIG into various EMRs.

Developing a CIG is a time-consuming and resource-dependent process that requires a group of informaticists and medical specialists to work together closely for a considerable period of time because they must engage in a significant number of discussions to eliminate ambiguity and misunderstanding of the narrative guideline during development [25,26]. In the conventional approach, CIG development is broken down into several phases, and the input and output artifacts in each phase are defined at a conceptual level. However, a clear specification of these artifacts has not yet been established, and the roles who should take part in each phase are unclear. Moreover, it is not possible to apply the conventional CIG development approach to COVID-19 because frequent face-to-face discussions are impractical during the pandemic. Few local medical specialists are available to take part in CIG development because these resources are currently scarce [3]. Even informaticists can no longer meet face-to-face in many countries because of local lockdown policies.

An approach that can standardize the input and output of CIGs and leverage existing resources would be helpful. The openEHR standard is a potential solution. From the technology point of view, openEHR aims to facilitate interoperability between information systems [27]. The openEHR archetype provides a standard information model that can be shared among organizations to avoid the "curly braces problem" [24]. From the domain knowledge perspective, openEHR aims to bring informaticists and medical specialists together. Specifically, openEHR uses an archetype to capture detailed and domain-specific clinical concepts that are modeled by clinical specialists [28]. The Guideline Definition Language (GDL) was proposed by the openEHR community to facilitate the use of openEHR archetypes to author CIGs [29]. Recently, GDL was upgraded to its second major version, known as GDL2 [29]. GDL improves the shareability of encoded CIGs among organizations across borders [30,31]. However, there is still a gap between interpreting a narrative guideline and using GDL to author a CIG, especially considering the current difficulties of efficient communication. A specification for informaticists to use GDL to rapidly capture narrative guideline knowledge is urgently required.

This paper proposes a rapid CIG development and dissemination approach using GDL. A sharable CIG enabling automatic diagnosis and treatment of COVID-19 has been developed and disseminated by applying the proposed approach. A prototype application has been developed and validated with public patient data to demonstrate the use of the CIG.

Methods

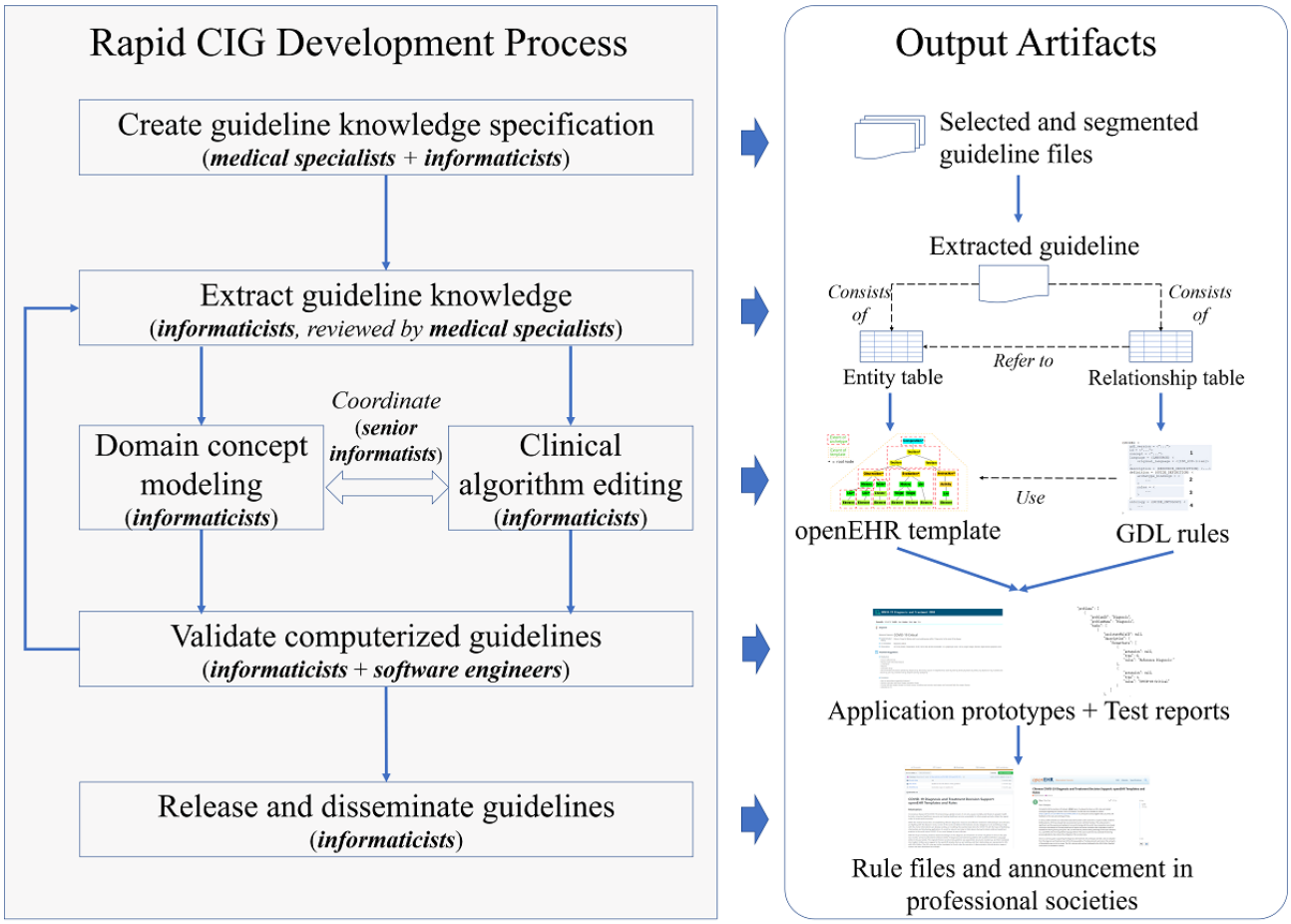
Referring to the CIG authoring approach proposed by Zhou et al [26], we designed a rapid CIG development approach that parallelizes data modeling and rule editing. In addition to the original approach, the output of each step was specified to eliminate ambiguities between different participants. Then, we reported the detailed process of applying the approach to develop a CIG for COVID-19 based on the seventh edition of the Chinese COVID-19 Diagnosis and Treatment Plan.

Design of the Rapid CIG Development Approach

Zhou et al [26] reported a CIG authoring approach with six steps, including (1) create a knowledge specification, (2) integrate with terminology, (3) author rules, (4) test rules, (5) publish rules, and (6) generate reports. These six steps should be carried out sequentially, and key artifacts are produced in each step. Two aspects of this approach should be optimized to support the rapid development of CIGs. First, among these six steps, the first three steps are time-consuming and dependent on medical resources. Second, although key artifacts are categorized in the approach, their contents are not specified, which may still cause ambiguities.

In this section, we propose a rapid approach to develop and disseminate CIGs by solving these two problems. The key steps, participants, and output artifacts of each step are specified in our approach. The rapid CIG development and dissemination approach contains six steps (see Figure 1): (1) create guideline knowledge specifications, (2) extract guideline knowledge, (3) model the domain concept, (4) edit the clinical algorithm, (5) validate the computerized guideline, and (6) release and disseminate the guideline. The approach is explained in detail as follows.

Figure 1. The scheme of the rapid CIG development approach. CIG: computer-interpretable guideline.



Step 1: Create Guideline Knowledge Specifications

In this step, a joint CIG development team consisting of both medical specialists and informaticists must be established. Medical specialists provide clinical requirements for decision support. According to the requirements, the informaticists select related guidelines, read through each narrative guideline, segment interesting sections in the guidelines, and finally confirm their selections with medical specialists. Selected and segmented human-readable guideline files are organized as the output of this step.

Step 2: Extract Guideline Knowledge

This step bridges the human-readable narrative guideline and the CIG by breaking narrative text into entities and relations. Informaticists read through the narrative guideline and break the guideline into small logic blocks, which can be represented by production rules. Each block contains a left-hand side representing the conditions and a right-hand side representing the consequent actions. These relationships are collected as a relationship table and delivered as an artifact. Both the left-hand side and right-hand side are further broken down from phrases into individual terms. Entities are marked up and extracted from those terms to form an entity table, which is also an artifact. The extraction results must be reviewed by medical specialists to ensure their correctness.

Steps 3 and 4: Model the Domain Concept Modeling and Edit the Clinical Algorithm

These steps are performed concurrently and in collaboration by two groups of informaticists to accelerate the development pace. This process should be coordinated by senior informaticists to ensure consistency in both groups. Domain concept modeling refers to the openEHR template development process. Based on the aforementioned entity table, informaticists search the openEHR archetype repository and select suitable archetypes that best represent the entities. An archetype should be created if there is no appropriate archetype for a specific entity. An openEHR template is developed to organize these archetypes. The detailed approach of openEHR template modeling is described elsewhere [32]. In the clinical algorithm editing step, another group of informaticists translates the relationship table into GDL rules. The translation is straightforward because GDL follows the general structure of production rules. When editing the clinical algorithms, the domain concept model is needed as the input. At the same time, the clinical algorithm editing raises domain concept requirements that must be created or refined. This bidirectional dependence must be coordinated by senior informaticists.

Step 5: Validate the Computerized Guideline

In this step, the CIG is validated with clinical data. Guideline authoring tools usually contain a CIG validation module, which receives data from manual input by informaticists. The CIG can be further validated by implementing clinical decision support

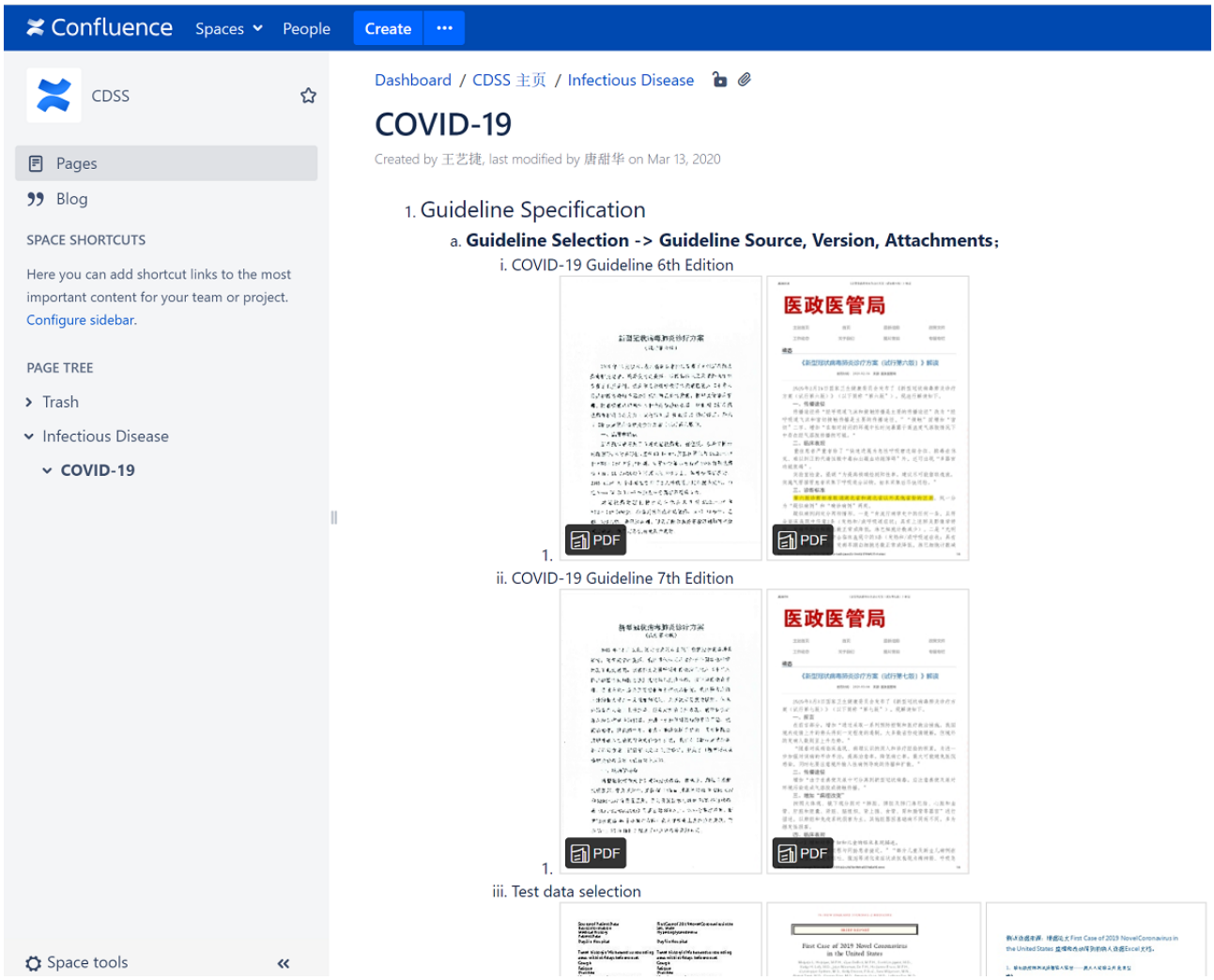
systems (CDSSs). This work requires the involvement of software engineers. If the validation results have any inconsistency with the original narrative guideline, the whole process from guideline extraction to authoring is reviewed.

Step 6: Release and Disseminate the Guideline

Finally, the well-developed guideline should be made publicly available by informaticists. Code-sharing platforms such as GitHub and forums of professional societies are good options for disseminating these guidelines.

To manage the proposed 6-step approach, especially considering online collaboration, we used the Atlassian Confluence team collaboration platform [33] as a tool to coordinate the work (see Figure 2). The six steps were elicited on the platform. Each participant was requested to submit their outcome artifacts on the platform. Version controls were implemented to track the history of documents. Archetypes were searched in the openEHR Clinical Knowledge Manager (CKM) [34]. The GDL rules were edited using the GDL2 Editor in docker [35].

Figure 2. Screenshot of the Confluence knowledge platform configured for the proposed approach.



COVID-19 CIG Development

In this section, we report the detailed process of rapid development of a CIG for COVID-19 based on the seventh edition of the COVID-19 Diagnosis and Treatment Plan published by the National Health Commission of the People's Republic of China (NHC) [36]. An English translation was used as a reference [37].

Create Guideline Knowledge Specifications

The Chinese COVID-19 guideline was published by the NHC to share the latest evidence-based best practices regarding COVID-19 diagnosis and treatment and to nationally standardize caregiver practices.

The seventh edition of the Chinese COVID-19 Diagnosis and Treatment Plan consists of 13 sections. Sections 1 to 4 provide general introductions to the pathogen, epidemiology, pathology, and patient features of SARS-CoV-2. Then, in sections 5 to 8, the diagnostic criteria are discussed in detail. In Section 9, guidelines are defined as to how cases should be reported. In Section 10, treatment plans are introduced. From Section 11 to Section 13, discharge criteria, transportation, and in-hospital infection are briefly discussed. The CIG development team spent three days reading and discussing the guideline. Because our primary goal was to develop an executable guideline, the CIG development team jointly decided to select the diagnosis and treatment sections (ie, sections 5 to 8 and sections 10 and 11). The traditional Chinese medicine section was not included in the treatment suggestions because it is not available in other

countries. Both the original copy published by the NHC and the English version translated by a public society were used in the next steps.

Extract Guideline Knowledge

One informaticist (TT) read the guideline and split the sentences into a spreadsheet (Multimedia Appendix 1). Four days were spent on this task. A color system was used in the splitting: blue indicated that the sentence was related to diagnosis or treatment; black indicated that the sentence was not suitable for translation to a clinical rule; and red indicated that the sentence contained vague content and required later consultation with medical specialists. The document was uploaded to the Confluence platform in a timely fashion. Each update was double-checked by two other informaticists (SN and HF).

Then, TT annotated the entities in the text and extracted them into another spreadsheet (Multimedia Appendix 2). Repeated entities were merged. TT uploaded the document to Confluence, and two other informaticists (HF and ML) double-checked the annotation and the extracted results. The final extracted results were reviewed and confirmed by an external medical specialist. Two additional days were spent on this task. In total, this step took six days.

Model the Domain Concept

Based on the entity table developed in the previous step, ML and HF mapped the listed entities to the openEHR concepts. While mapping, the openEHR template was expanded accordingly. The mapping was checked by a senior informaticist (SN) and an external openEHR expert. A group of archetypes in the format of archetype definition language (ADL) files were exported from the CKM. The detailed archetype searching and template development process is reported elsewhere [32]. Ten days were required for the domain concept modeling because the selection of proper archetypes required confirmation by the external openEHR expert. Three rounds of teleconferences were held to finalize the domain concept model.

Author the Clinical Algorithm

Sentences in the narrative guideline were broken down into the left-hand side and right-hand side blocks by TT in the guideline extraction step. In this step, HF used the GDL2 Editor to encode the clinical algorithms in GDL.

ADL files describing the COVID-19 data requirements were imported into the GDL2 Editor. Following the structure of the extracted guideline file, HF translated each left-hand side and right-hand side pair to a GDL rule in a *when-then* format.

The GDL rules were checked and confirmed by SN. SN and HF took part in both the domain concept modeling and clinical algorithm authoring. They bridged the two groups of informaticists and lowered the communication cost. The clinical algorithm authoring was performed at the same time as the domain concept modeling. Once a part of the domain concept model was finalized, the related GDL rules were created accordingly. The entire authoring process was synchronized with the domain concept modeling step.

Validate the Computerized Guideline

The CIG developed in the previous step was validated both by GDL2 Editor and a prototype CDSS for COVID-19. Our research team previously developed a configurable CDSS platform named Tracebook to develop CDSS applications rapidly [38]. In this study, we used the Tracebook platform to configure a fast prototype of a COVID-19 CDSS. Because there are no open-source or openly available GDL2 execution engines, we chose the Drools rule engine to execute the clinical rules [39]. Mapping was required between GDL2 and Drools at both the data model level and the language level. The mapping rule and mapping specification were defined jointly by SN, TT, and HF (Multimedia Appendix 2). Then, the mapping was performed manually by TT. An additional Drools rule for the user interface presentation was also developed. The additional rule mapping and system development were performed over five days.

Test patient data were adopted from patient case report published in a medical journal [40]. The patient was a 35-year-old man who had cough and fever symptoms and had recently traveled to Wuhan, China. The patient's demographic information, history, and observations were captured from the publication and entered into both the GDL2 Editor guideline validation module and our own CDSS. The output was compared with both the guideline and the reported diagnosis and prescription. Inconsistencies between these three outputs were reported to external medical specialists, and the CIG was reviewed.

Release and Disseminate the Guideline

Archetypes in ADL file format and the GDL rules in GDL2 format were exported from their editors, packaged together, and committed to GitHub [41]. Java data models and Drools rules were also committed to GitHub to benefit people who do not use openEHR. Then, the dissemination was reported on the openEHR disclosure forum [42].

Results

In this section, we illustrate our GDL COVID-19 guideline model and the validation results.

The Computerized COVID-19 Guideline

Domain Concept Model

The domain concept model is illustrated in detail in Table 1.

A total of 27 archetypes were used for the COVID-19 CIG, among which 26 were directly acquired from the CKM and 1 (openEHR-EHR-CLUSTER.imaging_result-COVID_19.v0) was acquired from the CKM and modified for the COVID-19 CDSS. These 27 archetypes were sorted into 9 categories: demographic, history, medical record, exam, vital sign, laboratory test, symptom, diagnosis, and order. The organization of the Java data models followed the concept categories.

Table 1. List of concept categories, used openEHR archetypes, and their associated models.

Concept category	openEHR archetypes	Object data model
Demographic	openEHR-EHR-OBSERVATION.age.v0	PatientInfo
History	openEHR-EHR-OBSERVATION.exposure_assessment.v0	EpidemicHistory
Medical record	openEHR-EHR-OBSERVATION.pf_ratio.v0 openEHR-EHR-OBSERVATION.story.v1	MedicalRecord
Exam	openEHR-EHR-CLUSTER.imaging_finding.v0 openEHR-EHR-CLUSTER.imaging_result-COVID_19.v0 openEHR-EHR-OBSERVATION.imaging_exam_result.v0	ImgExamResult
Vital sign	openEHR-EHR-CLUSTER.inspired_oxygen.v1 openEHR-EHR-CLUSTER.level_of_exertion.v0 openEHR-EHR-CLUSTER.problem_qualifier.v1 openEHR-EHR-OBSERVATION.body_temperature.v2 openEHR-EHR-OBSERVATION.pulse_oximetry.v1 openEHR-EHR-OBSERVATION.respiration.v2	PhysicalSign
Laboratory test	openEHR-EHR-CLUSTER.specimen.v0 openEHR-EHR-CLUSTER.laboratory_test_analyte.v1 openEHR-EHR-OBSERVATION.laboratory_test_result.v1	LabTestResult
Symptom	openEHR-EHR-CLUSTER.symptom_sign.v1 openEHR-EHR-COMPOSITION.encounter.v1 openEHR-EHR-OBSERVATION.symptom_sign_screening.v0 openEHR-EHR-OBSERVATION.condition_screening.v0	Symptom
Diagnosis	openEHR-EHR-EVALUATION.differential_diagnoses.v0 openEHR-EHR-EVALUATION.health_risk.v1 openEHR-EHR-EVALUATION.problem_diagnosis.v1	Diagnosis
Order	openEHR-EHR-EVALUATION.recommendation.v1 openEHR-EHR-INSTRUCTION.medication_order.v2 openEHR-EHR-INSTRUCTION.therapeutic_order.v0 openEHR-EHR-OBSERVATION.management_screening.v0	Order

Algorithm Model

The COVID-19 diagnosis and treatment rules are listed in Table 2.

Sections 5 to 8 and sections 10 and 11 of the Chinese COVID-19 Diagnosis and Treatment Plan were encoded in both GDL and Drools. These rules support diagnosis, classification, early warning, treatment, and discharge for caregivers.

For sections 5, 10, and 11, there are multiple GDL rules for one section. This is because GDL2 Editor now only allows one rule in a file, whereas these sections contain several rules. This limitation does not exist in Drools; therefore, we merged the rules for one purpose into one Drools rule file.

Table 2. List of created GDL and Drools rules for the associated sections of the Chinese COVID-19 Diagnosis and Treatment Plan.

Section	GDL[a] rules	Drools rule
5. Diagnostic Criteria	COVID_Confirmed_Diagnosis.v0.gdl2 COVID_Lymphocyte_count.v0.gdl2 COVID_Nucleic_acid_test_result.v0.gdl2 COVID_White_blood_cell_count.v0.gdl2 COVID_White_cell_count.v0.gdl2	Diagnosis_Confirmed
6. Clinical Classification	COVID_Classfication.v0.gdl2	Classification
7. Clinical Warning Sign	COVID_Clinical_Warning.v0.gdl2	Clinical_Warning
8. Differential Diagnosis	COVID_Suspected_Diagnosis.v0.gdl2	Diagnosis_Suspected
10. Treatment	COVID_Blood_Purification_Treatment.v0.gdl2 COVID_Circulation_support_Treatment.v0.gdl2 COVID_Continuous_Renal_Replacement_Therapy.v0.gdl2 COVID_Convalescent_plasma_Treatment.v0.gdl2 COVID_General_Treatment.v0.gdl2 COVID_Immunotherapy.v0.gdl2 COVID_Other_Treatment.v0.gdl2 COVID_Respiratory_support_Treatment.v0.gdl2	Treatment_Modern
11. Discharge	COVID_Body_Temperature_Monitor.v0.gdl2 COVID_Out_Hospital.v0.gdl2	Discharge

[a]GDL: Guideline Definition Language.

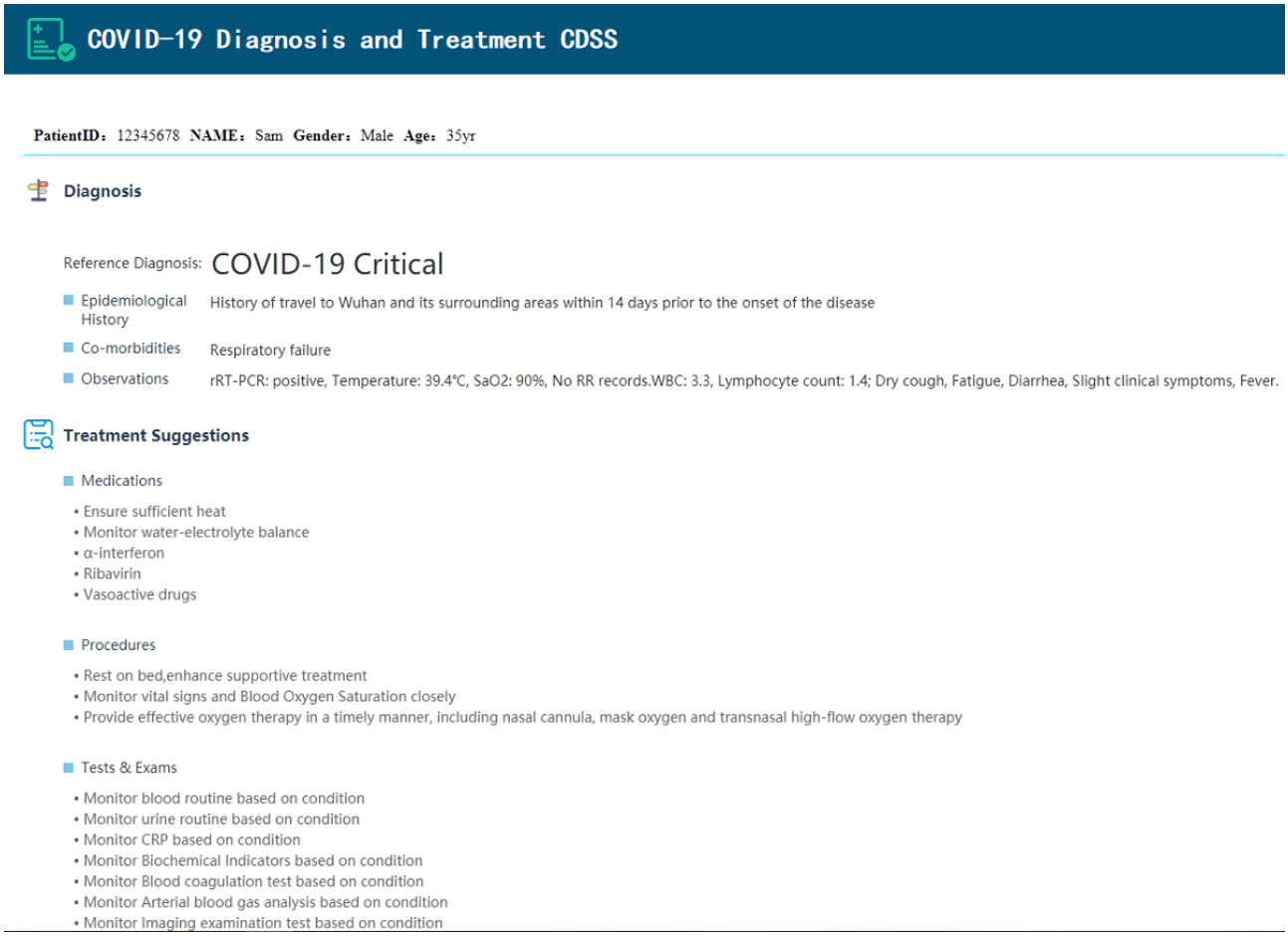
Validation of the Guideline

A prototype COVID-19 diagnosis and treatment CDSS was configured using the Tracebook platform [43]. The CDSS receives patient data from a web application program interface (API) and reasons with the COVID-19 CIG to provide evidence-based diagnosis and treatment suggestions. The user interface of the prototype is illustrated in Figure 3.

There are two blocks in the user interface: Diagnosis and Treatment Suggestions. The Diagnosis block is generated by the diagnosis, classification, and early warning rules. In the example in Figure 3, the patient is diagnosed with COVID-19, and the classification is critical. Data supporting the diagnosis are listed below the diagnosis. In the Treatment Suggestions block, medical suggestions, procedure suggestions, tests, and examination suggestions are provided according to the patient's specific situation.

The CIG in both GDL and Drools language was validated by the published patient case report. The patient was diagnosed with critical COVID-19, and a detailed history was available along with vital signs, symptoms, image examinations, laboratory tests, and medication prescriptions. The diagnosis and treatment suggestions fit both the diagnosis and treatment plan in the case report and the Chinese COVID-19 Diagnosis and Treatment Plan. A detailed validation test report, including input and output data, is provided in Multimedia Appendix 3.

Figure 3. User interface of the prototype clinical decision support system application implementing the COVID-19 computer-interpretable guideline.



Discussion

Principal Results

We described a rapid development and dissemination approach to establish CIGs and applied this approach to a COVID-19 CIG. The COVID-19 pandemic is a global crisis that requires worldwide contributions from every domain [8]. While medical researchers have been establishing efficient diagnostic measures and effective treatment methodologies, informaticists are obligated to accelerate the wide adoption of these valuable best practices [7]. The usual approach of CIG development does not specify the input and output of each key step; therefore, informaticists and clinical specialists must engage in intensive discussions to understand each other. In our approach, we hastened this process by formalizing and structuring the discussions and reducing ambiguity.

The rapid CIG development approach makes the maximum use of existing medical knowledge sources (ie, openEHR archetypes) and parallels the tasks of domain concept modeling and clinical algorithm editing to further accelerate the process. During this study, four informaticists, two software engineers, and one external medical specialist were able to interpret and model the COVID-19 guideline and develop a prototype system remotely in four weeks. The rapid development and dissemination approach worked well for developing the COVID-19 CIG.

Goud et al [25] proposed a parallel guideline development and formalization strategy that encourages guideline development teams and the CIG development team to work closely together. By applying this strategy, the quality of the guideline and the efficiency of development of the CIG can both be improved. However, we argue that this strategy is not applicable in the COVID-19 crisis because the guideline was developed by a temporal national committee that we are not able to work with. In fact, in most current cases, clinical guidelines are still published by authorized committees or societies without the participation of informaticists. Van Gorp et al [44] proposed a model-driven engineering approach to rapidly translate annotated guideline knowledge to decision support applications. However, the procedure of annotation was not specified in their study.

The openEHR approach also possesses several potential advantages for future implementation. The openEHR approach provides a standard information model (ie, an archetype) that enables sharing of data definitions among organizations so that the “curly braces problem” is avoided. Moreover, the output of GDL2 rules is built based on archetypes; therefore, it is theoretically sharable among organizations. Indeed, it is estimated that 58 healthcare providers in 14 countries are currently using openEHR solutions [45]. The openEHR-based techniques can be translated into standardized HL7 Fast Healthcare Interoperability Resources (FHIR) format, which can be adopted by more EHRs [46].

Limitations

The Guideline Elements Model (GEM) Cutter is a tool for annotating guidelines [47,48]. Due to the urgent development requirement, we did not use the GEM Cutter for the guideline annotation and extraction. A combination of GEM and GDL will be used in our future work.

Another limitation of our study is that the efficiency was not measured and compared with that of the usual approach. Because our primary goal was to rapidly develop and share a COVID-19 CIG, a comparison with other approaches was not performed. Moreover, for practical reasons, it is difficult to measure the exact time spent by each participant on each step. While developing this CIG, the researchers were locked down at home and working remotely. It is difficult to count the exact hours spent by each person because some of the researchers were using their spare time to perform this work, and working from home unavoidably scattered their working time.

For our prototype application, we did not use GDL as the execution language. There are three reasons for this. First, to the best of our knowledge, no dedicated execution engine for GDL2 is currently publicly available on the internet. Second, we did not manage to represent a time serial in GDL2 (eg, the last three nucleic acid tests were all negative); therefore, additional rules for data preprocessing in other languages were required. Last, there is a gap between the output of GDL2 rules and the actual requirements of the application. Thus, we used an open-source rule engine, Drools, instead. The GDL rules were manually translated to Drools rules with a set of predefined mapping rules. However, we believe that non-openEHR users can benefit from the object data model and Drools rules.

The rapid development and dissemination approach for CIGs has only been tested in the COVID-19 case. Although it worked well for our case, more tests are needed to determine its genericity. The COVID-19 guideline has been validated with a published patient case report. However, the clinical rules have not yet been applied to daily practice. When implementing these rules, it is likely that additional fine-tuning will be required to fit the local medical cultures and workflows of different health care providers.

Conclusions

A CIG for COVID-19 can help caregivers provide evidence-based diagnosis and treatment to patients with COVID-19 to improve the quality of care. As yet, no such CIG exists due to the difficulty of rapid development. In this paper, we proposed a rapid development and dissemination approach for CIGs and developed a COVID-19 guideline by applying this approach. We hope that the COVID-19 CIG that we developed can help clinical information system vendors and care providers build their own CDSSs for COVID-19. Further, we hope that our approach can help other informaticists rapidly develop their own CIGs and share them globally in the future.

Acknowledgments

This study was funded by the Chinese National Science and Technology Major Project (grant number 2016YFC0901703). The authors would like to thank Dr Heather Leslie for reviewing and commenting on the extracted guideline and the openEHR template, Mr Bin Qi for technical support while searching for related archetypes and developing the COVID-19 openEHR template, and Mr Kuai Yu for implementing the decision support system prototype.

Authors' Contributions

SN designed the approach, conducted the study, reviewed the clinical rules, designed the prototype, and drafted the manuscript. TT translated and structured the guideline. HF mapped the data elements to the openEHR template and encoded the structured guideline in GDL. YW codesigned the approach and supervised the rule development and software implementation. ML built the openEHR template and reviewed the extracted data elements. XL supervised the study and contributed to major revisions of this manuscript. HD supervised the study and coordinated the resources. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Segmented and structured Chinese COVID-19 guideline.
[XLS File (Microsoft Excel File), 138 KB - medinform_v8i10e21628_app1.xls]

Multimedia Appendix 2
Extracted data items and mapping between archetypes and object data models.
[XLS File (Microsoft Excel File), 138 KB - medinform_v8i10e21628_app2.xls]

Multimedia Appendix 3
Computer-interpretable guideline validation test report.
[DOC File , 190 KB - medinform_v8i10e21628_app3.doc]

References

1. Coronavirus disease 2019 (COVID-19) Situation Report 152. World Health Organization. 2020 Jun 20. URL: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200620-covid-19-sitrep-152.pdf?sfvrsn=83aff8ee_4 [accessed 2020-06-21]
2. WHO Coronavirus Disease (COVID-19) Dashboard. World Health Organization. URL: https://covid19.who.int/ [accessed 2020-06-21]
3. Ji Y, Ma Z, Peppelenbosch MP, Pan Q. Potential association between COVID-19 mortality and health-care resource availability. Lancet Glob Health 2020 Apr;8(4):e480 [FREE Full text] [doi: 10.1016/S2214-109X(20)30068-1] [Medline: 32109372]
4. Mitchell-Box K, Braun KL. Fathers' thoughts on breastfeeding and implications for a theory-based intervention. J Obstet Gynecol Neonatal Nurs 2012;41(6):E41-E50. [doi: 10.1111/j.1552-6909.2012.01399.x] [Medline: 22861175]
5. Coronavirus. World Health Organization. URL: https://www.who.int/health-topics/coronavirus [accessed 2020-06-20]
6. Bhagavathula AS, Aldhaleei WA, Rahmani J, Mahabadi MA, Bandari DK. Knowledge and Perceptions of COVID-19 Among Health Care Workers: Cross-Sectional Study. JMIR Public Health Surveill 2020 Apr 30;6(2):e19160 [FREE Full text] [doi: 10.2196/19160] [Medline: 32320381]
7. Bakken S. Informatics is a critical strategy in combating the COVID-19 pandemic. J Am Med Inform Assoc 2020 Jun 01;27(6):843-844 [FREE Full text] [doi: 10.1093/jamia/ocaa101] [Medline: 32501484]
8. Adams JG, Walls RM. Supporting the Health Care Workforce During the COVID-19 Global Epidemic. JAMA 2020 Apr 21;323(15):1439-1440. [doi: 10.1001/jama.2020.3972] [Medline: 32163102]
9. Fang Y, Nie Y, Penny M. Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis. J Med Virol 2020 Jun;92(6):645-659 [FREE Full text] [doi: 10.1002/jmv.25750] [Medline: 32141624]
10. Tosi D, Verde A, Verde M. Clarification of Misleading Perceptions of COVID-19 Fatality and Testing Rates in Italy: Data Analysis. J Med Internet Res 2020 Jun 17;22(6):e19825 [FREE Full text] [doi: 10.2196/19825] [Medline: 32490842]
11. Huang Q, Kang Y. Mathematical Modeling of COVID-19 Control and Prevention Based on Immigration Population Data in China: Model Development and Validation. JMIR Public Health Surveill 2020 May 25;6(2):e18638 [FREE Full text] [doi: 10.2196/18638] [Medline: 32396132]
12. Shen C, Chen A, Luo C, Zhang J, Feng B, Liao W. Using Reports of Symptoms and Diagnoses on Social Media to Predict COVID-19 Case Counts in Mainland China: Observational Infoveillance Study. J Med Internet Res 2020 May 28;22(5):e19421 [FREE Full text] [doi: 10.2196/19421] [Medline: 32452804]
13. Rajgor DD, Lee MH, Archuleta S, Bagdasarian N, Quek SC. The many estimates of the COVID-19 case fatality rate. Lancet Infect Dis 2020 Jul;20(7):776-777 [FREE Full text] [doi: 10.1016/S1473-3099(20)30244-9] [Medline: 32224313]
14. Obeid JS, Davis M, Turner M, Meystre SM, Heider PM, O'Bryan EC, et al. An artificial intelligence approach to COVID-19 infection risk assessment in virtual visits: A case report. J Am Med Inform Assoc 2020 Aug 01;27(8):1321-1325 [FREE Full text] [doi: 10.1093/jamia/ocaa105] [Medline: 32449766]
15. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020 May;20(5):533-534 [FREE Full text] [doi: 10.1016/S1473-3099(20)30120-1] [Medline: 32087114]
16. Pan F, Ye T, Sun P, Gui S, Liang B, Li L, et al. Time Course of Lung Changes at Chest CT during Recovery from Coronavirus Disease 2019 (COVID-19). Radiology 2020 Jun;295(3):715-721 [FREE Full text] [doi: 10.1148/radiol.2020200370] [Medline: 32053470]
17. Wosik J, Fudim M, Cameron B, Gellad ZF, Cho A, Phinney D, et al. Telehealth transformation: COVID-19 and the rise of virtual care. J Am Med Inform Assoc 2020 Jun 01;27(6):957-962 [FREE Full text] [doi: 10.1093/jamia/ocaa067] [Medline: 32311034]
18. Reeves JJ, Hollandsworth HM, Torriani FJ, Taplitz R, Abeles S, Tai-Seale M, et al. Rapid response to COVID-19: health informatics support for outbreak management in an academic health system. J Am Med Inform Assoc 2020 Jun 01;27(6):853-859 [FREE Full text] [doi: 10.1093/jamia/ocaa037] [Medline: 32208481]
19. Judson TJ, Odisho AY, Neinstein AB, Chao J, Williams A, Miller C, et al. Rapid design and implementation of an integrated patient self-triage and self-scheduling tool for COVID-19. J Am Med Inform Assoc 2020 Jun 01;27(6):860-866 [FREE Full text] [doi: 10.1093/jamia/ocaa051] [Medline: 32267928]
20. Yan A, Zou Y, Mirchandani DA. How hospitals in mainland China responded to the outbreak of COVID-19 using information technology-enabled services: An analysis of hospital news webpages. J Am Med Inform Assoc 2020 Jul 01;27(7):991-999 [FREE Full text] [doi: 10.1093/jamia/ocaa064] [Medline: 32311036]
21. Wang P, Lu W, Ko N, Chen Y, Li D, Chang Y, et al. COVID-19-Related Information Sources and the Relationship With Confidence in People Coping with COVID-19: Facebook Survey Study in Taiwan. J Med Internet Res 2020 Jun 05;22(6):e20021 [FREE Full text] [doi: 10.2196/20021] [Medline: 32490839]
22. Zhao Y, Cheng S, Yu X, Xu H. Chinese Public's Attention to the COVID-19 Epidemic on Social Media: Observational Descriptive Study. J Med Internet Res 2020 May 04;22(5):e18825 [FREE Full text] [doi: 10.2196/18825] [Medline: 32314976]

23. Miliard M. Epic pushes out software update to help spot coronavirus. Healthcare IT News. 2020 Jan 24. URL: https://www.medigy.com/news/2020/01/27/healthcareitnews.com-epic-pushes-out-software-update-to-help-spot-coronavirus/ [accessed 2020-06-21]
24. Wulff A, Haarbrandt B, Tute E, Marschollek M, Beerbaum P, Jack T. An interoperable clinical decision-support system for early detection of SIRS in pediatric intensive care using openEHR. Artif Intell Med 2018 Jul;89:10-23 [FREE Full text] [doi: 10.1016/j.artmed.2018.04.012] [Medline: 29753616]
25. Goud R, Hasman A, Strijbis A, Peek N. A parallel guideline development and formalization strategy to improve the quality of clinical practice guidelines. Int J Med Inform 2009 Aug;78(8):513-520. [doi: 10.1016/j.ijmedinf.2009.02.010] [Medline: 19375977]
26. Zhou L, Karipineni N, Lewis J, Maviglia SM, Fairbanks A, Hongsermeier T, et al. A study of diverse clinical decision support rule authoring environments and requirements for integration. BMC Med Inform Decis Mak 2012 Nov 12;12(1):128 [FREE Full text] [doi: 10.1186/1472-6947-12-128] [Medline: 23145874]
27. What is openEHR? openEHR. URL: https://www.openehr.org/about/what_is_openehr [accessed 2020-07-20]
28. Beale T, Heard S, Kalra D, Lloyd D. OpenEHR architecture overview. OpenEHR. URL: https://specifications.openehr.org/releases/BASE/Release-1.0.3/architecture_overview.html [accessed 2020-09-21]
29. Chen R. Guideline Definition Language v2 (GDL2). openEHR. URL: https://specifications.openehr.org/releases/CDS/latest/GDL2.html [accessed 2020-06-20]
30. Anani N, Chen R, Prazeres Moreira T, Koch S. Retrospective checking of compliance with practice guidelines for acute stroke care: a novel experiment using openEHR's Guideline Definition Language. BMC Med Inform Decis Mak 2014 May 10;14(1):39 [FREE Full text] [doi: 10.1186/1472-6947-14-39] [Medline: 24886468]
31. Kalliamvakos K. Evaluation of the Guideline Definition Language (GDL) in the clinical area of severe sepsis and septic shock. Dissertation. Karolinska Institutet. 2013. URL: https://ki.se/sites/default/files/migrate/evaluation_konstantinos_kalliamvakos.pdf [accessed 2020-09-21]
https://ki.se/sites/default/files/migrate/evaluation_konstantinos_kalliamvakos.pdf
32. Li M, Leslie H, Qi B, Nan S, Feng H, Cai H, et al. Development of an openEHR Template for COVID-19 Based on Clinical Guidelines. J Med Internet Res 2020 Jun 10;22(6):e20239 [FREE Full text] [doi: 10.2196/20239] [Medline: 32496207]
33. Confluence. Atlassian. URL: https://www.atlassian.com/software/confluence [accessed 2020-06-20]
34. GDL2 Editor. dockerhub. URL: https://hub.docker.com/r/cdsplatform/gdl2-editor [accessed 2020-06-20]
35. Docker Desktop for Windows. dockerhub. URL: https://store.docker.com/editions/community/docker-ce-desktop-windows [accessed 2020-06-20]
36. Novel coronavirus pneumonia diagnosis and treatment plan (provisional 7th edition). National Health Commission of the People's Republic of China. 2020 Mar 04. URL: http://www.nhc.gov.cn/yzygj/s7653p/202003/46c9294a7dfe4cef80dc7f5912eb1989.shtml [accessed 2020-06-20]
37. Novel coronavirus pneumonia diagnosis and treatment plan (provisional 7th edition). China Law Translate. 2020 Mar 04. URL: https://www.chinalawtranslate.com/en/coronavirus-treatment-plan-7/ [accessed 2020-06-20]
38. Nan S, Lu X, Van Gorp P, Korsten HHM, Vdovjak R, Kaymak U, et al. Design and implementation of a platform for configuring clinical dynamic safety checklist applications. Frontiers Inf Technol Electronic Eng 2018 Sep 14;19(7):937-946. [doi: 10.1631/fitee.1700623]
39. Drools documentation. Drools. URL: https://www.drools.org/learn/documentation.html [accessed 2020-06-20]
40. Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, Washington State 2019-nCoV Case Investigation Team. First Case of 2019 Novel Coronavirus in the United States. N Engl J Med 2020 Mar 05;382(10):929-936 [FREE Full text] [doi: 10.1056/NEJMoa2001191] [Medline: 32004427]
41. openEHR-COVID-19. GitHub. URL: https://github.com/ZJU-BME-VICO/openEHR-COVID-19 [accessed 2020-06-20]
42. Nan S. Chinese COVID-19 Diagnosis and Treatment Decision Support: openEHR Templates and Rules. openEHR. 2020 Mar 23. URL: https://discourse.openehr.org/t/chinese-covid-19-diagnosis-and-treatment-decision-support-openehr-templates-and-rules/516 [accessed 2020-06-20]
43. COVID-19 Diagnosis and Treatment CDSS. Cooperation VS. URL: http://mcp-cdss.vico-lab.com/lst/#/ [accessed 2020-06-20]
44. Gorp PV, Vanderfeesten I, Dalinghaus W, Sanden BVD, Kubben P, Van GP, et al. Towards generic MDE support for extracting purpose-specific healthcare models from annotated, unstructured texts. In: Weber J, Perseil I, editors. Foundations of Health Information Engineering and Systems. FHIES 2012. Lecture Notes in Computer Science, vol 7789. Berlin, Germany: Springer; 2013:213-221.
45. OpenEHR deployed solutions Internet. openEHR. URL: https://www.openehr.org/openehr_in_use/deployed_solutions/ [accessed 2020-06-22]
46. Fette G, Ertl M, Störk S. Translating openEHR Models to FHIR. Stud Health Technol Inform 2020 Jun;270:1415-1416. [doi: 10.3233/shti200469]
47. Koch K, Woodcock M, Harris M. AMIA Annu Symp Proc 2010 Nov 13;2010:397-401 [FREE Full text] [Medline: 21347008]
48. Hajizadeh N, Kashyap N, Michel G, Shiffman RN. GEM at 10: a decade's experience with the Guideline Elements Model. AMIA Annu Symp Proc 2011;2011:520-528 [FREE Full text] [Medline: 22195106]

Abbreviations

ADL: archetype definition language
API: application program interface
CDSS: clinical decision support system
CIG: computer-interpretable guideline
CKM: Clinical Knowledge Manager
EMR: electronic medical record
FHIR: Fast Healthcare Interoperability Resources
GDL: Guideline Definition Language
GEM: Guideline Elements Model
NHC: National Health Commission of the People's Republic of China

Edited by G Eysenbach; submitted 21.06.20; peer-reviewed by N Deng, E Poon; comments to author 20.07.20; revised version received 15.08.20; accepted 13.09.20; published 01.10.20.

Please cite as:
Nan S, Tang T, Feng H, Wang Y, Li M, Lu X, Duan H
A Computer-Interpretable Guideline for COVID-19: Rapid Development and Dissemination
JMIR Med Inform 2020;8(10):e21628
URL: https://medinform.jmir.org/2020/10/e21628
doi:10.2196/21628
PMID:32931443

Original Paper

Enabling External Inquiries to an Existing Patient Registry by Using the Open Source Registry System for Rare Diseases: Demonstration of the System Using the European Society for Immunodeficiencies Registry

Raphael Scheible[1,2], MSc; Dennis Kadioglu[3], MSc; Stephan Ehl[2], Prof Dr; Marco Blum[1], BSc; Martin Boeker[1], Prof Dr; Michael Folz[3], Dipl-Inf; Bodo Grimbacher[2,4,5,6], Prof Dr; Jens Göbel[3], BSc; Christoph Klein[7], Prof Dr Dr; Alexandra Nieters[2,8], PD Dr; Stephan Rusch[2,8]; Gerhard Kindle[2,8*], Dr, Dipl-Inf; Holger Storf[3*], Dr

[1]Institute of Medical Biometry and Statistics, Medical Center, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[2]Institute for Immunodeficiency, Center for Chronic Immunodeficiency, Medical Center, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[3]Medical Informatics Group, University Hospital Frankfurt, Frankfurt am Main, Germany

[4]German Center for Infection Research, Satellite Center Freiburg, Freiburg, Germany

[5]Centre for Integrative Biological Signalling Studies, University of Freiburg, Freiburg, Germany

[6]RESIST, Cluster of Excellence 2155 to Hanover Medical School, Satellite Center Freiburg, Freiburg, Germany

[7]Department of Pediatrics, Dr von Hauner Children's Hospital, University Hospital, Ludwig Maximilians Universität München, München, Germany

[8]FREEZE Biobank, Center for Biobanking, Medical Center, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[*]these authors contributed equally

Corresponding Author:
Raphael Scheible, MSc
Institute of Medical Biometry and Statistics
Medical Center, Faculty of Medicine
University of Freiburg
Stefan-Meier-Str. 26
Freiburg, 79104
Germany
Phone: 49 761 203 69272
Email: raphael.scheible@uniklinik-freiburg.de

Abstract

Background: The German Network on Primary Immunodeficiency Diseases (PID-NET) utilizes the European Society for Immunodeficiencies (ESID) registry as a platform for collecting data. In the context of PID-NET data, we show how registries based on custom software can be made interoperable for better collaborative access to precollected data. The Open Source Registry System for Rare Diseases (*Open-Source-Registersystem für Seltene Erkrankungen* [OSSE], in German) provides patient organizations, physicians, scientists, and other parties with open source software for the creation of patient registries. In addition, the necessary interoperability between different registries based on the OSSE, as well as existing registries, is supported, which allows those registries to be confederated at both the national and international levels.

Objective: Data from the PID-NET registry should be made available in an interoperable manner without losing data sovereignty by extending the existing custom software of the registry using the OSSE registry framework.

Methods: This paper describes the following: (1) the installation and configuration of the OSSE bridgehead, (2) an approach using a free toolchain to set up the required interfaces to connect a registry with the OSSE bridgehead, and (3) the decentralized search, which allows the formulation of inquiries that are sent to a selected set of registries of interest.

Results: PID-NET uses the established and highly customized ESID registry software. By setting up a so-called OSSE bridgehead, PID-NET data are made interoperable according to a federated approach, and centrally formulated inquiries for data can be received. As the first registry to use the OSSE bridgehead, the authors introduce an approach using a free toolchain to efficiently implement and maintain the required interfaces. Finally, to test and demonstrate the system, two inquiries are realized using the

graphical query builder. By establishing and interconnecting an OSSE bridgehead with the underlying ESID registry, confederated queries for data can be received and, if desired, the inquirer can be contacted to further discuss any requirements for cooperation.

Conclusions: The OSSE offers an infrastructure that provides the possibility of more collaborative and transparent research. The decentralized search functionality includes registries into one search application while still maintaining data sovereignty. The OSSE bridgehead enables any registry software to be integrated into the OSSE network. The proposed toolchain to set up the required interfaces consists of freely available software components that are well documented. The use of the decentralized search is uncomplicated to use and offers a well-structured, yet still improvable, graphical user interface to formulate queries.

(JMIR Med Inform 2020;8(10):e17420) doi:10.2196/17420

KEYWORDS

registry interoperability; collaboration in research; data findability; registry software

Introduction

Background

The German Network on Primary Immunodeficiency Diseases (PID-NET) [1,2] was initiated as a research program of the Pediatric Immunology Working Group (*Arbeitsgemeinschaft Pädiatrische Immunologie* [API], in German) [3], funded by the German Ministry for Education and Research (*Bundesministerium für Bildung und Forschung* [BMBF], in German). The API brings together clinicians and scientists interested in clinical care and clinical research on patients with inborn errors of the immune system. After funding for the BMBF concluded, PID-NET remained a research network of the API, representing a collaborative platform to address various aspects of primary immunodeficiency (PID) research, from clinical care to basic science. For rare diseases, the number of patients on a regional scale is relatively low. However, on a larger scale, for example, countrywide or worldwide, patients with rare diseases are an important group in health care. According to Mahlaoui et al [4], the estimated minimal prevalence of PID in Europe is 11 per 100,000 inhabitants. As these patients and the health care experts specializing in their diseases are spread over several countries, rare diseases need structures for patient support and for disseminating scientific advances that differ from those for frequent diseases. On this basis, Germany published a national plan for rare diseases in 2013, which includes 52 policy proposals to guide and structure actions for treating rare diseases within the German health and social system [5]. In this context, the Open Source Registry System for Rare Diseases (*Open-Source-Registersystem für Seltene Erkrankungen* [OSSE], in German) project [6] was funded by the German Federal Ministry of Health. An important achievement of the OSSE was the generation of freely usable and easily adaptable software for rare disease registries. The software with additional information is publicly available [7]. The open source software can be used by patient organizations, physicians, scientists, and other parties for the creation of patient registries. As a result, the national registry landscape is empowered to comply with European principles regarding the establishment of minimum datasets and compliance with data quality standards; this is summarized in the European Union Committee of Experts on Rare Diseases recommendation on rare disease registries [8]. Also, the necessary interoperability between different registries is supported from the outset and allows those registries to be confederated. For this, the concept of decentralized searches was implemented, which complies with data protection requirements and preserves data sovereignty [9,10]. The OSSE concept focuses on the interoperability of registries and facilitates the process of establishing research networks at various levels (ie, regional and national), which is very attractive in the field of rare diseases. While OSSE-based registries can be interconnected directly, registries based on other software solutions, like the European Society for Immunodeficiencies (ESID) registry, are supported in participating by using the so-called *OSSE bridgehead*. In this work, the ESID registry is used as an example to demonstrate the process of extending a registry that was not initially built using the OSSE framework. By using the OSSE bridgehead, the functionality to receive decentralized search inquiries is added to the registry. The presented work describes the process of how to connect such a registry to the OSSE network and further suggests an approach that uses a free toolchain to implement and maintain the required interfaces.

PID-NET and ESID Registry

Members of the interdisciplinary PID-NET consortium are working together to study inborn disorders of the innate and adaptive immune system. PID-NET focuses, in particular, on severe combined immunodeficiency diseases, autoimmune lymphoproliferative diseases, autoinflammatory diseases, and PID with colitis. One essential part of the PID-NET consortium is the registry in which data for analysis is stored. The registry was founded as part of the PID-NET consortium in 2009 and was funded by the BMBF until March 2018. The aim was to provide a tool to register PID patients for epidemiological and clinical research and to strengthen the network of PID researchers in Germany. Currently, more than 3000 patients from Germany are documented. When PID-NET decided to run a central register in 2009, the decision was made to use the existing ESID platform to document the German cases. The ESID registry uses custom software [11], which was completely re-engineered when the registry underwent a major redesign in 2014 [12].

The OSSE Concept

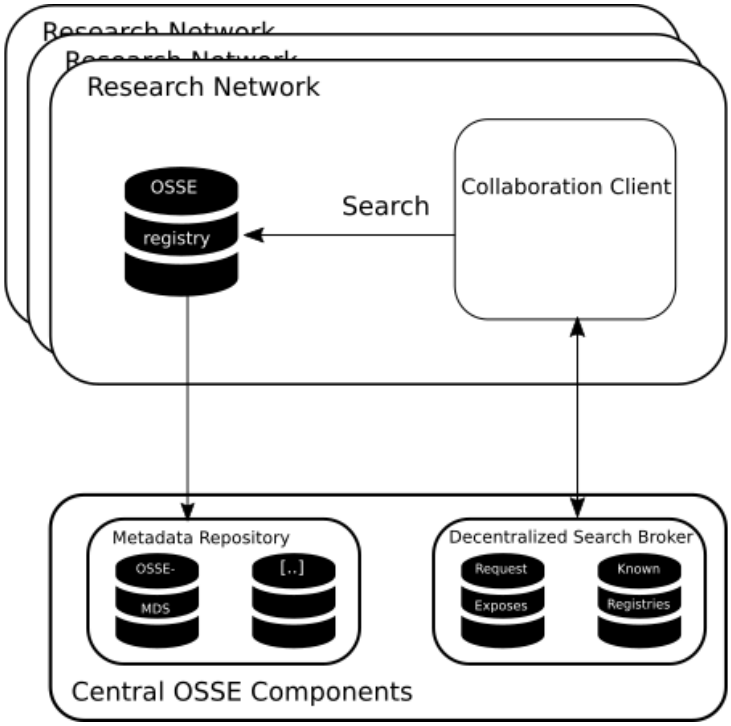
Comprised of specialized modules, OSSE is a registry software toolkit with the goal of enabling scientists with a basic information technology background to build a registry for a specific rare disease. A form editor allows electronic case report forms to be defined for basic and longitudinal medical data. Furthermore, the back end derives the corresponding data

schema from the structure of these forms. Each field or data element contained in these forms has to first be specified by metadata, including the data type, the measurement unit, and the value domain, among others, within a metadata repository (MDR) [13,14].

The integration of an MDR in the OSSE architecture facilitates the later process of integrating data from any registry, since all dataset specifications used in the respective registries can be retrieved from the MDR to ensure that third parties interpret data correctly.

Furthermore, existing data elements can be reused when a new registry is established, and only additionally required items have to be newly defined. At the regional, national, and international levels, common dataset specifications can also be published in the MDR. The MDR component also allows for the retrieval of metadata items from other MDRs. Apart from self-defined value sets, the MDR provides access to standardized classifications, such as the ICD-10-GM (German Modification of the 10th revision of the International Statistical Classification of Diseases and Related Health Problems) or the ICD-O-3 (3rd edition of the International Classification of Diseases for Oncology) [6]. Figure 1 depicts the OSSE concept.

Figure 1. This illustration shows the different components of the OSSE and how they work together. Each registry, based on the OSSE metadata repository (MDS), models its data. The decentralized search broker publishes search inquiries to all selected known registries. Finally, the collaboration client of the selected registries searches for matching data. OSSE: Open Source Registry System for Rare Diseases (*Open-Source-Registersystem für Seltene Erkrankungen*, in German).



The Distributed Search Principle

OSSE follows the principle that raw data, particularly of patients with rare diseases, should not leave the local registry. This is due to low acceptance among patients and data-owning scientists regarding the controllability of data collections and their usage by third parties, particularly outside the influence of national or Europe-wide data protection rules. Instead, OSSE provides a concept for a distributed search, which makes data from rare disease registries available, while respecting data sovereignty and privacy [9,15]. A central search broker allows for the definition of search queries based on the existing data elements in the MDR. Furthermore, the inquiring researcher has to include an abstract in the request describing the research question in detail.

The local request interface of each OSSE registry, called the *collaboration client*, downloads these queries and forwards them to the OSSE bridgehead, which then executes each query based on its local storage. If there is a nonempty result set, it is presented to the person in charge of each matching registry together with the exposé and the inquirer's contact information. Finally, the data owner can review the result set and, if cooperation is desirable, contact the inquirer.

Integration of Registries With the OSSE Bridgehead

Registries based on any non-OSSE software solution can be extended with the OSSE bridgehead to achieve the same interoperability. Hence, existing registries do not need to be converted to OSSE, which reduces barriers to collaboration. The OSSE bridgehead consists of the OSSE core components, including a local data storage component and the collaboration component. Data have to be imported from the registry by a periodical running process that can optionally access a local OSSE identifier management system based on the Mainzelliste [10,16], if the creation or modification of patient pseudonyms is required. To successfully import data into the bridgehead, each registry item has to be added as a data element defined in the MDR. The interface between the bridgehead and the existing registry uses XML following a specific XML Schema Definition

(XSD). Based on metadata from the MDR and a data structure defined in a form editor, the XSD is automatically generated. The existing registry solution either has to implement an XML-based export interface or its native export format has to be transformed in a second step.

Methods

Installation and Configuration of the OSSE Bridgehead

The OSSE bridgehead runs on its own server to increase security by physically distinguishing the registry's database from the local OSSE data storage. The bridgehead software is distributed as docker containers [17], which simplifies the installation and subsequent updates. After the installation, the bridgehead needs to be connected to at least one search broker. This is done by using an email verification process. After registration, one needs to specify the data elements and the data structure. In the MDR, all data elements of the registry's dataset need to be defined by entering a definition, designation, and language. Further, the specification of each element's data type is required. OSSE already offers data types, such as a list of permitted values, several numerical types for which ranges can be set, textual data, Boolean values, and types for date and time. Beyond that, for data elements using terminologies, a type catalog is created. These catalogs contain all the possible values. One can create new catalogs via an XML file upload.

Finally, represented by forms based on the MDR's data elements, the data structure is created. Generally, two types of forms are provided: one for baseline data and one for longitudinal data. In our case, these forms contain data recorded at the first visit and at follow-ups, respectively. These forms are designed with the help of a user interface.

Export Interface

Not every registry software offers a generic XML export, which is required by the OSSE bridgehead. With the help of the powerful, free, open source extract, transform, and load (ETL) tool, Pentaho [18,19], we suggest extending such registries by using an XML export. Further, based on an automatically generated XSD file downloaded from the OSSE bridgehead instance, which describes the dataset predefined in the MDR, an XSD transformation needs to be developed. This transformation translates data from the specialized XML export based on the registry's model to a valid XML file that is compatible with the bridgehead.

After Pentaho is set up to perform the transformation jobs, the result needs to be pushed into the bridgehead, calling its representational state transfer (REST) interface. This ETL procedure can be periodically triggered by a continuous integration process, which is usually integrated into a modern version control system, such as GitLab [20]. This architecture offers flexibility regarding any code change and fast deployment.

Querying OSSE Bridgeheads

The OSSE search broker enables the user to formulate search queries to connected registries. A centrally hosted user interface, for which an account is required, helps the user to create and manage queries. During the process of creating a new query, the submission of a proposal in PDF format is mandatory. Here, the user describes everything that could be relevant for the registry owner's decision in the event of a collaboration. Further, the user has to assign a name to the query and optionally can add a project description. The inquiry will be addressed to a selection of registries of interest. The European Rare Disease Registry Infrastructure (ERDRI) [21] provides an ERDRI Directory of Registries (ERDRI.dor), which offers an overview of participating registries with additional descriptive information. Based on this information, ERDRI.dor offers a search that helps the user to find suitable registries for the inquiry. Finally, after formulating a query and submitting the web form, the search broker publishes the query, which is then downloaded by the collaboration clients of the selected registries of interest. Each one forwards the query to the OSSE bridgehead, which processes it and provides the result. If the result contains one or more matching patients, the registry operator can inspect the result set and the exposé in a user interface. Additionally, the platform allows for reprocessing of the query in order to update the result set based on the most recent data. The registry operator can choose to contact the inquirer to begin to negotiate the requirements for collaboration. The user who sent the query to the registries does not get a direct answer from the system. All communications regarding actual transfer of data have to be done externally (eg, via phone or email), beyond the scope of OSSE. Consequently, data never leave the registry server automatically.

Results

Configuration of the OSSE Bridgehead

As the documented PID-NET dataset has a wide range of parameters, we started the integration of the OSSE bridgehead with a selected subset of data elements.

In general, data elements that are used by different registries and, therefore, already exist as an entity in the MDR should be reused instead of entered redundantly. A search function helps users to find those elements. In case of minor differences regarding data element properties, data elements can be duplicated and subsequently adopted. An example is the existing *gender* data element with the properties *male* and *female*. The ESID registry further offers the option *unknown*. In this case, the already-entered *gender* data element could be used and extended by the missing property. In our case, the classification of PIDs is hierarchically structured in main categories and subcategories. In order to increase usability, we defined a custom catalog that represents the structure and is used to specify the respective data element for the PID classification. For *country of birth*, a catalog of a list of countries predefined by the system was chosen.

In particular, some items in the ESID registry are polymorphic (ie, data elements that may have two kinds of data types). For the item *date of last news*, originally one could assign a given date or a special value. If there is no news, the value *no_news* as type string is stored. OSSE does not support polymorphic data. In the MDR, the option *no_news* was encoded as the date 1850-01-01. For other data elements such as *date of documentation*, the type *date* is simply used. Items like *gene*

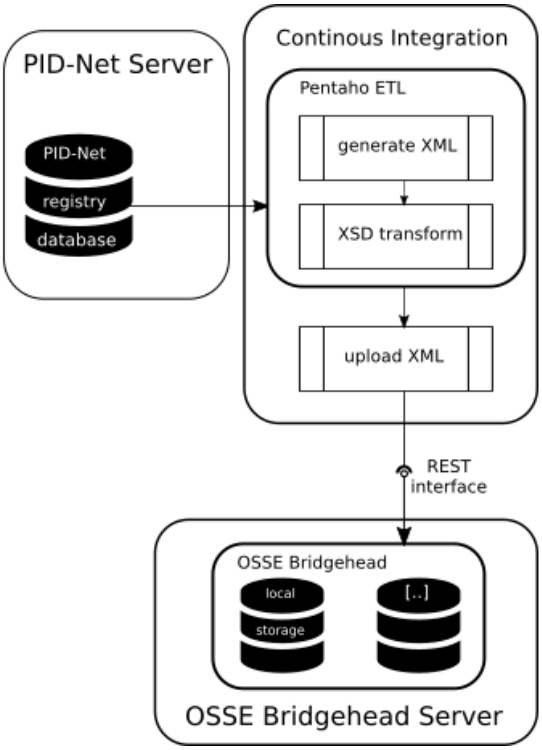
therapy and *current route of administration* are realized with a list of permitted values.

We further distinguished between data recorded at the first visit and data collected during follow-up visits. This is required to achieve a semantically correct definition of our data when defining the two types of forms. In our case, these forms contain data recorded at the first visit and at follow-ups, respectively. These forms are designed with the help of a user interface. Multimedia Appendix 1 shows an excerpt of the PID-NET dataset.

Export Interface

Our registry software does not offer a generic XML export. With the help of an ETL process, we extended the ESID registry by using an XML export functionality. The ETL process further transforms the resulting XML into the bridgehead's required format in order to periodically upload it to the OSSE bridgehead. The ETL process is triggered from within a continuous integration process in the version control software GitLab, which we installed on an internal server. In order to increase security, this server is safely located behind our corporate firewall and is only accessible from specific computers and by specific users. The continuous integration also uploads the export by using a shell script, which communicates with the bridgehead's REST interface. Figure 2 illustrates the design of our implementation.

Figure 2. Connecting PID-NET via an OSSE bridgehead. The same structure could be used for any non-OSSE registry. ETL: extract, transform, and load; OSSE: Open Source Registry System for Rare Diseases (*Open-Source-Registersystem für Seltene Erkrankungen*, in German); PID-NET: German Network on Primary Immunodeficiency Diseases; REST: representational state transfer; XSD: XML Schema Definition.



Querying OSSE Bridgeheads

With the help of a publicly accessible web platform [22], queries to test the system were graphically created. Generally, a search query is constructed by logically connecting parameters and their required values. We modeled and performed the following research questions:

1. Which male patients have ever received hematopoietic stem cell transplantation (HSCT)?
2. Which patients have a common variable immunodeficiency (CVID) as their most recently documented PID diagnosis, or are still receiving immunoglobulin (Ig) replacements at initial registration or during follow-up?

As described earlier, we distinguished between first visit and follow-up data. This enables the user to create more specific queries (eg, formulating queries only considering follow-up data). Such a data design decision has to be made by the registry operator while modeling the data structure in the MDR and form editor.

Any query is performed by concatenating Boolean expressions. Therefore, the user needs to formulate research questions with logical conjunction and disjunction. The part of the second question "at initial registration or during follow-up" implies that the appropriate information refers to the data element at initial registration (init) as well as to the data element at follow-up (fu). Consequently, according to our data model, this requires both data elements to be requested. In order to show the logical expression of the questions, transferable to the visual query builder, we formulated it in an s-expression; this is related to the functional programming language Lisp [23,24] (see Textbox 1). As the second question is more complex, the user needs to nest the formulation of the expressions. Again, the s-expression is shown in Textbox 1.

Textbox 1. Formulations of s-expressions for the first and second questions.

S-expression for the first query:

(AND (EQUALS HSCT “YES”) (EQUALS SEX “M”))

S-expression for the second query:

(AND

(EQUALS PID_{recent}, “CVID”)

(OR

(EQUALS Ig_{fu} “YES”)

(EQUALS Ig_{init} “YES”)

)

)

Figure 3 shows the first expression translated into the graphical query builder. Figure 4 shows the second resulting query in the graphical query builder. After submitting a query, the local bridgehead for every registry of interest receives and locally executes the query. The local collaboration client of each bridgehead installation gives an overview of incoming inquiries and shows the number of matching patients in the bridgehead’s local storage (see Figure 5). Data always stay in the bridgehead’s local storage and, therefore, within the registry’s server infrastructure. Every desired exchange of data requires manual user actions as the system does not automatically generate responses that could reveal information about data. In the case of an incoming query, the data owner decides how to proceed. All subsequent communication regarding possible data transfer to the inquiring party needs to take place outside of the OSSE.

Figure 3. The graphical implementation of the first question, "Which male patients have ever received hematopoietic stem cell transplantation (HSCT)?" MDR: metadata repository.

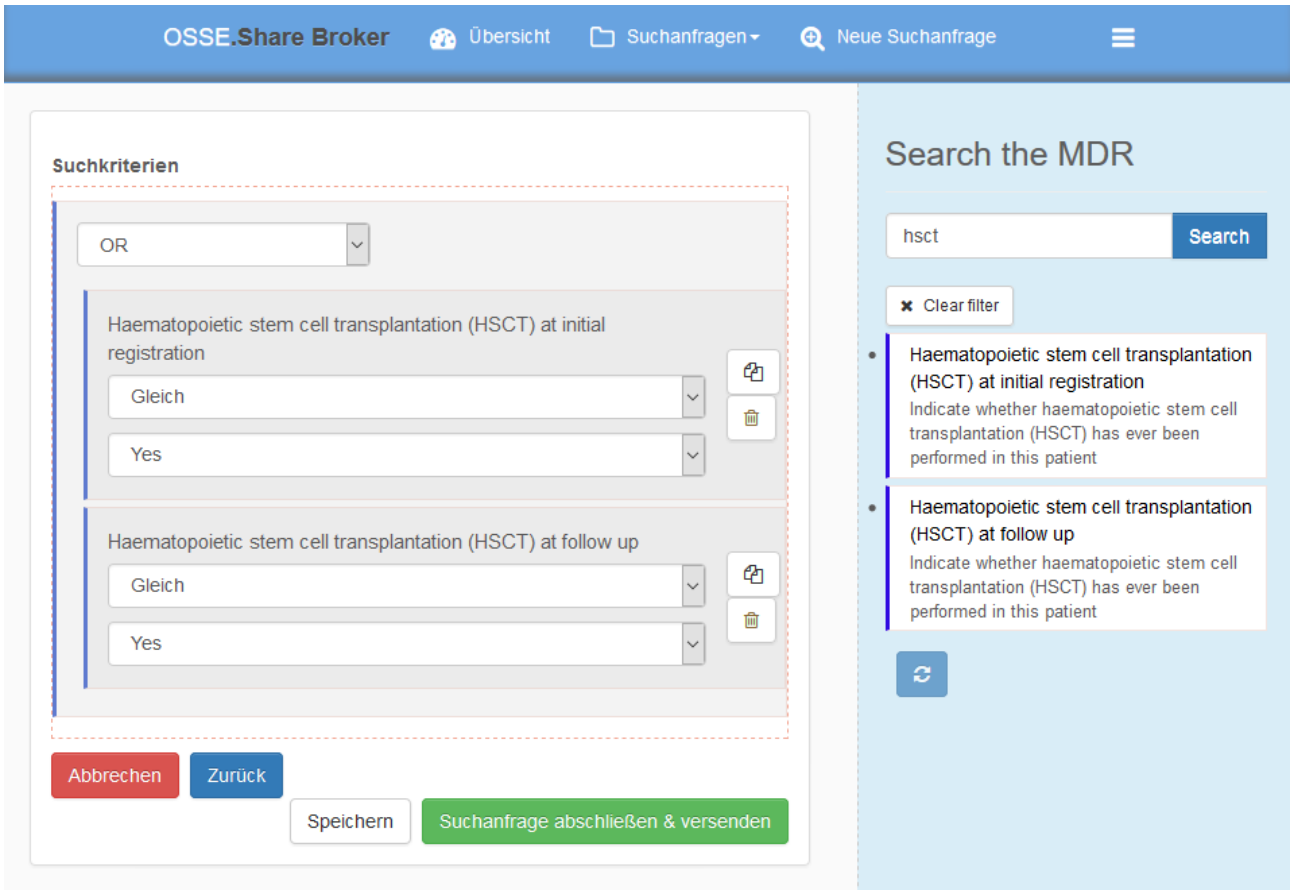


Figure 4. The graphical implementation of the second question requires nested constraints. The second question is "Which patients have a common variable immunodeficiency (CVID) as their most recently documented primary immunodeficiency (PID) diagnosis, or are still receiving immunoglobulin (Ig) replacements at initial registration or during follow-up?" MDR: metadata repository.

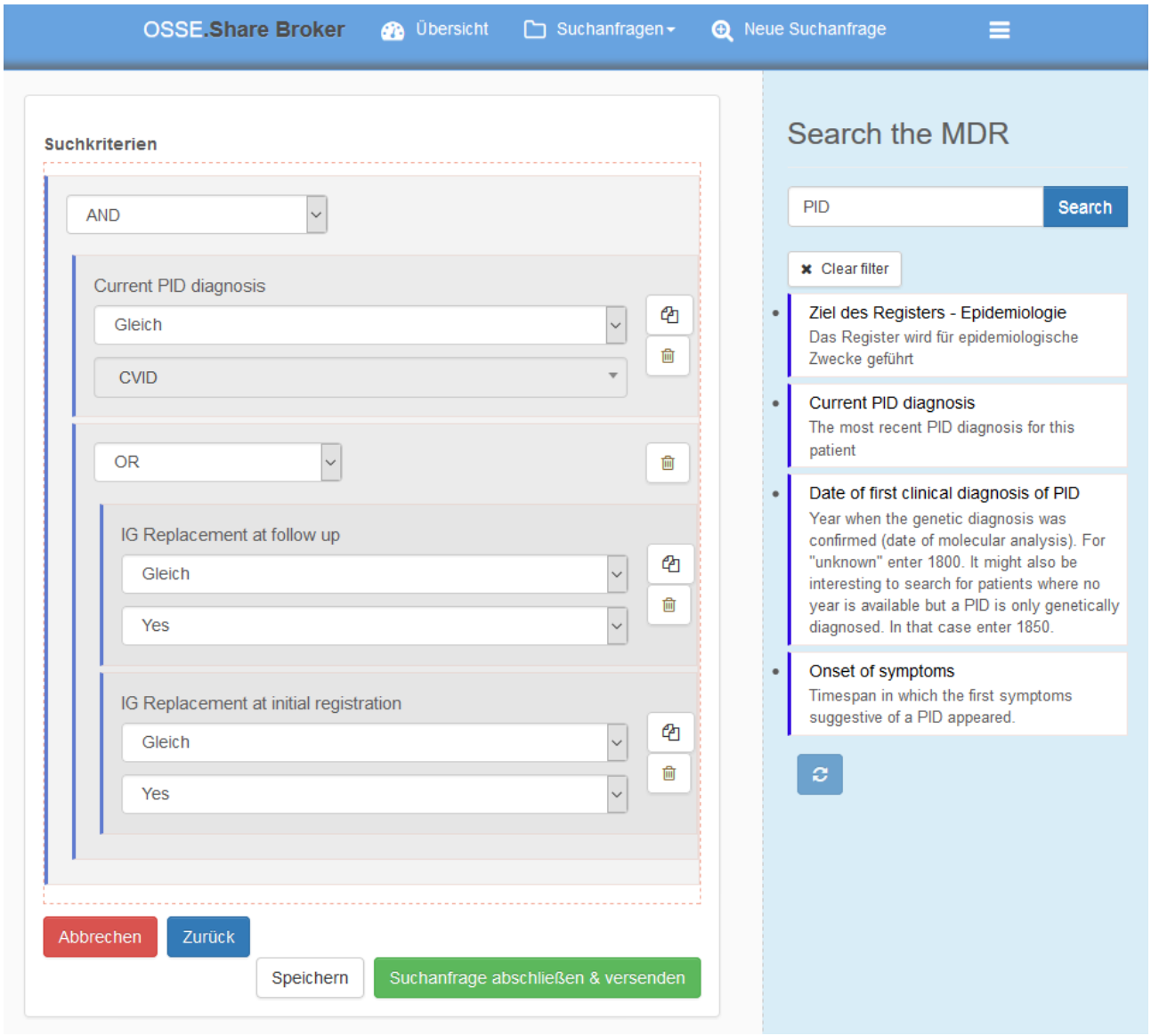
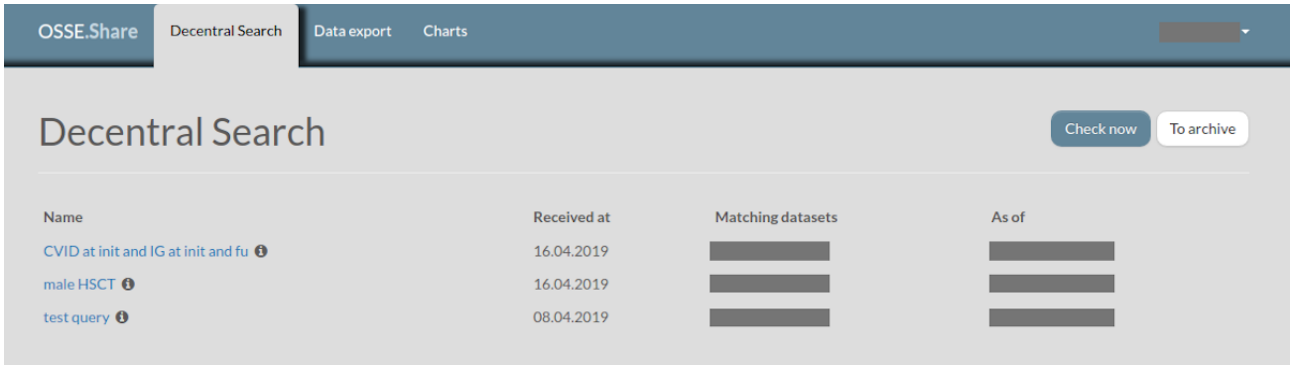


Figure 5. The local collaboration client, which is part of the OSSE bridgehead installation, lists all incoming inquiries. Further, one sees the date of last execution (As of) as well as the number of matching patients (Matching datasets). CVID: common variable immunodeficiency; fu: follow-up; HSCT: hematopoietic stem cell transplantation; IG: immunoglobulin; init: initial registration; OSSE: Open Source Registry System for Rare Diseases (*Open-Source-Registersystem für Seltene Erkrankungen*, in German).



Discussion

Principal Findings

Overview

This work demonstrates how a registry that is not implemented with the OSSE registry framework can be made available in an interoperable manner without losing data sovereignty. With the help of a proposed toolchain, the ESID registry was successfully extended with required data export functionality in order to connect to a local installation of the OSSE bridgehead that adds the decentralized search functionality. Finally, the new feature was demonstrated by executing two example queries.

Limitations

The OSSE bridgehead is provided as a docker container working out of the box with minimal configuration. There is no need to compile and install multiple individual software components. Entering the dataset into the MDR and creating the forms is carried out by web-based user interfaces, which does not require special software, but does require a web browser. However, data harmonization could be fostered by proposing data element candidates: while creating a data element and typing in the designation, the system could search the MDR for similar pre-existing data elements to use. The most complicated development step was to program the ETL process, which extends PID-NET via an OSSE-compatible XML export and uploads this data into the bridgehead. Writing this process was necessary, as our registry has a complex data model and no generic XML export. We highly recommend the use of the Pentaho software. For registry software, which already includes a generic XML export, one could directly develop an Extensible Stylesheet Language (XSL) transformation in order to generate the final XML file. Further, we propose the implementation of a REST interface to the MDR and to extend the bridgehead with functionality to allow the development of a semiautomatic refresh of the MDR dataset and the forms of the bridgehead. There are data fields that will change over time, especially catalogs such as the PID classification. Currently, maintaining the system requires continuous manual work. The graphical query builder offers drag-and-drop and a search functionality. We suggest increasing the usability even further by limiting the selection of data elements. At the time of writing this paper, the column in which these elements are displayed shows every data element from all registries, instead of a limited selection of the registries of interest. Especially as the number of participating registries grows, limiting this list would have a huge impact on its usability and in terms of clarity.

The work presented here is a *proof of principle* for the technical feasibility to amend an existing registry by using external search functionalities. So far it has been evaluated by the PID experts operating the registry with typical queries from their experience.

Outlook

Following the idea of the recently developed ERDRI, we registered the ESID registry in the ERDRI.dor and uploaded an export of the specifications of our data elements (ie, the metadata) from OSSE.MDR into the ERDRI Metadata Repository (ERDRI.mdr). A stronger link between OSSE and ERDRI would be useful (ie, data in ERDRI.dor and ERDRI.mdr could be automatically updatable). Possible further steps would be to bring this work to a broader audience in the field of PID and encourage them to use this tool, rather than inquire with the registries themselves, and to include other registries and eventually other disease domains.

Conclusions

In principle, the OSSE bridgehead allows registry software that is not created with the OSSE registry framework to be extended by a decentralized search functionality while maintaining data sovereignty. The only requirement is access to the raw data of the registry. This data access allows the registry to be extended by an ETL procedure that exports the data in the format the bridgehead requires. The decentralized search feature provides the possibility of more collaborative and transparent research. As the first registry to use the bridgehead, we successfully managed its integration into the OSSE network and successfully demonstrated the decentralized search functionality with two example queries. The setup of the OSSE bridgehead (ie, the installation, registration to a search broker, entering the dataset into the MDR, and creating the forms) could be further simplified. The most complex step was the implementation of the XML export interface, for which we suggest a specific and flexible free toolchain. Since this requires only a one-time effort, the expenditures are within justifiable limits.

Finally, we demonstrated that the OSSE can be used to interconnect registries, based on a federated search functionality, and ultimately made data from the ESID registry available in an interoperable manner without losing data sovereignty.

Acknowledgments

This work was supported by the BMBF (BMBF 01GM0896, 01GM1111B, 01GM1517C, 01EO1303, and 01ZZ1801B) and the ESID. The article processing charge was funded by the University of Freiburg through the Open Access Publishing funding program. Further, we would like to thank James Balmford for constructive criticism of the manuscript.

Authors' Contributions

RS, DK, M Blum, MF, and SR carried out the implementation. GK and HS jointly coordinated and supervised this work. GK further helped to formulate test queries and define the dataset we entered into the MDR. RS and DK wrote the manuscript with input from all authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Excerpt of the dataset of PID-NET (German Network on Primary Immunodeficiency Diseases).
[PDF File (Adobe PDF File), 43 KB - medinform_v8i10e17420_app1.pdf]

References

1. Gathmann B, Goldacker S, Klima M, Belohradsky BH, Notheis G, Ehl S, et al. The German national registry for primary immunodeficiencies (PID). Clin Exp Immunol 2013 Aug;173(2):372-380 [FREE Full text] [doi: 10.1111/cei.12105] [Medline: 23607573]
2. El-Helou SM, Biegner A, Bode S, Ehl SR, Heeg M, Maccari ME, et al. The German National Registry of Primary Immunodeficiencies (2012-2017). Front Immunol 2019;10:1272 [FREE Full text] [doi: 10.3389/fimmu.2019.01272] [Medline: 31379802]
3. Arbeitsgemeinschaft Pädiatrische Immunologie. URL: http://www.api-ev.eu [accessed 2019-12-04]
4. Mahlaoui N, Jais J, Brosselin P, Mignot C, Beaurain B, Brito C, CEREDIH Prevalence Study Collaborators. Prevalence of primary immunodeficiencies in France is underestimated. J Allergy Clin Immunol 2017 Dec;140(6):1731-1733. [doi: 10.1016/j.jaci.2017.06.020] [Medline: 28732644]
5. Nationales Aktionsbündnis für Menschen mit Seltenen Erkrankungen (NAMSE). URL: http://www.namse.de [accessed 2019-07-24]
6. Storf H, Schaaf J, Kadioglu D, Göbel J, Wagner TOF, Ückert F. Registries for rare diseases: OSSE - An open-source framework for technical implementation [Article in German]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2017 May;60(5):523-531. [doi: 10.1007/s00103-017-2536-7] [Medline: 28289778]
7. OSSE – Open Source Registry System for Rare Diseases. URL: https://www.osse-register.de/en/ [accessed 2019-11-28]
8. EUCERD Core Recommendations on Rare Disease Patient Registration and Data Collection. Paris, France: European Union Committee of Experts on Rare Diseases; 2013 Jun 05. URL: http://www.eucerd.eu/wp-content/uploads/2013/06/EUCERD_Recommendations_RDRegistryDataCollection_adopted.pdf [accessed 2019-11-27]
9. Lablans M, Kadioglu D, Muscholl M, Ückert F. Preserving the owner's autonomy in networks of patient registries and biobanks. Orphanet J Rare Dis 2014 Nov 11;9(Suppl 1):Article number P3 [FREE Full text] [doi: 10.1186/1750-1172-9-s1-p3]
10. Lablans M, Borg A, Ückert F. A RESTful interface to pseudonymization services in modern web applications. BMC Med Inform Decis Mak 2015 Feb 07;15:Article number 2 [FREE Full text] [doi: 10.1186/s12911-014-0123-5] [Medline: 25656224]
11. Guzman D, Veit D, Knerr V, Kindle G, Gathmann B, Eades-Perner AM, et al. The ESID Online Database network. Bioinformatics 2007 Mar 01;23(5):654-655. [doi: 10.1093/bioinformatics/btl675] [Medline: 17237056]
12. Scheible R, Rusch S, Guzman D, Mahlaoui N, Ehl S, Kindle G. The NEW ESID online database network. Bioinformatics 2019 Dec 15;35(24):5367-5369. [doi: 10.1093/bioinformatics/btz525] [Medline: 31263866]
13. Muscholl M, Lablans M, Wagner TO, Ückert F. OSSE – Open source registry software solution. Orphanet J Rare Dis 2014 Nov 11;9(Suppl 1):Article number O9 [FREE Full text] [doi: 10.1186/1750-1172-9-s1-o9]
14. Muscholl M, Lablans M, Hirche T, Ückert F. OSSE – Open-Source-Registersystem für Seltene Erkrankungen in der EU (OSSE - Open Source Registry System for Rare Diseases in the EU) [Article in German]. In: Proceedings of the 59th Annual Meeting of the German Society for Medical Informatics, Biometry and Epidemiology e.V (GMDS 2014). 2014 Sep 04 Presented at: 59th Annual Meeting of the German Society for Medical Informatics, Biometry and Epidemiology e.V (GMDS 2014); September 7-10, 2014; Göttingen, Germany URL: https://dx.doi.org/10.3205/14gmds125 [doi: 10.3205/14gmds125]
15. Lablans M, Kadioglu D, Muscholl M, Ückert F. Exploiting distributed, heterogeneous and sensitive data stocks while maintaining the owner's data sovereignty. Methods Inf Med 2015;54(4):346-352. [doi: 10.3414/ME14-01-0137] [Medline: 26196653]
16. Mainzelliste. Bitbucket. URL: http://www.mainzelliste.de [accessed 2012-04-19]
17. Anderson C. Docker [Software engineering]. IEEE Softw 2015 May;32(3):102-105. [doi: 10.1109/ms.2015.62]
18. Bouman R, van Dongen J. Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL. Indianapolis, IN: Wiley Publishing; 2009.
19. Casters M, Bouman R, van Dongen J. Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration. Indianapolis, IN: Wiley Publishing; 2010.
20. O'Grady A. GitLab Quick Start Guide: Migrate to GitLab for All Your Repository Management Solutions. Birmingham, UK: Packt Publishing; 2018.
21. European Rare Disease Registry Infrastructure (ERDRI). URL: https://eu-rd-platform.jrc.ec.europa.eu/erdri-description_en [accessed 2020-09-24]
22. Samply Share Broker - Decentral Search. URL: https://decentralsearch.osse-register.de/ [accessed 2019-12-11]

23. McCarthy J. Recursive functions of symbolic expressions and their computation by machine, Part I. Commun ACM 1960 Apr;3(4):184-195. [doi: 10.1145/367177.367199]
24. Shapiro SC. Common LISP: An Interactive Approach. New York, NY: WH Freeman and Company; 1991.

Abbreviations

API: Pediatric Immunology Working Group (Arbeitsgemeinschaft Pädiatrische Immunologie, in German)
BMBF: German Ministry for Education and Research (Bundesministerium für Bildung und Forschung, in German)
CVID: common variable immunodeficiency
ERDRI: European Rare Disease Registry Infrastructure
ERDRI.dor: ERDRI Directory of Registries
ERDRI.mdr: ERDRI Metadata Repository
ESID: European Society for Immunodeficiencies
ETL: extract, transform, and load
fu: follow-up
HSCT: hematopoietic stem cell transplantation
ICD-10-GM: German Modification of the 10th revision of the International Statistical Classification of Diseases and Related Health Problems
ICD-O-3: 3rd edition of the International Classification of Diseases for Oncology
Ig: immunoglobulin
init: initial registration
MDR: metadata repository
OSSE: Open Source Registry System for Rare Diseases (*Open-Source-Registersystem für Seltene Erkrankungen*, in German)
PID: primary immunodeficiency
PID-NET: German Network on Primary Immunodeficiency Diseases
REST: representational state transfer
XSD: XML Schema Definition
XSL: Extensible Stylesheet Language

Edited by G Eysenbach; submitted 11.12.19; peer-reviewed by A Vagelatos, B Åstrand; comments to author 19.02.20; revised version received 13.03.20; accepted 22.03.20; published 07.10.20.

Please cite as:
Scheible R, Kadioglu D, Ehl S, Blum M, Boeker M, Folz M, Grimbacher B, Göbel J, Klein C, Nieters A, Rusch S, Kindle G, Storf H Enabling External Inquiries to an Existing Patient Registry by Using the Open Source Registry System for Rare Diseases: Demonstration of the System Using the European Society for Immunodeficiencies Registry
JMIR Med Inform 2020;8(10):e17420
URL: http://medinform.jmir.org/2020/10/e17420/
doi:10.2196/17420
PMID:33026355

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org