JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 10 (2022), Issue 6 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Viewpoint

Quality Criteria for Real-world Data in Pharmaceutical Research and Health Care Decision-making: Austrian Expert Consensus (e34204)
Peter Klimek, Dejan Baltic, Martin Brunner, Alexander Degelsegger-Marquez, Gerhard Garhöfer, Ghazaleh Gouya-Lechner, Arnold Herzog, Bernd Jilma, Stefan Kähler, Veronika Mikl, Bernhard Mraz, Herwig Ostermann, Claas Röhl, Robert Scharinger, Tanja Stamm, Michael Strassnig, Christa Wirthumer-Hoche, Johannes Pleiner-Duxneuner. 4

Original Papers

Perspective of Information Technology Decision Makers on Factors Influencing Adoption and Implementation of Artificial Intelligence Technologies in 40 German Hospitals: Descriptive Analysis (e34678)
Lina Weinert, Julia Müller, Laura Svensson, Oliver Heinze. 13

Prevalence of Sensitive Terms in Clinical Notes Using Natural Language Processing Techniques: Observational Study (e38482)
Jennifer Lee, Samuel Yang, Cynthia Holland-Hall, Emre Sezgin, Manjot Gill, Simon Linwood, Yungui Huang, Jeffrey Hoffman. 24

Associations Between Family Member Involvement and Outcomes of Patients Admitted to the Intensive Care Unit: Retrospective Cohort Study (e33921)
Tamryn Gray, Anne Kwok, Khuyen Do, Sandra Zeng, Edward Moseley, Yasser Dbeis, Renato Umeton, James Tulsky, Areej El-Jawahri, Charlotta Lindvall. 35

Automatic International Classification of Diseases Coding System: Deep Contextualized Language Model With Rule-Based Approaches (e37557)
Pei-Fu Chen, Kuan-Chih Chen, Wei-Chih Liao, Feipei Lai, Tai-Liang He, Sheng-Che Lin, Wei-Jen Chen, Chi-Yu Yang, Yu-Cheng Lin, I-Chang Tsai, Chi-Hao Chiu, Shu-Chih Chang, Fang-Ming Hung. 46

A Clinical Decision Support System for Assessing the Risk of Cervical Cancer: Development and Evaluation Study (e34753)
Nasrin Chekin, Haleh Ayatollahi, Mojgan Karimi Zarchi. 59

Personalized Recommendations for Physical Activity e-Coaching (OntoRecoModel): Ontological Modeling (e33847)
Ayan Chatterjee, Andreas Prinz. 69

Experiences and Challenges of Emerging Online Health Services Combating COVID-19 in China: Retrospective, Cross-Sectional Study of Internet Hospitals (e37042)
Fangmin Ge, Huan Qian, Jianbo Lei, Yiqi Ni, Qian Li, Song Wang, Kefeng Ding. 87

Understanding the Relationship Between Mood Symptoms and Mobile App Engagement Among Patients With Breast Cancer Using Machine Learning: Case Study (e30712)
Anna Baglione, Lihua Cai, Aram Bahrini, Isabella Posey, Mehdi Boukhechba, Philip Chow. 122

Noninvasive Screening Tool for Hyperkalemia Using a Single-Lead Electrocardiogram and Deep Learning: Development and Usability Study (e34724)
Erdenebayar Urtnasan, Jung Lee, Byungjin Moon, Hee Lee, Kyuhee Lee, Hyun Youk. 137

Predicting Abnormal Laboratory Blood Test Results in the Intensive Care Unit Using Novel Features Based on Information Theory and Historical Conditional Probability: Observational Study (e35250)
Camilo Valderrama, Daniel Niven, Henry Stelfox, Joon Lee. 146

Noninvasive Diagnosis of Nonalcoholic Steatohepatitis and Advanced Liver Fibrosis Using Machine Learning Methods: Comparative Study With Existing Quantitative Risk Scores (e36997)
Yonghui Wu, Xi Yang, Heather Morris, Matthew Gurka, Elizabeth Shenkman, Kenneth Cusi, Fernando Bril, William Donahoo. 165

Medication-Wide Association Study Using Electronic Health Record Data of Prescription Medication Exposure and Multifetal Pregnancies: Retrospective Study (e32229)
Lena Davidson, Silvia Canelón, Mary Boland. 178

Error and Timeliness Analysis for Using Machine Learning to Predict Asthma Hospital Visits: Retrospective Cohort Study (e38220)
Xiaoyi Zhang, Gang Luo. 191

The Prediction of Preterm Birth Using Time-Series Technology-Based Machine Learning: Retrospective Cohort Study (e33835)
Yichao Zhang, Sha Lu, Yina Wu, Wensheng Hu, Zhenming Yuan. 200

Machine Learning Support for Decision-Making in Kidney Transplantation: Step-by-step Development of a Technological Solution (e34554)
François-Xavier Paquette, Amir Ghassemi, Olga Bukhtiyarova, Moustapha Cisse, Natanael Gagnon, Alexia Della Vecchia, Hobivola Rabearivelo, Youssef Loudiyi. 214

Multitask Learning With Recurrent Neural Networks for Acute Respiratory Distress Syndrome Prediction Using Only Electronic Health Record Data: Model Development and Validation Study (e36202)
Carson Lam, Rahul Thapa, Jenish Maharjan, Keyvan Rahmani, Chak Tso, Navan Singh, Satish Casie Chetty, Qingqing Mao. 227

Identifying the Risk of Sepsis in Patients With Cancer Using Digital Health Care Records: Machine Learning–Based Approach (e37689)
Donghun Yang, Jimin Kim, Junsang Yoo, Won Cha, Hyojung Paik. 246

Predicting Risk of Hypoglycemia in Patients With Type 2 Diabetes by Electronic Health Record–Based Machine Learning: Development and Validation (e36958)
Hao Yang, Jiaxi Li, Siru Liu, Xiaoling Yang, Jialin Liu. 257

Vaccine Adverse Event Mining of Twitter Conversations: 2-Phase Classification Study (e34305)
Sedigheh Khademi Habibabadi, Pari Delir Haghighi, Frada Burstein, Jim Buttery. 269

Predicting 30-Day Readmission Risk for Patients With Chronic Obstructive Pulmonary Disease Through a Federated Machine Learning Architecture on Findable, Accessible, Interoperable, and Reusable (FAIR) Data: Development and Validation Study (e35307)
Celia Alvarez-Romero, Alicia Martinez-Garcia, Jara Ternero Vega, Pablo Díaz-Jimènez, Carlos Jimènez-Juan, María Nieto-Martín, Esther Román Villarán, Tomi Kovacevic, Darijo Bokan, Sanja Hromis, Jelena Djekic Malbasa, Suzana Besla , Bojan Zaric, Mert Gencturk, A Sinaci, Manuel Ollero Baturone, Carlos Parra Calderón. 284

Virtual Specialist Care During the COVID-19 Pandemic: Multimethod Patient Experience Study (e37196)
Katie Dainty, M Seaton, Antonio Estacio, Lisa Hicks, Trevor Jamieson, Sarah Ward, Catherine Yu, Jeffrey Mosko, Charles Kassardjian. 295

Conditional Probability Joint Extraction of Nested Biomedical Events: Design of a Unified Extraction Framework Based on Neural Networks (e37804)
Yan Wang, Jian Wang, Huiyi Lu, Bing Xu, Yijia Zhang, Santosh Banbhrani, Hongfei Lin. 305

Review

Combating COVID-19 Using Generative Adversarial Networks and Artificial Intelligence for Medical Images: Scoping Review (e37365)
Hazrat Ali, Zubair Shah. 106

Viewpoint

Quality Criteria for Real-world Data in Pharmaceutical Research and Health Care Decision-making: Austrian Expert Consensus

Peter Klimek[1,2], PhD; Dejan Baltic[3], MD; Martin Brunner[4], MD; Alexander Degelsegger-Marquez[5], PhD; Gerhard Garhöfer[4], MD; Ghazaleh Gouya-Lechner[3], MD; Arnold Herzog[6], DI; Bernd Jilma[4], MD; Stefan Kähler[7], PhD; Veronika Mikl[3], MA; Bernhard Mraz[3], MA; Herwig Ostermann[5], PhD; Claas Röhl[8], Ing; Robert Scharinger[9], MA; Tanja Stamm[4], PhD; Michael Strassnig[10], PhD; Christa Wirthumer-Hoche[6], PhD; Johannes Pleiner-Duxneuner[3], MD

[1]Institute for Science of Complex Systems, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria

[2]Complexity Science Hub Vienna, Vienna, Austria

[3]Gesellschaft für Pharmazeutische Medizin, Vienna, Austria

[4]Medical University of Vienna, Vienna, Austria

[5]Gesundheit Österreich GmbH, Vienna, Austria

[6]Austrian Medicines and Medical Devices Agency (AGES Medizinmarktaufsicht), Vienna, Austria

[7]Verband der pharmazeutischen Industrie Österreichs (PHARMIG), Vienna, Austria

[8]EUPATI Austria, Vienna, Austria

[9]Federal Ministry of Social Affairs, Health, Care and Consumer Protection, Vienna, Austria

[10]Vienna Science and Technology Fund, Vienna, Austria

Corresponding Author:
Johannes Pleiner-Duxneuner, MD
Gesellschaft für Pharmazeutische Medizin
Engelhorngasse 3
Vienna, 1210
Austria
Phone: 43 1 40160 ext 36255
Email: johannes.pleiner-duxneuner@roche.com

Abstract

Real-world data (RWD) collected in routine health care processes and transformed to real-world evidence have become increasingly interesting within the research and medical communities to enhance medical research and support regulatory decision-making. Despite numerous European initiatives, there is still no cross-border consensus or guideline determining which qualities RWD must meet in order to be acceptable for decision-making within regulatory or routine clinical decision support. In the absence of guidelines defining the quality standards for RWD, an overview and first recommendations for quality criteria for RWD in pharmaceutical research and health care decision-making is needed in Austria. An Austrian multistakeholder expert group led by Gesellschaft für Pharmazeutische Medizin (Austrian Society for Pharmaceutical Medicine) met regularly; reviewed and discussed guidelines, frameworks, use cases, or viewpoints; and agreed unanimously on a set of quality criteria for RWD. This consensus statement was derived from the quality criteria for RWD to be used more effectively for medical research purposes beyond the registry-based studies discussed in the European Medicines Agency guideline for registry-based studies. This paper summarizes the recommendations for the quality criteria of RWD, which represents a minimum set of requirements. In order to future-proof registry-based studies, RWD should follow high-quality standards and be subjected to the quality assurance measures needed to underpin data quality. Furthermore, specific RWD quality aspects for individual use cases (eg, medical or pharmacoeconomic research), market authorization processes, or postmarket authorization phases have yet to be elaborated.

(JMIR Med Inform 2022;10(6):e34204) doi:10.2196/34204

KEYWORDS

real-world data; real-world evidence; data quality; data quality criteria; RWD quality recommendations; pharmaceutical research; health care decision-making; quality criteria for RWD in health care; Gesellschaft für Pharmazeutische Medizin; GPMed

Introduction

Real-world data (RWD) is an overarching term for data on patient's health (health status, effectiveness, medical treatment, the pattern of use of medicinal products, and resource use, etc) that are collected in routine health care processes and not in the context of clinical trials. RWD involve large and complex data sets such as data from electronic health records, pharmacy data, electronic smart devices, patient-reported outcomes, and digital applications or platforms [1,2]. When RWD are analyzed, they lead to real-world evidence (RWE) on the pattern of use and effectiveness of any kind of procedure, drug, or nonpharmacological intervention. The availability of RWD and evolving analytic techniques to generate RWE have created interest within the research and medical communities to use RWD and RWE to enhance clinical research and support regulatory decision-making [1,3]. On a European level, the European Medicines Agency (EMA) and Heads of Medicines Agencies fully recognize the value of health data and set up a joint task force to describe the health data landscape from a regulatory perspective and identify practical steps for the European medicines regulatory network to make the best use of health data in support of innovation and public health in the European Union [4].

The comprehensive work plan identifies 10 priorities [5], such as delivering a sustainable platform to access and analyze health care data from across the European Union (Data Analysis and Real World Interrogation Network [6]) or establishing an EU framework for data quality (European Health Data & Evidence Network [7] and Health Outcomes Observatory [8]) and representativeness. Despite many initiatives, there are still no guidelines for the quality criteria that RWD must meet in order to be able to use it for decision-making purposes within regulatory or routine clinical decision support. As a first example, the EMA Guideline on registry-based studies [9] provides considerations on good practice for registries to increase their usefulness for regulatory purposes.

The objective of this consensus statement of the Austrian Expert Group led by Gesellschaft für Pharmazeutische Medizin (GPMed; Austrian Society for Pharmaceutical Medicine) is to provide an overview and first recommendations for the quality criteria of RWD for primary and secondary research purposes to be adopted in medical or pharmacoeconomic research and health care decision-making processes. The consensus statement does not discuss the general use of RWD nor how to obtain RWE in general.

Methods

After EMA published a drafted guideline for registry-based studies, interested GPMed board members volunteered together with Austrian Medicines and Medical Devices Agency executive experts to assess how ready the Austrian research landscape is for registry-based studies.

The Austrian Medicines and Medical Devices Agency and GPMed invited Austrian RWD researchers and data experts to contribute voluntarily to the topic. The criteria to select working group members were those with scientific work in the field and longstanding expertise in using RWD for research purposes. After the kickoff meeting in April 2021, the expert group led by GPMed met on a monthly basis; reviewed guidelines, frameworks, use cases, or viewpoints; and derived a consensus statement on the quality criteria for RWD to be used more effectively for medical research purposes beyond the registry-based studies discussed in the EMA Guideline for registry-based studies [9].

Following agreement on a joint definition on RWD, experts from the group shared examples of RWD frameworks, guidelines, or viewpoints, which were discussed in the working group, and consensus was reached unanimously within the monthly meetings.

Results

Definition of RWD

Despite an increasing recognition of the value of RWD, a global consensus on the definition of RWD is lacking [10]. The definition of RWD can differ in various areas of application (eg, public health vs automotive industry). However, the expert group led by GPMed reviewed several definitions [7,8,10-15] and agreed on the following description.

Real-world data can be defined as data relating to patient health status or the delivery of health care that are routinely collected from a variety of sources (including patient-reported outcomes), such as:

- health care databases (systems into which health care providers routinely enter clinical and laboratory data; eg, electronic health records and pharmacist databases),
- health insurance and claims databases (maintained by payers for reimbursement purposes),
- patient registries (data on a group of patients with specific characteristics in common),
- disease registries (data on a particular disease or disease-related patient characteristic regardless of exposure to any medicinal product, other treatment, or a particular health service),
- data gathered from other sources that can inform on health status, such as mobile devices, wearables, or other smart medicinal products (eg, real-time continuous glucose monitoring devices),
- social media– and patient-powered research networks (eg, patient networks to share health information),
- biobanks, and
- observational studies.

Note that this definition includes data that are neither collected by licensed medical devices operated by health professionals in clinical settings nor observational data that are typically stored in public health registries and administrative databases. Namely, RWD also include health-related data that are generated by the patient by means of digital health technologies (sensors, wearables, and smartphones, etc). Hence, ethical and regulatory frameworks should also be applied to these health-related data and not only target health care databases and registries [16].

Examples of RWD Frameworks

Globally and Europe-wide, more and more examples of how RWD are used for research or regulatory purposes are being published. The expert group decided to illustrate some examples of how the quality of RWD is ensured along different approaches (Table 1). Further details to this overview can be found in the Multimedia Appendix 1.

Table 1. Examples and short descriptions of reviewed real-world data (RWD) frameworks.

RWD framework	Short description	Country
RWD for health systems research [17-23]	Nordic countries have set the worldwide gold standard for how RWD can be leveraged. Good RWD frameworks exist in Finland, Denmark, Sweden, Iceland, and Norway. The RWD quality and infrastructure built up in these countries can be seen as best practice examples for how to leverage RWD for research.	Denmark, Finland, Iceland, Norway, and Sweden
Danish Data Analytics Center [24]	The Danish DAC[a] has access to some of the most sophisticated and complete patient-level health data in the world and meets the highest requirements for data and IT security. DAC constitutes a unique possibility for the use of big data analytics to discover hidden patterns to benefit patients. It will reduce the entry barriers for new drugs to go to market while maintaining the high safety standards currently in place.	Denmark
EMA[b] submission supported by historical cohort patient data [25]	Based on the observed efficacy in Phase 2 studies (n=189 and n=36) and combined with an additional historical comparator study (1139 cases), conditional marketing authorization was granted with the need to better quantify the magnitude of the effect by submitting data from a Post Authorization Efficacy Study (Phase 3 randomized, comparative study of blinatumomab vs standard of care chemotherapy) as well as a noninterventional Post Authorization Safety Study in subsequent years.	European Union
Demonstrated the research potential of a clinico-genomic database [26,27]	In 2017, Foundation Medicine and Flatiron Health created a proof-of-concept study. Using a sample size of over 2000 patients with non–small cell lung cancer, they discovered that high versus low tumor mutation burden showed a far stronger association than high versus low PD-L1 levels after immunotherapy. Their results were nearly identical to those derived by a drug manufacturer from a post hoc analysis of a failed clinical trial. The validation study helped establish the groundwork for this data set to be used to advance cancer research.	United States
Multidatabase studies for medicines surveillance in real-world settings [28,29]	Postmarketing studies can be underpowered if outcomes or exposure of interest are rare, or the interest is in the subgroup effects. Combining several databases might provide the statistical power needed. Although many multidatabase studies have been performed in Europe in the past 10 years, there is a lack of clarity on the peculiarities and implications of the existing strategies to conduct them. Experts identified 4 strategies to execute multidatabase studies, classified according to specific choices in the execution.	European Union
EUnetHTA[c] REQueST[d] [30]	The Registry Evaluation and Quality Standards Tool (REQueST) aims to support health technology assessment organizations and other actors in guiding and evaluating registries for effective use in health technology assessment.	European Union

[a]DAC: Data Analytics Center.

[b]EMA: European Medicines Agency.

[c]EUnetHTA: European Network for Health Technology Assessment.

[d]REQueST: Registry Evaluation and Quality Standards Tool.

Legal Frameworks

The current legal framework in Austria with the Federal Statistics Act as well as the Research Organization Act recognizes the “use” of RWD—especially for research purposes [31-33].

Independently of the question of data availability, many RWD sources, as defined within this expert consensus paper, do not address data quality issues. Therefore, the need for high–data quality standards should be also recognized by legal frameworks. On a European level, data quality aspects are strongly embedded within the development of the European Health Data Space [34] and Data Analysis and Real World Interrogation Network [6]. Shared outcomes on data quality should be reflected within local legal frameworks as well.

Recommendations

Data Quality

RWD are often used for purposes that are different from the intention for which the data were collected originally. Therefore, it is of utmost importance to check upfront if the RWD are adequate in terms of clearly defined quality criteria and can, therefore, be used in general for primary or secondary research purposes as well. Due to the lack of guidelines defining the quality standards of RWD to be used for decision-making, it is even more important to be able to assess the suitability of RWD for research purposes by applying checklists and some standardized questionnaires [35-38].

RWD Should Follow High Standards and Be Subject to Quality Assurance

The value of the secondary use of RWD data (in particular, registries) for research purposes depends crucially on their quality as quantified by *completeness* and *accuracy* [39], next to *timeliness*, *comparability*, the technical prerequisite that the

size of the data source is sufficient (ie, the study does not become underpowered), and that the data is in principle accessible and can be mapped with other relevant data sets (well defined research question outlined in a research plan). An evaluation with regard to these factors is therefore recommended before using the data. Note that these quality criteria are not unique in the sense that alternative data quality concepts have also been described (eg, validity, consistency, and integrity).

Completeness is defined as the proportion of true cases of a variable (disease, treatment, and diagnose, etc) in all or a certain subgroup of patients that is correctly reported in the data. Completeness therefore captures the amount of missing data in a specific source—the extent to which all necessary data that could have been registered has been registered [40]. Very often there is no comprehensive reference source available for evaluating the completeness of a data set with regard to the general population. In that case, it might be advisable to identify studies that report the variables of interest for specific comparable subgroups and therefore allow for an assessment of data completeness [39]. These comparisons should ideally be performed on an individual level (eg, comparing data records from registries for certain diseases to administrative records) or, in cases where the required information is not available on an individual level, attempts should be made to examine completeness at least on an aggregate level (by comparing the expected number of cases across data sets).

Accuracy measures the proportion of patients with a certain property (diagnosis, prescription, and socioeconomic or demographic properties, etc) in a data set that truly have the property. Accuracy is typically assessed by comparing the data records with the reference standard used to confirm the specific variable [41]. In many cases, this reference could be the medical record; for certain areas, other references might be feasible as well. One strategy to perform such a comparison could be to randomly sample a given percentage (eg, 5%) or an absolute number (eg, 1000) manually. This helps to identify errors and whether they are systematic (as often happens through algorithmic problems when the data are collected in an automated way or if the data are collated from different reporting systems, regional or otherwise) or random (often resulting from manual data collection), thereby informing strategies to increase data accuracy.

Timeliness measures data quality with regard to the time at which the variable (disease and diagnosis, etc) was recorded (eg, the extent to which the time of the recorded disease corresponds to the true time of the disease). This can often be assessed together with completeness and accuracy and is of particular importance in longitudinal study designs.

Furthermore, *comparability* needs to be checked to ensure that variable definitions in a data set conform to international guidelines and other relevant references.

A comprehensive review of 114 data quality studies in the Danish registry network showed that both completeness and accuracy increased over time and accuracy varies substantially across different diseases, between less than 15% of correctly coded diagnoses to almost 100% [41]. This finding underscores the *need for data quality assurance of RWD for research use*.

High Research Standards Should Underpin the Quality of RWD

Study Protocol

Observational postmarketing studies are an important tool, using data obtained from routine clinical care, to provide data on medical treatment effect estimates and the tolerability of medicinal products in a real-world setting, as well as for medical devices as part of the postmarketing surveillance [42]. Nonrandomized studies may be used to complement the evidence base represented by randomized controlled trials [43], even though one cannot expect nonrandomized, observational studies to exactly reproduce randomized controlled trials as these are different study designs, and hence measure different types of effects [44]. Noncontrolled studies lack a comparison group, which means that inferences on the treatment effect and tolerability must rely on before-and-after comparisons of the outcome of interest. Treatment effect estimates and tolerability derived from nonrandomized studies are at greater risk of bias. Thus, data from routine clinical observation should be collected after the development of a study protocol where the population of interest, study outcome, methods for data generation and analysis, limitation of study data, and bias are defined in advance, as also defined in the EMA guideline for registry-based studies [9].

Informed Consent

The informed consent process of patients in observational, noninterventional studies are not discussed by Good Clinical Practice (ISO 14155) [45], and this topic is still dealt with heterogeneously throughout the European Union. Within the study protocol, the consent process and requirements of compliance to the General Data Protection Regulation (GDPR) should be specified. Data generated in an anonymized way would not require patient consent, though collection of pseudonymized data in observational studies requires the consent of patients prior to data collection, which should be limited only to the GDPR requirements, and not include any consent to medical treatment. The burden of obtaining informed consent to collect routine clinical data should be kept feasible to reduce bias of missing data from severely ill patients or patients incapable of consenting, such as in emergency situations. Since GDPR applies only to living people, a waiver for data collection from the deceased can be obtained if the purpose is sufficiently outlined in the study protocol.

Institutional Review Board and Ethics Committee

Within the study protocol, all interventions in the observational trial (ie, treatment, diagnostic or monitoring procedures) should fall within the standard of care or routine treatment, as interpreted by the competent authority or ethics committee in that member state. Thus, a review and approval from the respective ethics committee is required, as also indicated in the EMA guideline for registry-based studies [9].

Checklist on Quality Criteria for RWD

Following general recommendations and reflecting guidelines and checklists on registry-based research [9,37], the expert group suggests a minimum set of criteria summarized within

the checklist presented in Table 2 to ensure the quality of RWD for research purposes and health care decision-making processes.

Table 2. Gesellschaft für Pharmazeutische Medizin (GPMed) checklist for real-world data (RWD) quality.

Criteria	Description
Data management and stewardship	• "FAIR Data Principles" which formulate principles that sustainable, reusable research data and research data infrastructures must meet [38,46,47]
Governance framework	• Available policy for collaborations with external organizations • Involvement of patient organizations • Governance structure for decision-making on requests for collaboration • Templates for research and data-sharing contracts between partners and institutions
Quality requirements	• High–RWD quality standards are implemented, such as completeness, accuracy, timeliness, and comparability • Process in place for ongoing data quality assessments • Processes in place for quality planning, control, assurance, and improvement • Data verification (the method and frequency of verification) • Auditing practice
Data privacy and transparency	• Informed consent processes and its validity for research purposes according to General Data Protection Regulation and relevant national regulations • Data privacy officer
Research objectives	• Well-defined research question outlined in a research plan • Available documentation, protocol, or proposal that describes the purpose of RWD use and rational that the RWD sources adequately address the research questions (eg, study protocol) • Approval of RWD use from independent an institutional review board or ethics committee • Protocol should follow the Declaration of Helsinki, and furthermore, the Declaration of Taipei [48] on Research on Health Databases, Big Data and Biobanks should be taken into account
Data providers	• Adequate description of data providers, such as patients, caregivers, or health care professionals; their geographical area; and any selection process (inclusion and exclusion criteria) that may be applied for their acceptance as data providers
Patient population covered	• Adequate description of the type of patient population (disease, condition, time period covered, and procedure), which defines the criteria for patient eligibility • Relevance of setting and catchment area • Clarity on patients' inclusion and exclusion criteria • Methods applied to minimize selection bias and loss to follow-up • Ensure fair representations of minorities, sex, gender, and socially disadvantaged groups
Data elements	• Core RWD set collected for RWD use case or purpose • Definition, dictionary, and format of data elements • Standards and terminologies applied • Capabilities and plans for amendments of data elements
Infrastructure	• High-quality systems for RWD collection, recording, and reporting, including timelines • Capability (and experience) for expedited reporting and evaluation of severe suspected adverse reactions in RWD collection • Capability (and experience) for periodic reporting of clinical outcomes—ideally patient-reported outcomes—and adverse events reported by physicians, at the individual-patient level and aggregated data level • Capability (and experience) for data cleaning, extraction, transformation, and analysis • Capability (and experience) for data transfer to external organizations • Capabilities for amendment of safety reporting processes

Discussion

Principle Findings

Over the past months, EU and EMA strategies, workplans, and initiatives on health data use developed very quickly [34,49-51]. This paper shows the consensus of a multistakeholder expert group which summarizes a minimum set of the quality criteria of RWD for research and decision-making purposes in health care. The most important quality assurance measures identified are a profound data management and stewardship; established governance framework; standardized quality requirements; adhered data privacy and transparency measures; well-defined research objectives; adequate description of data providers; well-described patient population covered; outlined which data elements are required; and high-quality infrastructure for RWD collection, recording, and reporting.

Conclusions

To future-proof registry-based studies, the group strongly recommends that RWD should follow high standards and be subject to the quality assurance measures needed to underpin the quality of RWD. Furthermore, specific RWD quality aspects for individual use cases (eg, medical or pharmacoeconomic research), market authorization processes, or postmarket authorization phases have yet to be elaborated.

Acknowledgments

The work was supported in kind by the participating organizations. The authors declare no financial support or funding for this project.

Conflicts of Interest

DB is an employee of Amgen GmbH, Vienna, Austria. SK is an employee of Bristol Myers Squibb, Vienna, Austria. BM is an employee of Novartis Pharma GmbH, Austria. TS reports grants and personal fees from AbbVie, Roche, Sanofi, Takeda, and Novartis, all outside the submitted work. VM and JPD are employees of Roche Austria GmbH. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Examples of real-world data frameworks or use cases.
[DOCX File , 32 KB - medinform_v10i6e34204_app1.docx]

References

1. Submitting documents using real-world data and real-world evidence to FDA for drugs and biologics: guidance for industry. U.S. Food & Drug Administration. 2019 May. URL: https://www.fda.gov/media/124795/download [accessed 2021-09-17]
2. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence - what is it and what can it tell us? N Engl J Med 2016 Dec 08;375(23):2293-2297. [doi: 10.1056/NEJMsb1609216] [Medline: 27959688]
3. Babrak LM, Smakaj E, Agac T, Asprion PM, Grimberg F, der Werf DV, et al. RWD-Cockpit: application for quality assessment of real-world data. JMIR Form Res. Preprint posted online on Feb 19, 2022. [FREE Full text] [doi: 10.2196/29920] [Medline: 35266872]
4. Big Data Steering Group workplan 2021-2023. European Medicines Agency. URL: https://www.ema.europa.eu/en/documents/work-programme/workplan-2021-2023-hma/ema-joint-big-data-steering-group_en.pdf [accessed 2021-09-17]
5. Priority recommendations of the HMA-EMA joint Big Data Task Force. European Medicines Agency. URL: https://www.ema.europa.eu/en/documents/other/priority-recommendations-hma-ema-joint-big-data-task-force_en.pdf [accessed 2021-09-17]
6. Data Analysis and Real World Interrogation Network (DARWIN EU). European Medicines Agency. URL: https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu [accessed 2021-09-17]
7. Result of the 4th open call for data partners. European Health Data & Evidence Network. URL: https://us20.campaign-archive.com/?u=123c73def0355ab534c08baa9&id=214d8c4a9e [accessed 2021-09-17]
8. Stamm T, Bott N, Thwaites R, Mosor E, Andrews MR, Borgdorff J, et al. Building a value-based care infrastructure in Europe: the Health Outcomes Observatory. NEJM Catalyst 2021 Jun 09 [FREE Full text] [doi: 10.1056/CAT.21.0146]
9. Committee for Human Medicinal Products (CHMP). Guideline on registry-based studies. European Medicines Agency. 2021 Oct 22. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-registry-based-studies_en-0.pdf [accessed 2022-05-05]
10. Makady A, de Boer A, Hillege H, Klungel O, Goettsch W. What is real-world data? a review of definitions based on literature and stakeholder interviews. Value Health 2017 Jul 01;20(7):858-865 [FREE Full text] [doi: 10.1016/j.jval.2017.03.008] [Medline: 28712614]
11. HMA-EMA Joint Big Data Taskforce phase II report: 'evolving data-driven regulation'. European Medicines Agency. 2019. URL: https://www.ema.europa.eu/en/documents/other/hma-ema-joint-big-data-taskforce-phase-ii-report-evolving-data-driven-regulation_en.pdf [accessed 2021-05-11]
12. Read-world evidence. U.S. Food & Drug Administration. URL: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence [accessed 2021-05-11]
13. Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. JAMA 2018 Sep 04;320(9):867-868. [doi: 10.1001/jama.2018.10136] [Medline: 30105359]
14. MHRA position statement and guidance: electronic health records. Medicines & Healthcare Products Regulatory Agency. 2015 Sep. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/470228/Electronic_Health_Records_MHRA_Position_Statement.pdf [accessed 2022-05-05]
15. Here, no one goes at it alone. PatientsLikeMe. URL: https://www.patientslikeme.com/ [accessed 2021-09-21]

16. Vayena E, Haeusermann T, Adjekum A, Blasimme A. Digital health: meeting the ethical and policy challenges. Swiss Med Wkly 2018;148:w14571 [FREE Full text] [doi: 10.4414/smw.2018.14571] [Medline: 29376547]
17. Sørensen HT. Regional administrative health registries as a resource in clinical epidemiology: a study of options, strengths, limitations and data quality provided with examples of use. Int J Risk Saf Med 1997;10(1):1-22. [doi: 10.3233/JRS-1997-10101] [Medline: 23511270]
18. Sund R. Quality of the Finnish Hospital Discharge Register: a systematic review. Scand J Public Health 2012 Aug 16;40(6):505-515. [doi: 10.1177/1403494812456637] [Medline: 22899561]
19. Lynge E, Sandegaard JL, Rebolj M. The Danish National Patient Register. Scand J Public Health 2011 Jul 20;39(7 Suppl):30-33. [doi: 10.1177/1403494811401482] [Medline: 21775347]
20. Ludvigsson JF, Andersson E, Ekbom A, Feychting M, Kim J, Reuterwall C, et al. External review and validation of the Swedish national inpatient register. BMC Public Health 2011 Jun 09;11(1):450 [FREE Full text] [doi: 10.1186/1471-2458-11-450] [Medline: 21658213]
21. Gudbjornsson B, Thorsteinsson SB, Sigvaldason H, Einarsdottir R, Johannsson M, Zoega H, et al. Rofecoxib, but not celecoxib, increases the risk of thromboembolic cardiovascular events in young adults-a nationwide registry-based study. Eur J Clin Pharmacol 2010 Jun 16;66(6):619-625. [doi: 10.1007/s00228-010-0789-2] [Medline: 20157701]
22. Norsk pasientregister (NPR). Norwegian Directorate of Health. URL: https://www.helsedirektoratet.no/tema/statistikk-registre-og-rapporter/helsedata-og-helseregistre/norsk-pasientregister-npr [accessed 2022-06-09]
23. Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. Clin Epidemiol 2015 Nov 17;7:449-490 [FREE Full text] [doi: 10.2147/CLEP.S91125] [Medline: 26604824]
24. Data Analytics Centre. Danish Medicines Agency. URL: https://laegemiddelstyrelsen.dk/en/about/organisation/name/ [accessed 2021-09-21]
25. Assessment report: BLINCYTO, international non-proprietary name: blinatumomab. European Medicines Agency. URL: https://www.ema.europa.eu/en/documents/assessment-report/blincyto-epar-public-assessment-report_en.pdf [accessed 2022-05-30]
26. Getting closer to cancer research's holy grail: the Clinico-Genomic Database. Roche. 2019 Apr 16. URL: https://www.roche.com/about/priorities/personalised_healthcare/combining-data-to-advance-personalised-healthcare.htm [accessed 2021-05-21]
27. Singal G, Miller PG, Agarwala V, Li G, Kaushik G, Backenroth D, et al. Association of patient characteristics and tumor genomics with clinical outcomes among patients with non-small cell lung cancer using a clinicogenomic database. JAMA 2019 Apr 09;321(14):1391-1399 [FREE Full text] [doi: 10.1001/jama.2019.3241] [Medline: 30964529]
28. Gini R, Sturkenboom MCJ, Sultana J, Cave A, Landi A, Pacurariu A, Working Group 3 of ENCePP (Inventory of EU data sourcesmethodological approaches for multisource studies). Different strategies to execute multi-database studies for medicines surveillance in real-world setting: a reflection on the European model. Clin Pharmacol Ther 2020 Aug;108(2):228-235 [FREE Full text] [doi: 10.1002/cpt.1833] [Medline: 32243569]
29. Real world research on medicines: contribution of the European Network of Centres in Pharmacoepidemiology and Pharmacovigilance (ENCePP). European Medicines Agency. 2021 Mar 08. URL: https://www.ema.europa.eu/en/events/real-world-research-medicines-contribution-european-network-centres-pharmacoepidemiology#event-summary-section [accessed 2021-05-21]
30. REQueST Tool and its vision paper. EUnetHTA. URL: https://eunethta.eu/request-tool-and-its-vision-paper/ [accessed 2021-05-21]
31. Registerforschung. Austrian Federal Ministry of Education, Science and Research. URL: https://www.bmbwf.gv.at/Themen/Forschung/Forschung-in-%C3%96sterreich/Strategische-Ausrichtung-und-beratende-Gremien/Leitthemen/Registerforschung.html [accessed 2021-09-21]
32. König T, Schmoigl L. Erfolgreiche Registerforschung in Österreich: Welchen Mehrwert generiert die reglementierte Öffnung von Registerdaten für die wissenschaftliche Forschung? Eine Darstellung anhand von drei Beispielen. Österreichisches Institut für Wirtschaftsforschung (WIFO) und Institut für höhere Studien (IHS). 2020 Nov. URL: https://irihs.ihs.ac.at/id/eprint/5576/7/koenig-schmoigl-2020-erfolgreiche-registerforschung-in-oesterreich.pdf [accessed 2021-09-21]
33. König T, Strassnig M, Schwarz G, Oberhofer H. Zugang zu Register- und Individualdaten für die wissenschaftliche Forschung in Österreich. fteval Journal for Research and Technology Policy Evaluation 2020;50:11-15. [doi: 10.22163/fteval.2020.464]
34. European Health Data Space. European Commission. URL: https://ec.europa.eu/health/ehealth/dataspace_en [accessed 2022-05-04]
35. A validated checklist for evaluating the quality of observational cohort studies for decision-making support. Grace Principles. URL: https://www.graceprinciples.org/doc/GRACE-Checklist-031114-v5.pdf [accessed 2021-07-12]
36. Motheral B, Brooks J, Clark MA, Crown WH, Davey P, Hutchins D, et al. A checklist for retrospective database studies--report of the ISPOR Task Force on Retrospective Databases. Value Health 2003 Mar;6(2):90-97 [FREE Full text] [doi: 10.1046/j.1524-4733.2003.00242.x] [Medline: 12641858]

37. Determining real-world data's fitness for use and the role of reliability. Duke-Margolis Center for Health Policy. 2019 Sep 26. URL: https://healthpolicy.duke.edu/sites/default/files/2019-11/rwd_reliability.pdf [accessed 2021-09-21]
38. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3(1):160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
39. Sorensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. Int J Epidemiol 1996 Apr;25(2):435-442. [doi: 10.1093/ije/25.2.435] [Medline: 9119571]
40. Arts DGT, De Keizer NF, Scheffer G. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc 2002;9(6):600-611 [FREE Full text] [doi: 10.1197/jamia.m1087] [Medline: 12386111]
41. Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. CLEP 2015 Nov:449. [doi: 10.2147/clep.s91125]
42. Directive of the European Parliament and of the Council of 6 October 1997 Amending Council Directives 90/387/EEC and 92/44/EEC for the Purpose of Adaptation to a Competitive Environment in Telecommunications (97/51/EC). The European Parliament and the Council of the European Union. 1997 Oct 29. URL: https://doi.org/10.1163/9789004481466_018 [accessed 2022-05-30]
43. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol 2016 Apr 15;183(8):758-764 [FREE Full text] [doi: 10.1093/aje/kwv254] [Medline: 26994063]
44. Groenwold RHH. Trial emulation and real-world evidence. JAMA Netw Open 2021 Mar 01;4(3):e213845 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.3845] [Medline: 33783521]
45. ISO 14155:2020 clinical investigation of medical devices for human subjects — good clinical practice. International Organization for Standardization. 2020 Jul. URL: https://www.iso.org/standard/71690.html [accessed 2022-05-30]
46. FAIR Principles. Go Fair. URL: https://www.go-fair.org/fair-principles/ [accessed 2021-09-21]
47. Facile R, Muhlbradt EE, Gong M, Li Q, Popat V, Pétavy F, et al. Use of Clinical Data Interchange Standards Consortium (CDISC) standards for real-world data: expert perspectives from a qualitative Delphi survey. JMIR Med Inform 2022 Jan 27;10(1):e30363 [FREE Full text] [doi: 10.2196/30363] [Medline: 35084343]
48. Declaration of Taipei: research on health databases, big data and biobanks. World Medical Association. URL: https://www.wma.net/what-we-do/medical-ethics/declaration-of-taipei/ [accessed 2021-09-21]
49. Joint action towards the European Health Data Space - TEHDAS. TEHDAS. URL: https://tehdas.eu/ [accessed 2022-05-04]
50. Big Data Steering Group Workplan 2021-2023. European Medicines Agency. URL: https://www.ema.europa.eu/en/documents/work-programme/workplan-2021-2023-hma/ema-joint-big-data-steering-group_en.pdf [accessed 2022-05-04]
51. Electronic cross-border health services. European Commission. URL: https://ec.europa.eu/health/ehealth/electronic_crossborder_healthservices_en [accessed 2022-05-04]

Abbreviations

EMA: European Medicines Agency
GDPR: General Data Protection Regulation
GPMed: Gesellschaft für Pharmazeutische Medizin
RWD: real-world data
RWE: real-world evidence

Edited by C Lovis; submitted 11.10.21; peer-reviewed by W Van Biesen, X Wu, N Wickramasekera; comments to author 27.04.22; revised version received 16.05.22; accepted 17.05.22; published 17.06.22.

Please cite as:
Klimek P, Baltic D, Brunner M, Degelsegger-Marquez A, Garhöfer G, Gouya-Lechner G, Herzog A, Jilma B, Kähler S, Mikl V, Mraz B, Ostermann H, Röhl C, Scharinger R, Stamm T, Strassnig M, Wirthumer-Hoche C, Pleiner-Duxneuner J
Quality Criteria for Real-world Data in Pharmaceutical Research and Health Care Decision-making: Austrian Expert Consensus
JMIR Med Inform 2022;10(6):e34204
URL: https://medinform.jmir.org/2022/6/e34204
doi:10.2196/34204
PMID:35713954

Original Paper

Perspective of Information Technology Decision Makers on Factors Influencing Adoption and Implementation of Artificial Intelligence Technologies in 40 German Hospitals: Descriptive Analysis

Lina Weinert[1], MSc; Julia Müller[1], MSc; Laura Svensson[1], MSc; Oliver Heinze[1], Dr sc hum

Institute of Medical Informatics, Heidelberg University Hospital, Heidelberg, Germany

Corresponding Author:
Lina Weinert, MSc
Institute of Medical Informatics
Heidelberg University Hospital
Im Neuenheimer Feld 130.3
Heidelberg, 69120
Germany
Phone: 49 622156 ext 34367
Email: lina.weinert@med.uni-heidelberg.de

Abstract

Background: New artificial intelligence (AI) tools are being developed at a high speed. However, strategies and practical experiences surrounding the adoption and implementation of AI in health care are lacking. This is likely because of the high implementation complexity of AI, legacy IT infrastructure, and unclear business cases, thus complicating AI adoption. Research has recently started to identify the factors influencing AI readiness of organizations.

Objective: This study aimed to investigate the factors influencing AI readiness as well as possible barriers to AI adoption and implementation in German hospitals. We also assessed the status quo regarding the dissemination of AI tools in hospitals. We focused on IT decision makers, a seldom studied but highly relevant group.

Methods: We created a web-based survey based on recent AI readiness and implementation literature. Participants were identified through a publicly accessible database and contacted via email or invitational leaflets sent by mail, in some cases accompanied by a telephonic prenotification. The survey responses were analyzed using descriptive statistics.

Results: We contacted 609 possible participants, and our database recorded 40 completed surveys. Most participants agreed or rather agreed with the statement that AI would be relevant in the future, both in Germany (37/40, 93%) and in their own hospital (36/40, 90%). Participants were asked whether their hospitals used or planned to use AI technologies. Of the 40 participants, 26 (65%) answered “yes.” Most AI technologies were used or planned for patient care, followed by biomedical research, administration, and logistics and central purchasing. The most important barriers to AI were lack of resources (staff, knowledge, and financial). Relevant possible opportunities for using AI were increase in efficiency owing to time-saving effects, competitive advantages, and increase in quality of care. Most AI tools in use or in planning have been developed with external partners.

Conclusions: Few tools have been implemented in routine care, and many hospitals do not use or plan to use AI in the future. This can likely be explained by missing or unclear business cases or the need for a modern IT infrastructure to integrate AI tools in a usable manner. These shortcomings complicate decision-making and resource attribution. As most AI technologies already in use were developed in cooperation with external partners, these relationships should be fostered. IT decision makers should assess their hospitals’ readiness for AI individually with a focus on resources. Further research should continue to monitor the dissemination of AI tools and readiness factors to determine whether improvements can be made over time. This monitoring is especially important with regard to government-supported investments in AI technologies that could alleviate financial burdens. Qualitative studies with hospital IT decision makers should be conducted to further explore the reasons for slow AI.

(JMIR Med Inform 2022;10(6):e34678) doi:10.2196/34678

KEYWORDS

artificial intelligence; AI readiness; implementation; decision-making; descriptive analysis; quantitative study

Introduction

Background

In recent years, artificial intelligence (AI) in medicine has gained significant attention, with innovative technologies promising better quality of diagnosis [1-3], treatment [1], advancements in personalized medicine [1,4], and improvements in workflow [5]. Simultaneously, these technologies have the potential to save time and cost [1,6]. The use of AI could free health care workers from repetitive and tedious tasks and enable them to allocate their attention and time more effectively [7]. However, fears surrounding AI in health care persist. Common fears include possible job losses because of automation and negative effects on the patient-physician relationship [2,8,9]. For this study, we used the definition by He et al [10]. They define AI as "a branch of applied computer science wherein computer algorithms are trained to perform tasks typically associated with human intelligence" [10]. There are different relevant subcategories of AI, such as machine learning and deep learning, with different implications for professional users and health care organizations. However, in this study, we focused on the general concept of AI in hospitals.

A recent systematic review by Yin et al [5] demonstrated hesitancy and slow uptake of AI technologies. The authors reported on real-life implementations of AI in health care. Their search retrieved 51 real-life clinical implementations of AI worldwide, with most studies conducted in the United States. The most common applications of AI tools were in the field of decision support. These technologies mainly focus on specific diseases such as sepsis, breast cancer, and diabetic retinopathy [5]. Diverging outcome measures and low-quality studies were prevalent in the review, making it difficult for decision makers to compare and evaluate AI effectiveness, advantages, and disadvantages. Furthermore, they found that outcome evaluation and acceptance measures only included patients and health care workers [5]. Their search strategy retrieved only one paper from Germany, which is in contrast with the German government's AI strategy [11] and recent political efforts to increase the use of AI in hospitals [12]. Hence, we identified a need to investigate the current spread of AI technologies in hospitals and their stage of development as well as AI readiness factors in Germany.

The transfer of new and innovative technologies into practice is usually associated with barriers and requires employees' and institutions' ability to adapt to change [13,14]. Recently, existing frameworks and learnings on the dissemination of innovative technologies have been applied to AI [15]. Three main components can be outlined: (1) adoption, which entails the decision to use an innovation [16]; (2) readiness, encompassing the assessment of the conditions needed to engage in an activity [17]; and (3) implementation, describing an innovation's transfer into practice [15].

Although new AI technologies are being developed at a high speed, strategies and practical experiences surrounding the adoption and implementation of AI in health care are lacking [10,18]. This is partly because of the high implementation complexity of AI, as it is neither easy to use nor easy to deploy [17,19]. Furthermore, AI can be difficult to understand and has been described as a *black box*, meaning a machine with nontransparent workings and inexplicable results of automated algorithms. This has the potential to lower trust and discourage decision makers and users [4,20,21].

Aims

This study presents the first large-scale web-based survey on the current adoption and implementation of AI technologies in randomly selected German hospitals. We further aimed to gain insights into the number, type, and developmental stage of the AI technologies currently in use. In addition to the literature on AI readiness and adoption, we examined the applicability of existing AI readiness factors to the German health care sector.

Methods

Study Design

A quantitative study design was used to obtain a general overview of the situation in Germany. Data were collected using an anonymous web-based questionnaire. We invited chief information officers (CIOs) from randomly selected German hospitals. We identified CIOs as important intermediaries because their position is linked to the clinical implementation of AI as well as to developers, technology companies, and regulatory authorities. Anonymity was ensured throughout the study.

Ethics Approval

The study was approved by the Ethics Committee of Heidelberg University Hospital (S-490/2020). The study was conducted according to the Checklist for Reporting Results of Internet E-Surveys checklist for quantitative research [22].

Instrument Development and Design

After consulting existing literature on AI readiness, implementation, and adoption, the authors conducted a creative brainstorming process to develop preliminary survey items. The preliminary items were compared with existing theoretical frameworks.

Jöhnk et al [15] developed a model that focused on organizational AI readiness. They described AI readiness both as a predecessor and a constant influence on AI adoption and implementation [15]. Jöhnk et al [15] identified 18 organizational readiness factors in 5 categories (strategic alignment, resources, knowledge, culture, and data) and pointed out that these factors continuously foster AI adoption [15]. Awareness of these factors can improve the adoption and implementation outcomes, as a higher level of organizational readiness is believed to increase the success of innovation adoption while lowering the risk of failure [20,23]. For example, knowledge and awareness of AI were shown to be prerequisites for successful AI adoption [15,24,25].

The technological-organizational-environmental framework by DePietro et al [26] describes the adoption, implementation, and use of technology in firms as dependent on the technological, organizational, and the environmental context [27]. Pumplun et al [24] first applied this framework to AI and discussed that challenges to AI readiness can be observed at all of these levels.

Observed technological challenges often stem from data accessibility issues owing to AI's need for extensive databases and adjacent data privacy considerations [24,28]. Environmental challenges include questions about consumer and patient trust in AI, regulatory acceptance, and in some cases, mandated work councils (mandated institutions of nonunion employee representation) [6,24,29,30]. Concerning organizational challenges, a lack of (top) managerial support has been identified as very relevant [17]. A further challenge is the need for highly skilled and trained staff (eg, data scientists, a very sought-after group of professionals) [15,17]. Financial aspects, such as unclear reimbursement processes for health care delivered by AI and liability issues, contribute to hesitancy in AI adoption and implementation [1].

On the basis of these theoretical considerations, LW, JM, and LS refined the survey design and wording of the questions. In the first section, the questionnaire focused on participants' general professional opinions on AI in hospitals to assess the hospital's strategic alignment and their stance in the AI adoption phase. The second section asked participants to state their hospital's use of AI technologies, which helped us gain insight into the dissemination of AI technologies. In the following sections, the survey presents items on known perceived barriers, opportunities, and resources needed for the implementation of AI in hospitals. In addition to these questions, the questionnaire also asked for sociodemographic data of the participants, hospital size, and hospital ownership (private, public, or nonprofit). A translated English version of the survey can be found in Multimedia Appendix 1.

The survey was pretested by 6 researchers from the field of medical informatics, using a cognitive pretesting method [31]. The pretest participants suggested changes in the wording and order of questions. These suggestions were implemented, and the final survey was created.

The final survey did not include any randomized or alternated items. Adaptive questioning was used to reduce the length of the questionnaires. On average, the 10-page questionnaire contained 6.3 items per page. Possible answers were either presented on a 5-point Likert scale or as *yes or no*, with *I don't know* and *prefer not to say* as alternative options. Few questions were asked for further elaboration of answers in open-text formats. Automatic checks for completeness were performed, and participants were required to choose an answer for each question. Cookies were used to assign unique user IDs. Participants were offered the option to return and modify their answers. They were also able to leave the survey and continue it later. IP addresses of participants were neither saved nor checked. REDCap (Research Electronic Data Capture; Vanderbilt University) [32,33] hosted at the Heidelberg University Hospital was used for data collection and management. REDCap is a secure web-based software platform designed to support data capture for research studies [32,33].

Data Collection and Analysis

From a publicly available database of all hospitals in Germany provided by the German Hospital Federation [34], we randomly selected the hospitals we wanted to include in our recruitment process by performing a spreadsheet calculation. We aimed for an equal, realistic representation of hospital size (measured through the number of hospital beds) in each sample. We then checked whether the selected hospitals were actually in operation. Other specific inclusion and exclusion criteria were not applied, as we wanted to depict a realistic reflection of all the hospitals in Germany. In addition to this random selection, we included all academic hospitals in Germany in our recruitment efforts. CIOs and their contact details were manually retrieved from the websites of hospitals. We recruited participants from 609 hospitals in 4 rounds of recruitment. Initially, participants were invited via email to participate in the study. The emails contained a link to access the open survey and information about the study (eg, purpose of the study, length of questionnaire, data protection guidelines, and investigators). As participation in this study was voluntary and anonymous, we regarded survey completion as consent for study participation and data use.

Although all 4 rounds followed the same administrative process, we used additional measures in recruitment rounds 3 and 4 to increase the number of participants. In round 3, we used telephonic prenotifications when an office telephone number was publicly available. In round 4 of recruitment, we designed invitational leaflets that were sent via mail. The leaflets encompassed a short informational text and a QR code, leading to the open survey. For each round, we sent 2 reminders via email. Our survey was not advertised elsewhere, as we wanted to include only members of our specific target group in the sample. No incentives were offered to the study participants.

Data were collected from October 2020 to February 2021. After completion, all data were exported from REDCap to SPSS statistical software (version 27, IBM). All data were checked for plausibility and analyzed by LW. Descriptive analyses were conducted. For open-item responses, recurring keywords and phrases were paraphrased and summarized.

Results

Overview

Our database recorded 50 surveys, of which 10 were terminated early, usually in the first third of the survey. A total of 40 surveys were fully completed and were included in the analysis, resulting in a response rate of 6.6%. Timeframes were analyzed, but no unusual timeframes were observed. No statistical corrections were performed.

Demographic Characteristics

A total of 40 fully completed surveys were included in the analysis. Table 1 provides information on participant characteristics. Most participants were aged between 46 and 55 years (23/40, 58%), and most of the participants were male (33/40, 83%). Of the 40 participants, 26 (65%) said they were CIOs or leaders of the IT department of their institution. Other commonly mentioned professions included IT department employee (7/40, 18%) and research associate (4/40, 1%). Participants stated the ownership of their hospitals as follows: public hospital (30/40, 75%), nonprofit hospital (8/40, 20%), private hospital (2/40, 5%), and hospital with an academic affiliation (15/40, 38%)

Table 1. Participant characteristics (N=40).

Characteristics	Participants
Gender, n (%)	
Female	5 (13)
Male	33 (83)
Prefer not to say	2 (5)
Age group (years), n (%)	
26 to 35	2 (5)
35 to 45	8 (20)
46 to 55	23 (58)
56 to 65	5 (13)
>65	2 (5)
Hospital ownership, n (%)	
Public	30 (75)
Nonprofit	8 (20)
Private	2 (5)
Academic affiliation, n (%)	
Academic	15 (38)
Nonacademic	25 (63)
Number of beds in hospital, n (%)	
1 to 199	3 (8)
200 to 399	5 (13)
400 to 599	7 (18)
600 to 799	4 (10)
>800	21 (52)
Position[a], n (%)	
Chief information officer or head of IT	26 (65)
Chief data officer	1 (3)
Chief medical officer	1 (3)
IT department employee	7 (18)
Research associate	4 (10)
Data scientist	3 (8)
No answer	1 (3)
Other	3 (8)

[a]Selection of multiple items possible.

Participants' Professional Opinions and Assessments

Most participants were either undecided or said they rather disagreed with the statement that AI is relevant for the current health care provision in their hospital and in Germany. However, most participants agreed or rather agreed that AI would be relevant in the future, both in Germany (37/40, 93%) and in their own hospital (36/40, 90%). This fits well with most participants fully agreeing or rather agreeing that AI plays a role in their hospital's strategy (22/40, 55%). On the topic of information about the possible application of AI in hospitals, the participants were more undecided. In all, 13% (5/40) of the participants fully agreed with the statement that they were well informed, and 38% (15/40) of the participants rather agreed that they were well informed. A total of 38% (15/40) of the respondents were undecided, and 13% (5/40) of the respondents said they were rather uninformed. Overall, the participants were rather optimistic about the use of AI technologies in their hospitals. Of the 40 participants, 14 (35%) rather agreed that their hospital was ready for AI, 14 (35%) were undecided, 7 (18%) said they were rather not ready, and only 4 (10%) stated

that their hospital was not ready at all. One participant did not respond to this question.

AI Technologies in Use or in Planning

The next section of the questionnaire focused on AI tools and technologies. In the first subcategory, participants were asked whether their hospital used or planned to use AI technologies. Of the 40 participants, 26 (65%) answered "yes." Through the following questions, participants were asked to describe these technologies in more detail. Most AI technologies were used or planned for patient care, followed by biomedical research, administration, and logistics and central purchasing. Other areas mentioned by the participants in free text were marketing, malware detection, and pathology. Participants were presented with a list of common AI technologies when they answered "yes" to the first question in this subcategory (Multimedia Appendix 1 provides the full list of technologies). For every listed AI technology, they could categorize their hospital's current stance on this technology. The options included the following: in planning, in research and developmental stage, implementation phase, routine care, and not applicable. The most commonly chosen technologies overall were as follows: speech recognition and text analysis systems (20/26, 77%, assigned one of the stances other than *not applicable*), systems for picture recognition (17/26, 65%), and robotics and autonomous systems (17/26, 65%).

Sensorics and communication systems were the least picked (10/26, 38%). Most technologies were in the planning phase.

Concerning the integration of these technologies into the overarching system architecture, 27% (7/26) of the participants stated that technologies in their hospital were integrated, in 23% (6/26) of hospitals, technologies were not integrated but integration was planned, 38% (10/26) were partly integrated, and 12% (3/26) were not integrated. In free text, participants provided reasons for the lack of integration, which included missing interfaces; missing standards for interfaces, processes, and organization; unfavorable cost-benefit relationship; missing evaluation and overall concepts; and immaturity of the AI technology.

In a question allowing for multiple choice, participants stated that some or all AI technologies in their institution were commonly developed with industry partners (23/26, 88%) or university-based research partners (9/26, 35%). Only 12% (3/26) of the participants stated that some or all of their AI technologies were developed within their own institutions.

Barriers to AI Use and Possible Opportunities Associated With AI

The second subcategory included questions about perceived barriers to the use of AI (Table 2). Through a matrix design, we presented the participants with a list of known barriers compiled from the literature. The barrier most participants (36/40, 90%) agreed or partly agreed with was *lacking resources (staff, knowledge, financial)*. Other relevant barriers were *lacking compatibility or interoperability with existing IT infrastructure* (33/40, 83%) and *quality of data* (30/40, 75%). Participants also disagreed or rather disagreed with some of the barriers derived from the literature. Here, the barriers with the least agreement were *leadership acceptance* (4/40, 10%, agreed or rather agreed with the statement) and *patient acceptance* (4/40, 10%). Other barriers with low agreement were *user (eg, physicians and nurses) acceptance* (9/40, 23%) and *corporate culture* (13/40, 33%). In free text, some participants described additional barriers. These contained immaturity of available AI technologies, fear of high expenses in the training and learning phase of AI, and cloud strategies of AI producers.

Table 2. Perceived barriers to implementation and use of artificial intelligence (N=40).

Ranking	Barrier	Total participants in agreement and sample percentages, n (%)[a]
1	Lacking resources (staff, knowledge, and financial)	36 (90)
2	Lacking compatibility or interoperability with existing IT infrastructure	33 (83)
3	Quality of data	30 (75)
4	Availability of data	26 (65)
5	Ethical aspects (eg, liability issues)	24 (60)
6	Product range on the market	23 (58)
7	Data protection	22 (55)
7	Quantity of data	22 (55)
8	Legal regulations	19 (48)
9	Consent of the work council	15 (38)
10	Corporate culture	13 (33)
11	User (eg, physicians, nurses, and administration) acceptance	9 (23)
12	Leadership acceptance	4 (10)
12	Patient acceptance	4 (10)

[a]Responses of "agree" or "rather agree" were grouped together.

In the third subcategory, participants were asked about positive prospects possibly associated with AI (Table 3). Then, they had to state their agreement with these opportunities on a 5-point Likert scale. The opportunity with the highest agreement was *increase in efficiency due to time-saving effects* (29/40, 73% agreed or rather agreed with the statement). Other statements also yielded high agreement rates. The opportunity participants agreed with least was *financial savings*. Only 40% (16/40) of the participants said they agreed or rather agreed with the statement that AI could lead to financial savings in their hospital, whereas 40% (16/40) of the participants disagreed or rather disagreed. Overall, this subcategory yielded homogeneous results. No further opportunities were raised in free text.

A detailed presentation and graphs presenting the results of these 2 subcategories can be found in Multimedia Appendix 2.

Table 3. Perceived opportunities associated with the implementation and use of artificial intelligence (N=40).

Ranking	Opportunity	Total participants in agreement and sample percentages, n (%)[a]
1	Increase in efficiency due to time-saving effects	29 (73)
2	Competitive advantage	27 (69)
3	Increase in quality of care	25 (66)
4	Easing the workload of employees	21 (53)
5	Financial savings	16 (40)

[a]Responses of “agree” or “rather agree” were grouped together.

Resources and Requirements for AI Use in Hospitals

For the fourth subcategory, we focused on the resources required for the use of AI technologies in hospitals. Again, the participants were presented with a list of known critical resources for AI implementation, and they had to indicate their level of agreement with these findings from literature (Table 4). The resource most people needed was *staffing resources* (35/40, 90% agreed or rather agreed with the statement). The resource with the least relevance was *organizational frameworks* (25/40, 64%). As seen in the other subcategories, the distribution of answers was homogeneous. A detailed presentation and graphs presenting the results of this subcategory can be found in Multimedia Appendix 2.

Table 4. Resources needed for use and implementation of artificial intelligence (N=40).

Ranking	Resource	Total participants in agreement and sample percentages, n (%)[a]
1	Staffing resources	35 (90)
2	Time	34 (87)
3	Knowledge	33 (85)
4	Financial resources	32 (84)
5	Technical resources	31 (79)
6	Data base	27 (69)
7	Organizational frameworks	25 (64)

[a]Responses of “agree” or “rather agree” were grouped together.

The next item asked participants whether their hospital needed to fulfill any further requirements or resources besides those already mentioned in a yes or no format. A total of 60% (24/40) of the participants answered “yes” and provided explanations in free text. Here, organizational aspects were most common (eg, competencies and responsibilities), followed by workflow and legal issues. Technical aspects were described in detail, such as lacking hardware and software, interoperability, difficulties with data transfer from old to new systems, need for additional modules for data capture, and Wi-Fi availability and speed.

Considering the tech industry and its offerings on the market, the participants were highly undecided. Furthermore, 58% (23/40) of the participants said that they did not know if the supply met the demand for AI technologies in their hospital. Only 7% (3/40) of the participants stated that offerings on the market were sufficient.

Discussion

Principal Findings

This study provided insights into the current and planned dissemination of AI tools as well as perceived barriers and opportunities for the implementation and adoption of AI tools in 40 hospitals in Germany. We designed a web-based survey based on existing literature on the implementation of AI in hospitals. Our participants were mainly from an IT background, with 28 decision makers in leadership positions. Two-thirds of the participants said that they used or planned to use AI tools in their institution. Speech recognition and text analysis systems, systems for picture recognition, and robotics and autonomous systems were the tools or systems most commonly used, or their use was planned. We did not find differing opinions among hospitals of different sizes or ownership. The results showed

that most participants recognized the implementation of AI in hospitals as a relevant, forthcoming part of their IT strategy. However, lack of resources and compatibility or interoperability with the existing IT infrastructure were identified as barriers to implementation. Staffing resources, time, knowledge, financial resources, and technical resources required for the implementation of AI were all highly relevant resources. A possible increase in efficiency because of time-saving effects, competitive advantage, and increase in quality of care was seen as the most important opportunity associated with AI use. We conclude that AI readiness factors derived from the literature are applicable to the hospital context in Germany. The following discussion highlights the most relevant barriers to AI readiness, adoption, and implementation while also presenting possible ways to overcome these barriers.

AI in Hospital Strategies

AI readiness as a concept has been described recently [15,24]. *Strategic alignment* was identified as 1 of 5 key aspects of organizational AI readiness. Our survey included a question addressing whether AI was a part of the participants' hospital IT strategy. To this question, 55% (22/40) of the participants agreed or rather agreed that AI was a part of their strategy. In addition, most participants agreed or rather agreed that AI would be relevant in the future, both in Germany (37/40, 93%) and in their own hospital (36/40, 90%). However, this also means that there are decision makers who recognize the relevance of AI in the future but do not consider it a part of their hospitals' strategy. First, this could be because of the complexity of AI implementation (eg, uncertainties surrounding the workings of the technology, acceptance of the technology, and an unclear regulatory situation) [1,10,17,19,29,35]. Second, the hesitancy to include AI in a hospital's IT strategy could be explained by high costs and unclear reimbursement schemes [1]. In our study, 80% (32/40) of the participants agreed that their institution lacked financial resources, and 90% (36/40) said that a lack of resources overall was a barrier for AI implementation. At the same time, only 40% (16/40) of the participants agreed with the statement that AI holds a potential for financial savings. This paints a picture of AI as a resource-intensive technology with limited financial rewards. To overcome this barrier and compensate for the financial burden because of investments in digital technologies, the German government recently introduced a new law, the *hospital future act* (ie, Krankenhauszukunftsgesetz). Through this law, hospitals trying to implement digital technologies, including decision support systems, can apply for financial support to facilitate necessary acquisitions [12]. The law went into effect during our data collection period; thus, we cannot report on the possible impacts of this law. However, as the financial aspects were reported as a relevant barrier in our study, it could be of interest for future research to evaluate the effects of the new law.

Although there are both expectations and observations of AI as a possible tool to save cost and generate high revenue [1,6,29], for example, through higher efficiency, high-quality evidence analyzing the cost and benefits of AI implementation in hospitals is missing [7]. Hence, decision makers lack evidence and information, and the business case for AI in hospitals remains unclear [10,29], which in turn inhibits organizational AI readiness [15].

AI Acceptability

With regard to further barriers to the implementation of AI, *soft* factors such as user, patient, and leadership acceptance were seen as less relevant barriers by the participants in our survey. This impression might be caused by limited contact of IT department members with users, leadership, and especially with patients. Acceptance issues might also become more obvious to decision makers over time, as most participants in our study had not yet implemented AI in their hospital. Nonetheless, it is important to consider the evidence that acceptability is a relevant antecedent of AI adoption and implementation. For example, a paper reviewing 9 studies on the acceptance of AI in health care concluded that consumers have a robust reluctance toward medical care delivered by AI compared with human providers [36]. In another study, only 3% of patients found that the possible negative aspects of AI outweighed the potential benefits [37]. Overall, there is mixed evidence regarding patients' acceptance of AI in the medical context, and further research is needed [5].

Leadership acceptance and support have been identified as important antecedents for AI implementation [15,24]. The acceptance of AI users, such as physicians and hospital employees, has also been identified as relevant in other studies [9,38]. Following the technology acceptance model, perceived ease of use and usefulness can positively affect favorable attitudes toward a new technology, which in turn improves its acceptability and use [13,39]. Hence, special attention should be paid to these aspects when deciding on acquiring and implementing new AI tools in a hospital.

Finally, the issue of AI acceptability can be addressed by investing in the concept of *explainable AI*, meaning a more transparent, understandable AI with high performance levels [40]. Although little evidence exists, it is reasonable to expect that this new approach could increase AI acceptability by increasing understanding and trust in the new technology [13,40-42]. IT decision makers should not underestimate the issue of AI acceptability and should take the fears and perceptions surrounding AI seriously when planning to implement new AI technology.

Possible Mismatch in Supply and Demand

Another finding in our study was that only 7% (3/40) of the participants said that the supply of applicable AI solutions to the tech market was sufficient for their needs. Another 58% (23/40) of participants reported that they were unsure. One reason could be that we did not reach the right people in the institution, and they were thus unable to assess the tech market. Another possibility could be that our participants did not spend time researching the offerings in the tech market. This could be especially true for those who are not using or planning to use AI tools. However, it could also be possible that the offerings on the market do not fit the requirements of their potential clients. This result could be of value for tech companies trying to reach decision makers in hospitals. This finding is especially important considering that only 12% of the AI tools were

developed within the hospitals in our survey. Hence, partnerships for the development of AI tools are common and must be fostered.

Generalizability

We created this survey instrument based on an extensive literature research and theoretical frameworks and used cognitive pretesting to ensure understandability. Participants usually completed the survey in <10 minutes. Hence, our survey instrument enabled us to collect data both efficiently and in a theoretically informed manner. This survey could serve as a template for other studies, especially in countries with a similar level of dissemination of AI technologies. Country-specific items, such as the work council, should be adapted to the context in question. Although our survey included these country-specific aspects, they did not appear to be of high relevance in our sample. However, we think that these aspects should be surveyed, as their importance in other contexts is not predictable.

Strengths and Limitations

This study investigated the status quo of AI technologies in 40 German hospitals and the applicability of AI readiness factors derived from the literature. Owing to the low response rate and resulting small sample size, our results are not representative but describe a first impression. We surveyed hospital CIOs, a group we identified as important intermediaries for digital innovation adoption and implementation. While other studies about the perceptions, barriers, and issues surrounding AI questioned users (eg, physicians and health professionals), patients, or other stakeholders [37,43-45], we focused on the seldom studied group of IT decision makers. Although focusing exclusively on one stakeholder group may introduce a bias, we believe that the focus on this seldom studied group makes our study unique and relevant, thus warranting the risk of bias. The presented perspective of hospital CIOs depicts barriers to AI use and acceptance on a decision or leadership level. Our results can further the holistic discussion about the real-world implementation of AI and AI readiness.

We analyzed the differences in opinions of hospitals differing in size and ownership, which did not produce relevant results. This finding should be interpreted cautiously, as our sample size could be too small to produce significant results.

Owing to technical limitations, we were unable to report the number of unique site visitors. This impedes the calculation of correct survey response rates. Although we used various recruitment methods (emails, letters, and telephone calls) over a prolonged period, our sample size remained small compared with the number of hospitals in Germany (1914 hospitals in 2019 [35]). The small number of respondents may be explained by a general lack of interest in the survey's topic [46], time constraints because of the COVID-19 pandemic, or because of the requirements of a leadership position and by a hesitancy to click on links sent via email owing to fear of security breaches. We tried to reduce this fear by establishing an *offline* contact with possible participants (letters and telephonic prenotification), but the effect is unclear. At the same time, people who chose to participate in the survey might have a stronger interest or profound experience with AI. We tried to minimize this effect by pointing out in invitations that no knowledge or experience with AI is necessary for participation. However, there was a risk of nonresponse bias in our study.

Considering the demographics of the survey respondents, the sample was very homogeneous, as most participants were middle-aged and male. This distribution was expected and represents the composition of IT departments in Germany [47]. As we included all academic hospitals in our recruitment efforts, larger hospitals were overrepresented in our sample. We expected academic hospitals to be more involved in AI research and thus wanted to invite them to participate in our study. In addition, very small hospitals sometimes do not have a CIO position or outsource their IT services. In such situations, it is possible that our survey invitations did not reach the right person.

As AI is a new and complex technology, it is possible that our participants misunderstood some questions or falsely claimed that they had used AI in their hospital. We managed this risk by closely aligning our survey design with the results from the 6 pretests. Pretest participants suggested not to include a general definition of AI but to give examples for the specific tools in question 2 ("Please assess the current stage of implementation of these AI tools in your hospital"). To keep the survey as short as possible and by keeping in mind that our target group consisted of experts in a related field, we followed this suggestion. However, this risk must be considered when comparing our results.

Conclusions

This study paints a mixed picture of the status quo of AI in German hospitals. In our sample, few tools have been implemented in routine care, and many hospitals do not use or plan to use AI in the future. This can likely be explained by missing or unclear business cases, which complicates decision-making and resource attribution. We also observed a mismatch or lack of information about AI offerings in the tech market. This is another important aspect to be monitored, as most AI technologies that are already in use were developed in cooperation with external partners. Therefore, these relationships should be fostered. IT decision makers in hospitals should assess their hospitals' readiness for AI individually with a focus on resources. Further research should continue to monitor the dissemination of AI tools and AI readiness factors to determine whether improvements can be made over time, especially with regard to government-supported investments in AI technologies that could alleviate the financial burden. Qualitative studies with hospital IT decision makers should be conducted to explore the reasons for slow AI adoption in more detail. The results of our study may infer that AI adoption is not only a topic solely for the IT department but also for the whole hospital as an enterprise, including management, medical staff, and business in terms of an important building block of the digital transformation.

Acknowledgments

The authors would like to thank all study participants for their contributions to this study. The authors would also like to thank Carolin Anders (Heidelberg University Hospital, Institute of Medical Informatics) for her support. This study was funded by the Baden-Wuerttemberg (Germany) Ministry of Science, Research and the Arts, under reference number 42-04HV.MED(19)/15/1 as part of the project ZIV (*Zentrum fuer Innovative Versorgung*).

Authors' Contributions

LW drafted and prepared the original manuscript. OH was the principal investigator of the study. LW and JM were responsible for study design and protocol. All authors contributed to the concept and design of the study and preparation of the manuscript. LW, JM, and LS constructed and tested the survey design and the quantitative data collection tool. LW analyzed survey data. LW interpreted and phrased the results of the quantitative data. All the authors provided substantial comments and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Translated survey.
[DOCX File , 47 KB - medinform_v10i6e34678_app1.docx]

Multimedia Appendix 2
Bar graphs.
[DOCX File , 64 KB - medinform_v10i6e34678_app2.docx]

References

1. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr Cardiol Rep 2014 Jan;16(1):441. [doi: 10.1007/s11886-013-0441-8] [Medline: 24338557]
2. Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. NPJ Digit Med 2018;1:5 [FREE Full text] [doi: 10.1038/s41746-017-0012-2] [Medline: 31304291]
3. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, Reader study level-Ilevel-II Groups, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018 Aug 01;29(8):1836-1842 [FREE Full text] [doi: 10.1093/annonc/mdy166] [Medline: 29846502]
4. Fröhlich H, Balling R, Beerenwinkel N, Kohlbacher O, Kumar S, Lengauer T, et al. From hype to reality: data science enabling personalized medicine. BMC Med 2018 Aug 27;16(1):150 [FREE Full text] [doi: 10.1186/s12916-018-1122-7] [Medline: 30145981]
5. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. J Med Internet Res 2021 Apr 22;23(4):e25759 [FREE Full text] [doi: 10.2196/25759] [Medline: 33885365]
6. Agarwal Y, Jain M, Sinha S, Dhir S. Delivering high‐tech, AI‐based health care at Apollo Hospitals. Glob Bus Organ Excell 2020;39(2):20-30. [doi: 10.1002/joe.21981]
7. Wolff J, Pauling J, Keck A, Baumbach J. The economic impact of artificial intelligence in health care: systematic review. J Med Internet Res 2020 Feb 20;22(2):e16866 [FREE Full text] [doi: 10.2196/16866] [Medline: 32130134]
8. Rubeis G. The disruptive power of Artificial Intelligence. Ethical aspects of gerontechnology in elderly care. Arch Gerontol Geriatr 2020 Jul 15;91:104186. [doi: 10.1016/j.archger.2020.104186] [Medline: 32688106]
9. Castagno S, Khalifa M. Perceptions of artificial intelligence among healthcare staff: a qualitative survey study. Front Artif Intell 2020;3:578983 [FREE Full text] [doi: 10.3389/frai.2020.578983] [Medline: 33733219]
10. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36 [FREE Full text] [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
11. Strategie Künstliche Intelligenz der Bundesregierung. Nationale Strategie für Künstliche Intelligenz. Berlin: Die Bundesregierung; 2018 Nov. URL: https://www.bundesregierung.de/resource/blob/975226/1550276/3f7d3c41c6e05695741273e78b8039f2/2018-11-15-ki-strategie-data.pdf?download=1 [accessed 2022-05-27]
12. Riedel W, Riedel H. Krankenhauszukunftsgesetz: Die große Digitalisierungsoffensive. kma - Klinik Management aktuell 2020;25(12):55-57. [doi: 10.1055/s-0040-1722470]
13. Romero-Brufau S, Wyatt KD, Boyum P, Mickelson M, Moore M, Cognetta-Rieke C. A lesson in implementation: a pre-post study of providers' experience with artificial intelligence-based clinical decision support. Int J Med Inform 2020 May;137:104072. [doi: 10.1016/j.ijmedinf.2019.104072] [Medline: 32200295]

14. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018 Dec 04;320(21):2199-2200. [doi: 10.1001/jama.2018.17163] [Medline: 30398550]
15. Jöhnk J, Weißert M, Wyrtki K. Ready or not, AI comes— an interview study of organizational AI readiness factors. Bus Inf Syst Eng 2020 Dec 22;63(1):5-20. [doi: 10.1007/s12599-020-00676-7]
16. Frambach RT, Schillewaert N. Organizational innovation adoption: a multi-level framework of determinants and opportunities for future research. J Bus Res 2002 Feb;55(2):163-176 [FREE Full text] [doi: 10.1016/S0148-2963(00)00152-1]
17. Lokuge S, Sedera D, Grover V, Dongming X. Organizational readiness for digital innovation: development and empirical calibration of a construct. Inf Manag 2019 Apr;56(3):445-461. [doi: 10.1016/j.im.2018.09.001]
18. Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. NPJ Digit Med 2019;2:77 [FREE Full text] [doi: 10.1038/s41746-019-0155-4] [Medline: 31453372]
19. Gallivan MJ. Organizational adoption and assimilation of complex technological innovations: development and application of a new framework. Data Base Adv Inf Syst 2001 Jul;32(3):51-85. [doi: 10.1145/506724.506729]
20. Weiner BJ. A theory of organizational readiness for change. Implement Sci 2009 Oct 19;4:67 [FREE Full text] [doi: 10.1186/1748-5908-4-67] [Medline: 19840381]
21. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc 2020 Mar 01;27(3):491-497 [FREE Full text] [doi: 10.1093/jamia/ocz192] [Medline: 31682262]
22. Eysenbach G. Improving the quality of Web surveys: the checklist for reporting results of Internet E-surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
23. Snyder-Halpern R. Indicators of organizational readiness for clinical information technology/systems innovation: a Delphi study. Int J Med Inform 2001 Oct;63(3):179-204. [doi: 10.1016/s1386-5056(01)00179-4] [Medline: 11502432]
24. Pumplun L, Tauchert C, Heidt M. A new organizational chassis for artificial intelligence - exploring organizational readiness factors. In: Proceedings of the 27th European Conference on Information Systems. 2019 Presented at: ECIS '19; June 8-14, 2019; Stockholm, Sweden p. 8-14 URL: https://aisel.aisnet.org/ecis2019_rp/106
25. Alsheibani S, Cheung Y, Messom C. Artificial intelligence adoption: AI-readiness at firm-level. In: Proceedings of 2018 Pacific Asia Conference in Information Systems. 2018 Presented at: PACIS '18; June 26-30, 2018; Yokohama, Japan URL: https://researchmgt.monash.edu/ws/portalfiles/portal/273209396/254798983_oa.pdf
26. DePietro R, Wiarda E, Fleisher M. The processes of technological innovation. In: Tornatzky LG, Fleisher M, editors. The Context for Change: Organization, Technology and Environment. Washington, DC, USA: Lexington Books; 1990:151-175.
27. Zhu K, Kraemer KL. Post-adoption variations in usage and value of e-business by organizations: cross-country evidence from the retail industry. Inf Syst Res 2005 Mar;16(1):61-84. [doi: 10.1287/isre.1050.0045]
28. Cypko M, Emmert MY, Falk V, Meyer A. [Artificial intelligence in cardiac surgery]. Chirurg 2020 Mar;91(3):235-239. [doi: 10.1007/s00104-020-01132-8] [Medline: 32144448]
29. Alsheiabni S, Cheung Y, Messom C. Factors inhibiting the adoption of artificial intelligence at organizational-level: a preliminary investigation. In: Proceedings of the 2019 Americas Conference on Information Systems. 2019 Presented at: AMCIS '19; August 15-17, 2019; Cancun, Mexico URL: https://researchmgt.monash.edu/ws/portalfiles/portal/287736273/287674072_oa.pdf
30. Jirjahn U, Smith SS. Nonunion employee representation: theory and the German experience with mandated works councils. Ann Public Cooperative Econ 2018 Jan 31;89(1):201-233. [doi: 10.1111/apce.12191]
31. Lenzner T, Neuert C, Otto W. Cognitive Pretesting (Version 2). GESIS Survey Guidelines. GESIS - Leibniz Institute for the Social Sciences. 2016. URL: https://www.ssoar.info/ssoar/bitstream/handle/document/56369/ssoar-2016-lenzner_et_al-Cognitive_Pretesting_Version_20.pdf?sequence=3 [accessed 2022-05-27]
32. 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]
33. 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]
34. Mission and Objectives. The German Hospital Federation. Berlin: The German Hospital Federation URL: https://www.dkgev.de/englisch/the-german-hospital-federation/mission-and-objectives/ [accessed 2022-05-27]
35. Medical facilities, hospital beds and movement of patient. Statistisches Bundesamt. 2021. URL: https://www.destatis.de/EN/Themes/Society-Environment/Health/Hospitals/Tables/gd-hospitals-years.html [accessed 2022-05-27]
36. Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. J Consum Res 2019 Dec;46(4):629-650. [doi: 10.1093/jcr/ucz013]
37. Tran VT, Riveros C, Ravaud P. Patients' views of wearable devices and AI in healthcare: findings from the ComPaRe e-cohort. NPJ Digit Med 2019;2:53 [FREE Full text] [doi: 10.1038/s41746-019-0132-y] [Medline: 31304399]
38. Strohm L, Hehakaya C, Ranschaert ER, Boon WP, Moors EH. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol 2020 Oct;30(10):5525-5532 [FREE Full text] [doi: 10.1007/s00330-020-06946-y] [Medline: 32458173]

39. Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010 Feb;43(1):159-172 [FREE Full text] [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
40. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 2018 Sep 17;6:52138-52160 [FREE Full text] [doi: 10.1109/access.2018.2870052]
41. Reddy S. Use of artificial intelligence in healthcare delivery. In: Heston TF, editor. eHealth - Making Health Care Smarter. London, UK: IntechOpen; 2018. [doi: 10.5772/intechopen.74714]
42. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195 [FREE Full text] [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
43. Laï MC, Brian M, Mamzer MF. Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France. J Transl Med 2020 Jan 09;18(1):14 [FREE Full text] [doi: 10.1186/s12967-019-02204-y] [Medline: 31918710]
44. Ongena YP, Haan M, Yakar D, Kwee TC. Patients' views on the implementation of artificial intelligence in radiology: development and validation of a standardized questionnaire. Eur Radiol 2020 Feb;30(2):1033-1040 [FREE Full text] [doi: 10.1007/s00330-019-06486-0] [Medline: 31705254]
45. Pinto Dos Santos D, Giese D, Brodehl S, Chon SH, Staab W, Kleinert R, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
46. Keusch F. Why do people participate in Web surveys? Applying survey participation theory to Internet survey data collection. Manag Rev Q 2015 Jan 9;65(3):183-216. [doi: 10.1007/s11301-014-0111-y]
47. Weitzel T, Eckhardt A, Laumer S, Maier C, Oelhorn C, Wirth J, et al. Women in IT: Eine empirische Untersuchung mit den Top 1000 Unternehmen aus Deutschland sowie den Top 300 Unternehmen aus den Branchen Finanzdienstleistung, Health Care und IT. Recruiting Trends. 2017. URL: https://www.uni-bamberg.de/fileadmin/uni/fakultaeten/wiai_lehrstuehle/isdl/4_Women_in_IT_20170210_WEB.pdf [accessed 2022-05-27]

Abbreviations

AI: artificial intelligence
CIO: chief information officer
REDCap: Research Electronic Data Capture

Edited by C Lovis; submitted 04.11.21; peer-reviewed by D Pinto dos Santos, K Wyatt, A Joseph, M Sedlmayr; comments to author 02.01.22; revised version received 15.02.22; accepted 11.03.22; published 15.06.22.

Please cite as:
Weinert L, Müller J, Svensson L, Heinze O
Perspective of Information Technology Decision Makers on Factors Influencing Adoption and Implementation of Artificial Intelligence Technologies in 40 German Hospitals: Descriptive Analysis
JMIR Med Inform 2022;10(6):e34678
URL: https://medinform.jmir.org/2022/6/e34678
doi:10.2196/34678
PMID:35704378

Original Paper

Prevalence of Sensitive Terms in Clinical Notes Using Natural Language Processing Techniques: Observational Study

Jennifer Lee[1,2*], MD; Samuel Yang[1,2*], MD; Cynthia Holland-Hall[1,2], MD; Emre Sezgin[1], PhD; Manjot Gill[2], MD; Simon Linwood[1], MD, MSc; Yungui Huang[1], PhD; Jeffrey Hoffman[1,2], MD

[1]Nationwide Children's Hospital, Columbus, OH, United States

[2]The Ohio State University College of Medicine, Columbus, OH, United States

[*]these authors contributed equally

Corresponding Author:
Jennifer Lee, MD
Nationwide Children's Hospital
700 Children's Drive
Columbus, OH, 43205
United States
Phone: 1 614 722 2000
Email: jennifer.lee2@nationwidechildrens.org

Abstract

Background: With the increased sharing of electronic health information as required by the US 21st Century Cures Act, there is an increased risk of breaching patient, parent, or guardian confidentiality. The prevalence of sensitive terms in clinical notes is not known.

Objective: The aim of this study is to define sensitive terms that represent the documentation of content that may be private and determine the prevalence and characteristics of provider notes that contain sensitive terms.

Methods: Using keyword expansion, we defined a list of 781 sensitive terms. We searched all provider history and physical, progress, consult, and discharge summary notes for patients aged 0-21 years written between January 1, 2019, and December 31, 2019, for a direct string match of sensitive terms. We calculated the prevalence of notes with sensitive terms and characterized clinical encounters and patient characteristics.

Results: Sensitive terms were present in notes from every clinical context in all pediatric ages. Terms related to the mental health category were most used overall (254,975/1,338,297, 19.5%), but terms related to substance abuse and reproductive health were most common in patients aged 0-3 years. History and physical notes (19,854/34,771, 57.1%) and ambulatory progress notes (265,302/563,273, 47.1%) were most likely to include sensitive terms. The highest prevalence of notes with sensitive terms was found in pain management (950/1112, 85.4%) and child abuse (1092/1282, 85.2%) clinics.

Conclusions: Notes containing sensitive terms are not limited to adolescent patients, specific note types, or certain specialties. Recognition of sensitive terms across all ages and clinical settings complicates efforts to protect patient and caregiver privacy in the era of information-blocking regulations.

(JMIR Med Inform 2022;10(6):e38482) doi:10.2196/38482

KEYWORDS

adolescent; child; privacy; patient portals; natural language processing; eHealth

Introduction

With the increased sharing of electronic health information (EHI) through patient portals as the result of the US 21st Century Cures Act information blocking regulations, there is an increased risk of sharing sensitive information with the wrong person [1]. For pediatric patients and their parents or guardians, there are two major types of risk. The first is disclosure of sensitive information to the patient, which a parent or guardian wants to remain private. In a recent position statement, the Society for Adolescent Health and Medicine supports the parent's right to withhold "certain family information" such as HIV status, substance use disorders, or consanguinity with the child [2]. The second type of risk is disclosure of sensitive information that the child or patient desires (and may be legally entitled) to withhold from a parent or guardian such as documentation of

certain types of reproductive health care, mental health care, or substance abuse treatment [3-5].

Institutions rely on providers to manually flag notes that contain sensitive information [6,7]. At one pediatric institution, Parsons et al [7] manually reviewed notes flagged as sensitive (which accounted for only 2.3% of the total note volume) and found that 16% of them did not have discernable sensitive information. This aligns with the findings of prior work that providers often do not have awareness of the relevant adolescent consent laws in their state [8].

The percentage of notes flagged as containing sensitive information should be higher than indicated in the study conducted by Parsons et al [7]. In newborn patient notes, it is routine practice to document intrauterine drug use or exposure to infectious diseases such as maternal HIV. Bright Futures Guidelines from the American Academy of Pediatrics recommends providers perform psychosocial screening on patients of all ages, and substance use and sexual health screening in all adolescent patients [9,10]. In a recent survey of 3533 high school–aged adolescents, 71% confirmed that providers interviewed them without a parent present [11], in which case it should lead to the documentation of that private interview. Because current electronic health record (EHR) systems are not able to automatically identify sensitive information in clinical notes, the overall prevalence of sensitive information documented in pediatric clinical notes cannot be easily ascertained. The aim of this study is to define a keyword set of sensitive terms and characterize the prevalence of provider notes that contain sensitive terms across different clinical settings.

Methods

Setting

This study was a single-center retrospective review of provider notes in patients aged 0-21 years from January 1, 2019, to December 31, 2019, at an urban, academic, not-for-profit, freestanding children's hospital with over 50 subspecialties and 1.6 million patient visits annually. The note types included were history and physical (H&P) notes, progress notes (inpatient, emergency care, and ambulatory), consultation notes, and discharge summaries authored by physicians (residents, fellows, and attendings) and advanced practice providers (nurse practitioners and physician assistants) documented in the local Epic EHR (Epic Systems Corporation).

Study Design

We used natural language processing (NLP) term expansion to create a representative list of sensitive terms or phrases in the following four categories of sensitive information as determined by local experts: substance use, mental health, reproductive health, and home environment. These categories were created based on prior work to categorize confidential content to incorporate the most common types of sensitive content that may warrant protection from disclosure [7]. Mental health and substance use disorder records are subject to additional Health Insurance Portability and Accountability Act privacy protections [12,13]. Similarly, adolescents may consent to several elements of their own reproductive health care without parental involvement (though specific elements of care vary by state) [14-17]. Home environment includes topics the disclosure of which may place a child in danger in the home, such as parental discord or domestic violence [18]. Similar term expansion methods have been used before, such as for identification of smoking status in free-text data [19,20].

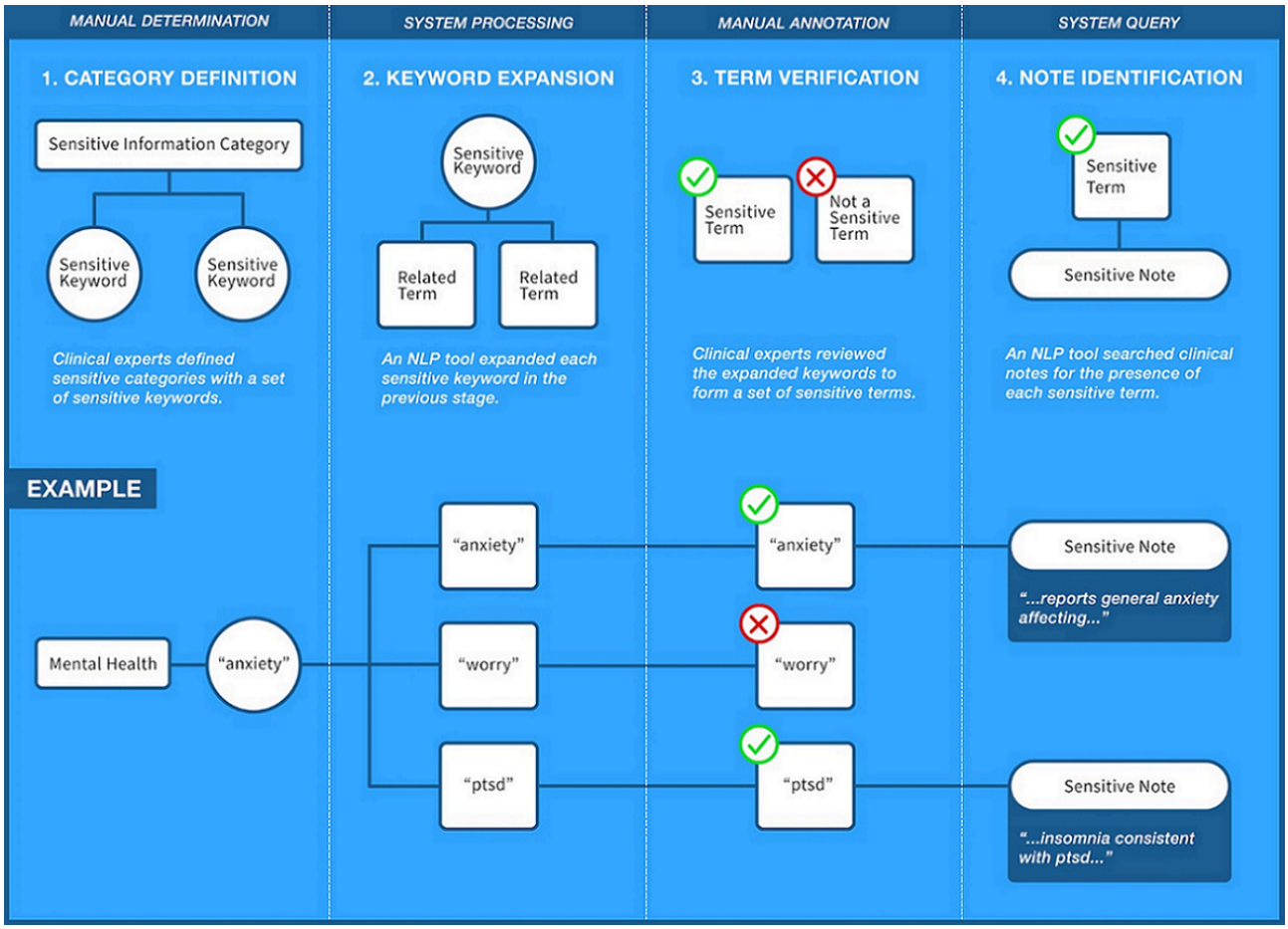
For each category, subject matter experts (SY, CH, and JL) identified 5 to 10 representative terms. We then employed a locally developed NLP tool dubbed DeepSuggest for term expansion. DeepSuggest was trained on approximately 93 million clinical notes in the EHR data set. For each term, DeepSuggest identified 60 additional potentially related terms or phrases, as well as common abbreviations and misspellings [21]. For example, for the term "alcohol," DeepSuggest produced related terms such as "etoh" (abbreviation), "drug" (related term), and "alchol" (misspelling).

Two subject matter experts (JL and CH) manually annotated each term provided by DeepSuggest as either "sensitive" or "not sensitive" (Figure 1). A term was defined as "sensitive" if its presence in a clinical note could indicate documentation of a topic that might reflect sensitive information. For example, as shown in Figure 1, the initial term "anxiety" is expanded by DeepSuggest to include terms such as "worry" and "ptsd." "Worry" is related to "anxiety" but was deemed not likely to represent sensitive content and was thus not included in the final vocabulary. Disagreements between the 2 subject matter experts was resolved through discussion.

We used direct string matching to query for the presence of any sensitive term or phrase in the selected note types. Notes documented in the EHR could contain free text, dictated or transcribed text, templated text, or dynamic links that insert discrete data from elsewhere in the EHR (eg, family history, tobacco use screening, or problem list). However, because the notes are saved as plain text, we were able to use direct string matching to screen for the presence of a term or phrase regardless of how each portion of the note was populated. Moreover, as exact string matching was used, manual review to confirm the accuracy or recall of the search parameters was unnecessary. Parsons et al [7] included the presence of psychiatric or substance use screening questions regardless of positive or negative status as confidential information. Similarly, we determined the presence of a sensitive term, regardless of negation status, as sensitive information. Figure 1 describes the study design.

Clinical notes identified by the search were stratified by note type, author type, and patient age. Because most patient encounters were ambulatory encounters, these notes were also stratified by specialty.

Figure 1. Combining natural language processing (NLP) and expert definitions to identify sensitive notes in the electronic health record. Overview of sensitive term identification protocol. Four categories and sensitive keywords representative for each category were identified by local subject matter experts. A natural language processing tool trained on the entire cohort of notes at the organization was used for keyword expansion. Each sensitive keyword was expanded to 60 potentially related terms. Each related term was manually annotated as a "sensitive" or "not sensitive" term by board-certified pediatricians and adolescent medicine specialists. Exact string word matching was used to determine if a sensitive term was documented in a clinical note; ptsd: posttraumatic stress disorder.



Analysis

Descriptive analysis was performed for all notes by patient cohort, encounter type, and author provider type. We identified the top 10 frequently occurring sensitive terms by category and compared the prevalence of clinical notes with sensitive terms among note types by age using the Fisher exact test ($P<.05$). We used the Cohen kappa to quantify interrater agreement for sensitive term identification. We then determined the prevalence of clinical notes with sensitive terms written in the ambulatory setting and compared them by clinical specialty, also using the Fisher exact test ($P<.05$).

Ethics Approval

The study was approved by the Nationwide Children's Hospital Institutional Review Board (STUDY00000611).

Results

In the study period, there were 763,133 clinical encounters among 279,737 unique patients. In total, 70.3% (536,201/763,133) of the encounters occurred in an ambulatory setting; 20.7% (70,378/763,133) of the patients were 13 years or older. Most patients were White (151,988/279,737, 54.3%), and there was a slight male predominance (142,539/279,737, 51.0% male vs 137,180/279,737, 49.0% female). During the study period, a total of 1,338,297 notes were written by 2342 unique providers, with 501,762/1,338,297 (37.5%) notes containing at least one sensitive term (Table 1).

Table 1. Patient, encounter, and provider characteristics.

Populations and characteristics	Values, n (%)
Patients (n=279,737)	
Age (years)	
Less than 13	209,359 (74.84)
13 to 18	59,415 (21.23)
18 to 21	10,963 (3.92)
Legal sex	
Male	142,539 (50.95)
Female	137,180 (49.04)
Unknown	18 (0.01)
Race	
White	151,988 (54.33)
Black or African American	66,995 (23.95)
Latino or Hispanic	19,053 (6.81)
Other or unknown	41,701 (14.91)
Encounters	763,133
Ambulatory care	536,201 (70.3)
Emergency care	188,204 (24.7)
Inpatient care	38,728 (5.1)
Providers (n=2342)	
Resident	888 (37.92)
Attending	828 (35.35)
Fellow	393 (16.78)
Advanced practice provider	233 (9.94)
Notes (n=1,338,297)	
Notes with sensitive terms	501,762 (37.49)

DeepSuggest expanded 27 sensitive keywords to 1620 new candidate terms. Of those 1620 terms, 478 (30%) were duplicates; 781 (68%) of the 1142 unique candidate terms were determined to be sensitive with an interrater reliability (kappa score) of 0.944. Of the 781 sensitive terms, 698 (89%) were found in the study period (supplemental Table for list of initial keywords and full list of terms are presented in Multimedia Appendix 1).

"Anxiety" was the most frequent sensitive term, with 418,766 total occurrences among 143,968 notes. "Depression" had fewer mentions than anxiety, occurring in 150,934 notes. Abbreviations such as "thc" (tetrahydrocannabinol), "cps" (child protective services), "si" (suicidal ideation), and "hiv" (human immunodeficiency virus) were among the terms most commonly found in the notes. Table 2 describes the 10 most frequent terms in each of the 4 categories.

Table 2. Most frequently used sensitive terms by category.

Category and term	Term frequency, n	Note frequency, n
Substance use		
tobacco	190,547	119,764
alcohol	143,945	107,871
substance	101,183	78,997
smoker	51,572	50,538
cigarettes	36,444	35,443
substance abuse	28,970	23,131
thc[a]	21,216	14,153
marijuana	16,625	10,985
smoked	14,508	14,271
cocaine	13,747	8618
Mental health		
anxiety	418,766	143,968
depression	270,661	150,934
mood	267,706	122,293
suicidal	224,989	72,709
suicidal ideation	140,918	57,057
suicide	109,123	46,463
si[b]	66,977	35,713
panic	52,040	32,025
bipolar	46,729	35,539
depressive	41,511	26,025
Reproductive health		
sexual	238,310	84,710
pregnancy	118,872	77,337
hiv	80,306	56,072
partner	62,456	33,155
sexually	44,491	33,902
sexually active	37,149	29,817
sexual abuse	36,000	16,030
sti[c]	33,904	22,679
sex	30,612	21,714
partners	23,406	13,461
Home environment		
abuse	156,957	70,712
food insecurity	21,108	14,848
bullying	17,259	10,712
conflict	14,997	9657
cps[d]	13,962	9081
weapons	12,016	9485
abuse or neglect	11,671	6212

Category and term	Term frequency, n	Note frequency, n
emotional abuse or neglect	11,195	5784
perpetration	11,017	5403
ycsu[e]	10,511	6488

[a]thc: tetrahydrocannabinol.

[b]si: suicidal ideation.

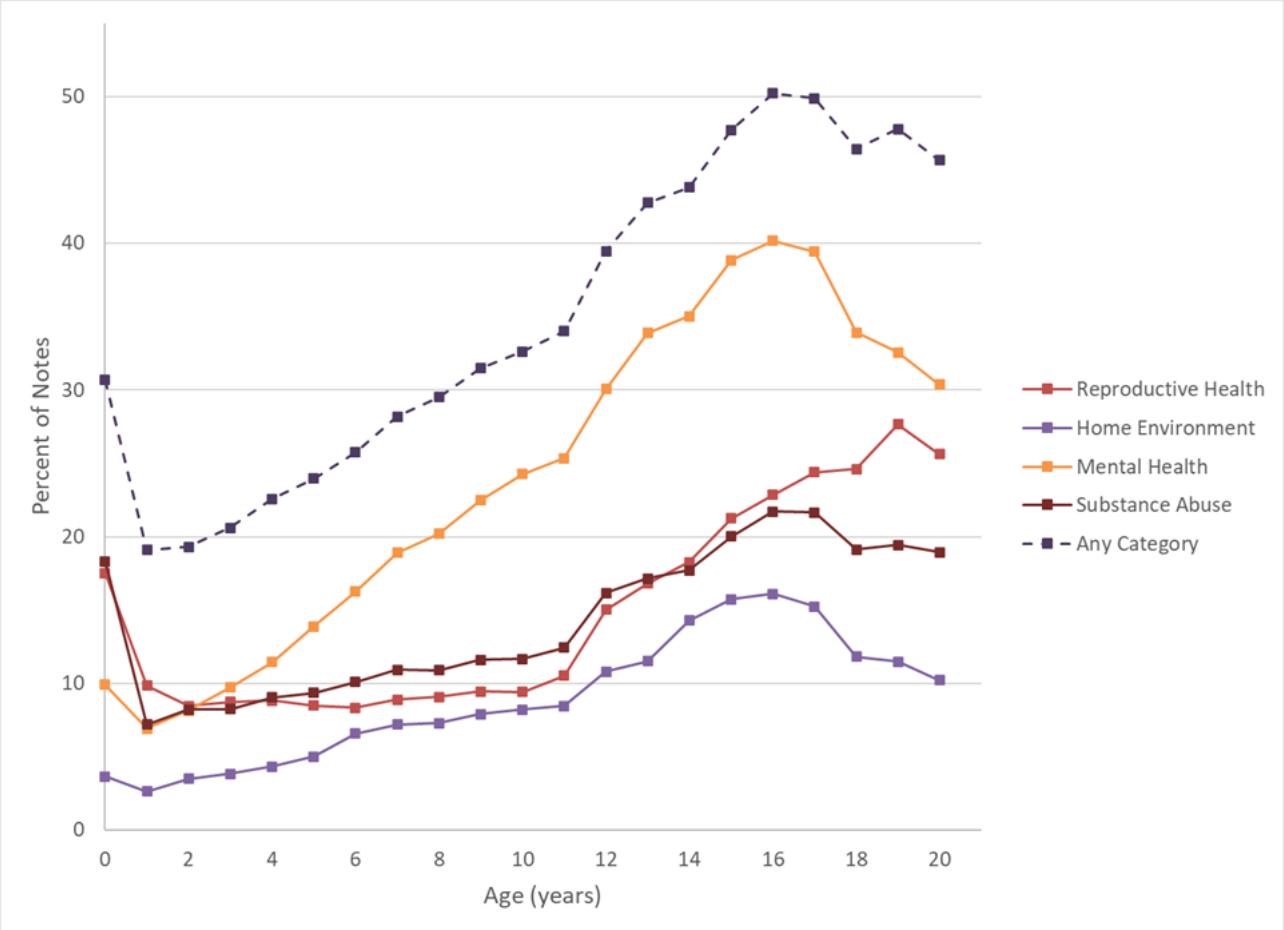
[c]sti: sexually transmitted infection.

[d]cps: Child Protective Services.

[e]ycsu: Youth Christian Social Union.

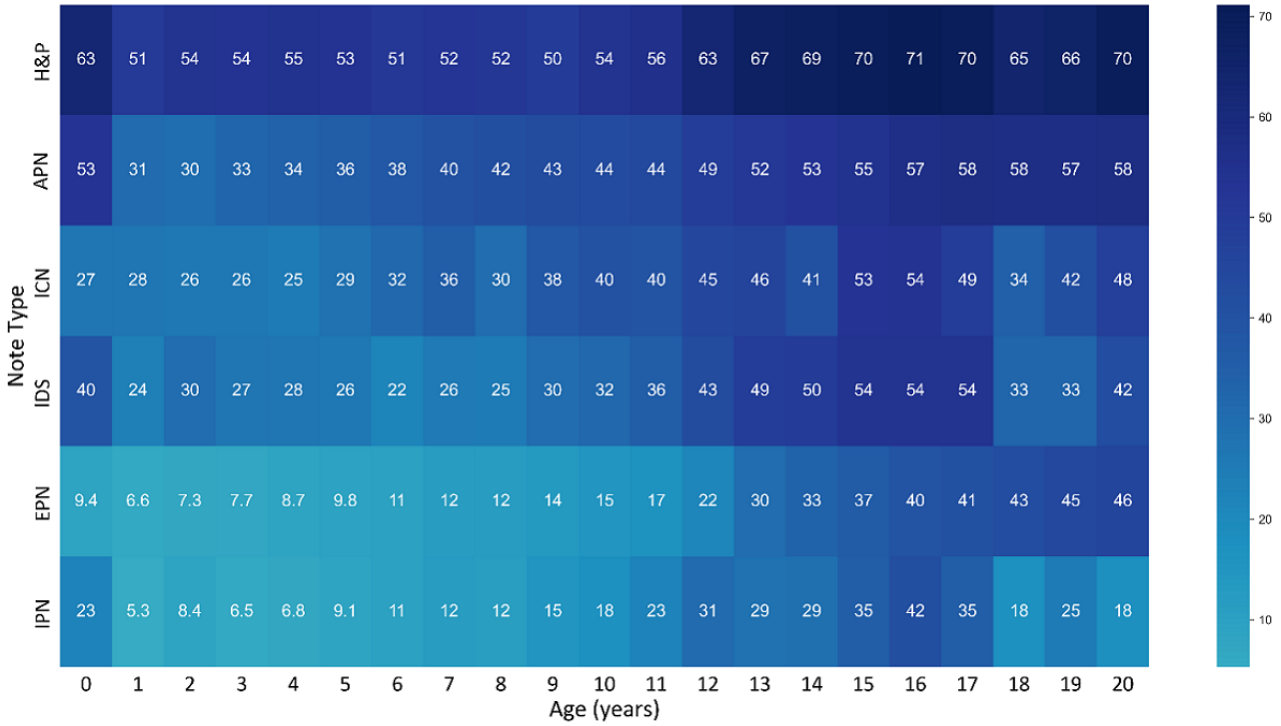
Mental health terms were documented most, occurring in 254,975 (19.5%) of notes, followed by reproductive health in 184,720 (14.2%), substance use in 184,342 (14.2%), and home environment in 95,598 (7.4%). The difference in term prevalence between mental health and the next closest category (reproductive health) was statistically significant (P<.05). This difference was most notable in the adolescent years. In the first year of life, substance use and reproductive health terms are more frequently documented than terms from the other two categories. Figure 2 demonstrates the prevalence of any term from different categories by age.

Figure 2. Percent of notes containing sensitive terms by age of patient and category. Line graph depicting percent of clinical notes containing at least one sensitive term over age. Sensitive terms are found in a portion of clinical notes for all patient ages. This figure demonstrates that while all categories show an upward trend during adolescent age, in the first year of life, reproductive health and substance abuse categories are the most frequently documented.



The prevalence of sensitive terms varied by note type. The inpatient H&P note type contained at least one sensitive term 57.1% (19,854/34,771) of the time, whereas ambulatory progress notes contained sensitive terms in 46.7% (265,302/563,273) of cases. Figure 3 shows a heat map of the percentage of notes containing a sensitive term by note type and age. The notes with the highest percentage sensitive terms are H&P notes among adolescent patients aged 12-20 years (66%-73%).

Figure 3. Percent of notes containing sensitive terms by age and note type. This heat map demonstrates the specific note types that contain at least one sensitive term of any category. Sensitive terms are found in a portion of all clinical note types examined in all age groups. This figure demonstrates that while all categories show an upward trend of including sensitive notes during adolescent age, the history and physical note is most likely to contain sensitive term overall. APN: ambulatory progress note; EPN: emergency care and urgent care progress note; H&P: history and physical note; ICN: inpatient consult note; IDS: inpatient discharge summary; IPN: inpatient progress note.



In terms of specialty, the highest prevalence of ambulatory progress notes with at least one sensitive term occurred in pain management (950/1112, 85.4%) child abuse pediatrics (1092/1282, 85.2%), obstetrics or gynecology (5701/6707, 85.0%), and behavioral health (2128/2589, 82.2%). The difference between obstetrics or gynecology and behavioral health was statistically significant (P<.05). Pediatric primary care had an overall prevalence of 44.0% (175,173/398,120) of notes with sensitive terms.

Discussion

Principal Findings

To our knowledge, our study is one of the first studies to define sensitive terms to represent categories of confidential information using NLP. In this study, sensitive terms were identified in notes from every clinical context, provider type, specialty, and in all ages included in the study cohort. Prior to the 21st Century Cures Act information blocking regulations, sharing of notes through the portal was not mandatory. Organizations voluntarily sharing notes (eg, the OpenNotes initiative) commonly prevented the release of notes in specific specialties and, in particular, among the adolescent age group to protect confidentiality [22,23]. Now, federal regulations limit the circumstances under which health care providers may withhold EHI. Moreover, our data show that sensitive terms are present diffusely across all notes in our system, making approaches that restrict notes within specific specialties or certain age groups no longer viable options.

Institutions rely on manual notation by the author [7] even though research has shown that providers may not be aware of the confidentiality laws in their state [8]. Our work indicates that there is no generalizable rule that can be applied to prevent unintended disclosure of sensitive terms in clinical notes. It is important to note that the presence of a sensitive term in a clinical note is not equivalent to an EHI that should be withheld. Instead, providers need to be cognizant of sensitive term documentation before sharing with a patient, parent, or guardian. Future work is being carried out at our organization to alert authors of the presence of sensitive terms before releasing to a portal.

We found that sensitive terms related to mental health were the most common overall, but in the first years of life, terms related to reproductive health and substance abuse were more prevalent. This is most likely due to the documentation of maternal history [24]. Disclosure of maternal history may lead to privacy violations when viewed by another legal guardian or by the patient at an older age [25].

Inpatient H&P and ambulatory progress note types have a higher prevalence of sensitive terms across all age groups. This may be due to the documentation of various screening tools used during patient encounters. For example, the American Academy of Pediatrics recommends universal psychosocial and depression screening beginning at the age of 11 years and risk assessment for alcohol and drug use during well-child exams [9,10]. The US Preventive Services Task Force recommends routinely screening adolescents for HIV starting at the age of 15 years [26,27]. To facilitate compliance with screening for billing purposes, clinical note templates often include these screening

questions, along with their requisite sensitive terms, to prompt clinicians during the visit. Adolescent patients in ambulatory settings report having frequent private conversations with their provider [11]; however, in the pediatric inpatient setting, nonobservance of privacy protections is often reported [28].

The notes with highest prevalence of sensitive terms were adolescent patient H&P notes. Studies have shown adolescents would forgo care if confidentiality regarding sensitive issues was not assured [29,30]. In addition to missed care, the release of sensitive information for adolescents may constitute a breach of state or federal privacy law. Individual health systems define different types of portal access, often giving adolescent patients full or limited access [31]. Under the Health Insurance Portability and Accountability Act, parents and legal guardians are considered personal representatives of patients under 18 years (ie, minors) and are thus afforded proxy access to the patient's EHI, including access through patient portals [16,23,31]. However, a recent work by Ip et al [32] demonstrated that parents are often active users of adolescent portal accounts, making it even more crucial that note authors recognize sensitive content in their notes and take into consideration who can see what in their patient portal.

Clinical Implications

The presence of sensitive terms in a clinical note does not necessarily indicate that a note is to be considered confidential. However, confidential notes likely contain sensitive terms. Providers need to be educated on what information is protected by federal and state laws, and they should determine, on a case-by-case basis, which notes are not to be shared. Furthermore, patients and guardians should be informed regarding who has access to what information in a patient portal, and proxy access policies should be regularly reviewed and updated as needed. Ideally, sensitive conversations with patients or guardians should also include discussion about whether this information should be kept confidential or shared through the patient portal.

Our approach identified sensitive terms anywhere in the body of a clinical note regardless of whether it was entered manually by the provider or added to the note from discrete data sources elsewhere in the EHR, such as prior family or social history documentation. For example, if a patient's problem list contains a sensitive term such as "prior suicide attempt" and the problem list is included in a note template, it may be added automatically to a clinical note for a visit unrelated to a sensitive condition, thus rendering the note inadvertently confidential. Similarly, copy and paste behavior can result in unrecognized inclusion of sensitive information in otherwise nonconfidential notes. For these reasons, additional work is needed to identify the source of the sensitive terms found in Figure 3.

Policy Considerations

This study demonstrates that sensitive terms are documented in clinical notes across all ages, including an increase in mental health–related terms starting at the age of 10 years. Laws and institutional policies are often designed to protect adolescent privacy; however, there is often a lack of protection for other age groups. Current law and policies might need to be revisited in light of this research.

Future Development

Further technological development is needed for EHRs and other health information technologies to support improved protection of patient and guardian privacy. Tools based on NLP techniques may now be possible, which could provide real-time feedback to note authors in situations where sensitive content present in clinical documentation may not otherwise be recognized and protected from inadvertent disclosure. Several challenges may be encountered when considering the implementation of similar NLP-supported tools [33]. Prior to implementation, a health institution must ensure data privacy and integrity, consider the necessities of information system infrastructure, model, and system performance, as well as performing assessment for algorithmic bias [33-35]. From a provider standpoint, as many institutions are working on reducing provider alert burden [36], they should be cautious toward implementing such tools not to increase provider alerting, which has been associated with provider burnout. As such, provider acceptance of these tools should be monitored over time [33].

Limitations

We defined sensitive terms broadly to increase the likelihood of identifying notes that might contain information that should remain confidential. However, the presence of a sensitive term by itself does not equate to confidential content. For example, "partner" (or "partners") is a very common term in the reproductive health category. This word could be used in phrases that indicate confidential content, such as "sexual partner," but also in nonconfidential content such as "partners with teachers to assess behavior at school." This highlights the need for providers to make the final determination of whether a clinical note contains confidential content.

Our list of sensitive terms is not comprehensive. DeepSuggest expands a single keyword to up to 60 potentially related terms in an unsupervised manner. However, given that less than 50% of the expanded terms were considered sensitive, expanding the potentially related term set may not improve the identification of additional sensitive terms. In future work, these sensitive term findings may be used to develop a specific algorithm to locate sensitive terms with a greater degree of precision. For instance, the deep learning algorithms have been successfully used to identify adverse events [37].

Conclusion

Clinical notes often contain sensitive terms and thus pose a challenge in complying with new regulations that require more timely and transparent disclosure of clinical notes to patients, parents, and legal guardians. Confidential information protected by law and ethical standards should be withheld from disclosure. The presence of sensitive terms in a clinical note may indicate documentation of confidential information requiring protection from inadvertent disclosure. To the best of our knowledge, this is the first study that defines sensitive terms in this context using an iterative process of expert opinion and NLP techniques, thus allowing an approximation of the actual prevalence of sensitive

terms in provider clinical notes in a pediatric population. We hope this work is the first step toward developing tools to assist providers in identifying potentially confidential information present in their clinical notes, thereby avoiding accidental disclosure to the wrong person.

Acknowledgments

We would like to acknowledge the contribution of Richard Hoyt, Dan Digby, and Rajesh Ganta for their assistance in data acquisition and search. We are also thankful to Brandon Abbott for his help on the illustrations.

This study is (partially) supported through a Patient-Centered Outcomes Research Institute Award (ME-2017C1-6413) under the name of “Unlocking Clinical Text in EMR by Query Refinement Using Both Knowledge Bases and Word Embedding.”

All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute, its Board of Governors, or Methodology Committee.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental table of sensitive terms.
[XLSX File (Microsoft Excel File), 40 KB - medinform_v10i6e38482_app1.xlsx]

References

1. Lee J, Miller S, Mezoff E, Screws J, Sauer C, Huang JS. The 21st Century CURES Act in Pediatric Gastroenterology: Problems, Solutions, and Preliminary Guidance. J Pediatr Gastroenterol Nutr 2021 May 01;72(5):700-703. [doi: 10.1097/MPG.0000000000003117] [Medline: 33720090]
2. Society for Adolescent HealthMedicine, Gray SH, Pasternak RH, Gooding HC, Woodward K, Hawkins K, et al. Recommendations for electronic health record use for delivery of adolescent health care. J Adolesc Health 2014 Apr;54(4):487-490. [doi: 10.1016/j.jadohealth.2014.01.011] [Medline: 24656534]
3. Bayer R, Santelli J, Klitzman R. New challenges for electronic health records: confidentiality and access to sensitive health information about parents and adolescents. JAMA 2015 Jan 06;313(1):29-30. [doi: 10.1001/jama.2014.15391] [Medline: 25562260]
4. English A, Bass L, Boyle A, Eshragh F. State Minor Consent Laws: A summary (3rd edition). FreeLists. 2010. URL: https://www.freelists.org/archives/hilac/02-2014/pdftRo8tw89mb.pdf [accessed 2022-05-24]
5. English A, Wilcox B. Work group vi: exploring the influence of law and public policy on adolescent health. Journal of Adolescent Health 2002 Dec;31(6):293-295. [doi: 10.1016/s1054-139x(02)00491-3]
6. Bourgeois FC, DesRoches CM, Bell SK. Ethical Challenges Raised by OpenNotes for Pediatric and Adolescent Patients. Pediatrics 2018 Jun 18;141(6):e20172745. [doi: 10.1542/peds.2017-2745] [Medline: 29776979]
7. Parsons C, Hron J, Bourgeois F. Preserving privacy for pediatric patients and families: use of confidential note types in pediatric ambulatory care. J Am Med Inform Assoc 2020 Nov 01;27(11):1705-1710 [FREE Full text] [doi: 10.1093/jamia/ocaa202] [Medline: 32989446]
8. Riley M, Ahmed S, Reed BD, Quint EH. Physician Knowledge and Attitudes around Confidential Care for Minor Patients. J Pediatr Adolesc Gynecol 2015 Aug;28(4):234-239. [doi: 10.1016/j.jpag.2014.08.008] [Medline: 26024938]
9. Levy SJL, Williams JF, Committee on Substance Use and Prevention. Substance Use Screening, Brief Intervention, and Referral to Treatment. Pediatrics 2016 Jul;138(1):1-15. [doi: 10.1542/peds.2016-1211] [Medline: 27325634]
10. Committee on Practice and Ambulatory Medicine, Bright Futures Steering Committee. Recommendations for Preventive Pediatric Health Care. Pediatrics 2007;120(6):1376. [doi: 10.1542/peds.2007-2901]
11. Klein JD, Wilson KM, McNulty M, Kapphahn C, Scott Collins K. Access to medical care for adolescents: results from the 1997 Commonwealth Fund Survey of the Health of Adolescent Girls. Journal of Adolescent Health 1999 Aug;25(2):120-130. [doi: 10.1016/s1054-139x(98)00146-3]
12. Substance Abuse Confidentiality Regulations. Substance Abuse and Mental Health Services Administration. URL: https://www.samhsa.gov/about-us/who-we-are/laws-regulations/confidentiality-regulations-faqs [accessed 2022-05-24]
13. Questions and Answers about HIPAA and Mental Health. Department of Health and Human Services. URL: https://www.hhs.gov/sites/default/files/hipaa-privacy-rule-and-sharing-info-related-to-mental-health.pdf [accessed 2022-05-24]
14. Akinbami L, Gandhi H, Cheng T. Availability of adolescent health services and confidentiality in primary care practices. Pediatrics 2003 Feb;111(2):394-401. [doi: 10.1542/peds.111.2.394] [Medline: 12563069]
15. Butler PW, Middleman AB. Protecting Adolescent Confidentiality: A Response to One State's "Parents' Bill of Rights". J Adolesc Health 2018 Sep;63(3):357-359. [doi: 10.1016/j.jadohealth.2018.03.015] [Medline: 30077547]

16. Maslyanskaya S, Alderman EM. Confidentiality and Consent in the Care of the Adolescent Patient. Pediatr Rev 2019 Oct 01;40(10):508-516. [doi: 10.1542/pir.2018-0040] [Medline: 31575802]
17. Thompson LA, Martinko T, Budd P, Mercado R, Schentrup AM. Meaningful Use of a Confidential Adolescent Patient Portal. J Adolesc Health 2016 Feb;58(2):134-140. [doi: 10.1016/j.jadohealth.2015.10.015] [Medline: 26802988]
18. Committee on Child AbuseNeglect. Policy statement--Child abuse, confidentiality, and the health insurance portability and accountability act. Pediatrics 2010 Jan;125(1):197-201. [doi: 10.1542/peds.2009-2864] [Medline: 20026490]
19. Young-Wolff KC, Klebaner D, Folck B, Carter-Harris L, Salloum RG, Prochaska JJ, et al. Do you vape? Leveraging electronic health records to assess clinician documentation of electronic nicotine delivery system use among adolescents and adults. Prev Med 2017 Dec;105:32-36 [FREE Full text] [doi: 10.1016/j.ypmed.2017.08.009] [Medline: 28823688]
20. Wu C, Chang C, Robson D, Jackson R, Chen S, Hayes RD, et al. Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register. PLoS One 2013 Sep 12;8(9):e74262 [FREE Full text] [doi: 10.1371/journal.pone.0074262] [Medline: 24069288]
21. Moosavinasab S, Sezgin E, Sun H, Hoffman J, Huang Y, Lin S. DeepSuggest: Using Neural Networks to Suggest Related Keywords for a Comprehensive Search of Clinical Notes. ACI Open 2021 Jun 06;05(01):e1-e12. [doi: 10.1055/s-0041-1729982]
22. Patients and Clinicians on the Same Page. OpenNotes. URL: http://www.opennotes.org/ [accessed 2022-05-24]
23. Wilcox L, Sharko M, Hong M, Hollberg J, Ancker JS. The need for guidance and consistency in adolescent privacy policies: a survey of CMIOs. AMIA Annu Symp Proc 2018;2018:1084-1092 [FREE Full text] [Medline: 30815150]
24. Abhyankar S, Demner-Fushman D. A simple method to extract key maternal data from neonatal clinical notes. AMIA Annu Symp Proc 2013;2013:2-9 [FREE Full text] [Medline: 24551317]
25. Scibilia J. How to protect maternal health information in newborn's medical record. American Academy of Pediatrics 2014;35(12):4 [FREE Full text]
26. Moyer VA, U.S. Preventive Services Task Force. Screening for HIV: U.S. Preventive Services Task Force Recommendation Statement. Ann Intern Med 2013 Jul 02;159(1):51-60 [FREE Full text] [doi: 10.7326/0003-4819-159-1-201307020-00645] [Medline: 23698354]
27. US Preventive Services Task Force, Owens DK, Davidson KW, Krist AH, Barry MJ, Cabana M, et al. Screening for HIV Infection: US Preventive Services Task Force Recommendation Statement. JAMA 2019 Jun 18;321(23):2326-2336. [doi: 10.1001/jama.2019.6587] [Medline: 31184701]
28. Talib H, Silver E, Alderman E. Challenges to Adolescent Confidentiality in a Children's Hospital. Hosp Pediatr 2016 Aug;6(8):490-495. [doi: 10.1542/hpeds.2016-0011] [Medline: 27461762]
29. Ford CA. Influence of Physician Confidentiality Assurances on Adolescents' Willingness to Disclose Information and Seek Future Health Care. JAMA 1997 Sep 24;278(12):1029. [doi: 10.1001/jama.1997.03550120089044]
30. Reddy DM, Fleming R, Swain C. Effect of mandatory parental notification on adolescent girls' use of sexual health care services. JAMA 2002 Aug 14;288(6):710-714. [doi: 10.1001/jama.288.6.710] [Medline: 12169074]
31. Bourgeois FC, Taylor PL, Emans SJ, Nigrin DJ, Mandl KD. Whose Personal Control? Creating Private, Personally Controlled Health Records for Pediatric and Adolescent Patients. Journal of the American Medical Informatics Association 2008 Nov 01;15(6):737-743. [doi: 10.1197/jamia.m2865]
32. Ip W, Yang S, Parker J, Powell A, Xie J, Morse K, et al. Assessment of Prevalence of Adolescent Patient Portal Account Access by Guardians. JAMA Netw Open 2021 Sep 01;4(9):e2124733 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.24733] [Medline: 34529064]
33. Verma AA, Murray J, Greiner R, Cohen JP, Shojania KG, Ghassemi M, et al. Implementing machine learning in medicine. CMAJ 2021 Aug 30;193(34):E1351-E1357 [FREE Full text] [doi: 10.1503/cmaj.202434] [Medline: 35213323]
34. Sun W, Nasraoui O, Shafto P. Evolution and impact of bias in human and machine learning algorithm interaction. PLoS One 2020 Aug 13;15(8):e0235502 [FREE Full text] [doi: 10.1371/journal.pone.0235502] [Medline: 32790666]
35. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model. JMIR Med Inform 2022 Feb 10;10(2):e32875 [FREE Full text] [doi: 10.2196/32875] [Medline: 35142635]
36. Chaparro JD, Hussain C, Lee JA, Hehmeyer J, Nguyen M, Hoffman J. Reducing Interruptive Alert Burden Using Quality Improvement Methodology. Appl Clin Inform 2020 Jan 15;11(1):46-58 [FREE Full text] [doi: 10.1055/s-0039-3402757] [Medline: 31940671]
37. Borjali A, Magnéli M, Shin D, Malchau H, Muratoglu OK, Varadarajan KM. Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: A case study of detecting total hip replacement dislocation. Comput Biol Med 2021 Feb;129:104140. [doi: 10.1016/j.compbiomed.2020.104140] [Medline: 33278631]

Abbreviations

EHI: electronic health information
EHR: electronic health record
H&P: history and physical

NLP: natural language processing

Edited by C Lovis; submitted 04.04.22; peer-reviewed by J Kim; comments to author 28.04.22; revised version received 09.05.22; accepted 10.05.22; published 10.06.22.

Please cite as:
Lee J, Yang S, Holland-Hall C, Sezgin E, Gill M, Linwood S, Huang Y, Hoffman J
Prevalence of Sensitive Terms in Clinical Notes Using Natural Language Processing Techniques: Observational Study
JMIR Med Inform 2022;10(6):e38482
URL: https://medinform.jmir.org/2022/6/e38482
doi:10.2196/38482
PMID:35687381

Original Paper

Associations Between Family Member Involvement and Outcomes of Patients Admitted to the Intensive Care Unit: Retrospective Cohort Study

Tamryn F Gray[1,2,3*], RN, MPH, PhD; Anne Kwok[3*], BS; Khuyen M Do[3*], BS, MPH; Sandra Zeng[3*], BA; Edward T Moseley[3*], BS; Yasser M Dbeis[4*]; Renato Umeton[4,5,6,7*], PhD; James A Tulsky[1,2,3*], MD; Areej El-Jawahri[1,8*], MD; Charlotta Lindvall[1,2,3*], MD, PhD

[1]Department of Medicine, Harvard Medical School, Boston, MA, United States

[2]Division of Palliative Medicine, Brigham and Women's Hospital, Boston, MA, United States

[3]Department of Psychosocial Oncology and Palliative Care, Dana-Farber Cancer Institute, Boston, MA, United States

[4]Department of Informatics & Analytics, Dana-Farber Cancer Institute, Boston, MA, United States

[5]Department of Biological Engineering and Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States

[6]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, United States

[7]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, United States

[8]Department of Medicine, Massachusetts General Hospital Cancer Center, Boston, MA, United States

[*]all authors contributed equally

Corresponding Author:
Tamryn F Gray, RN, MPH, PhD
Department of Medicine
Harvard Medical School
25 Shattuck Street
Boston, MA, 02115
United States
Phone: 1 617 582 9186
Email: tamryn_gray@dfci.harvard.edu

Abstract

Background: Little is known about family member involvement, by relationship status, for patients treated in the intensive care unit (ICU).

Objective: Using documentation of family interactions in clinical notes, we examined associations between child and spousal involvement and ICU patient outcomes, including goals of care conversations (GOCCs), limitations in life-sustaining therapy (LLST), and 3-month mortality.

Methods: Using a retrospective cohort design, the study included a total of 858 adult patients treated between 2008 and 2012 in the medical ICU at a tertiary care center in northeastern United States. Clinical notes generated within the first 48 hours of admission to the ICU were used with standard machine learning methods to predict patient outcomes. We used natural language processing methods to identify family-related documentation and abstracted sociodemographic and clinical characteristics of the patients from the medical record.

Results: Most of the 858 patients were White (n=650, 75.8%); 437 (50.9%) were male, 479 (55.8%) were married, and the median age was 68.4 (IQR 56.5-79.4) years. Most patients had documented GOCC (n=651, 75.9%). In adjusted regression analyses, child involvement (odds ratio [OR] 0.81; 95% CI 0.49-1.34; *P*=.41) and child plus spouse involvement (OR 1.28; 95% CI 0.8-2.03; *P*=.3) were not associated with GOCCs compared to spouse involvement. Child involvement was not associated with LLST when compared to spouse involvement (OR 1.49; 95% CI 0.89-2.52; *P*=.13). However, child plus spouse involvement was associated with LLST (OR 1.6; 95% CI 1.02-2.52; *P*=.04). Compared to spouse involvement, there were no significant differences in the 3-month mortality by family member type, including child plus spouse involvement (OR 1.38; 95% CI 0.91-2.09; *P*=.13) and child involvement (OR 1.47; 95% CI 0.9-2.41; *P*=.12).

Conclusions: Our findings demonstrate that statistical models derived from text analysis in the first 48 hours of ICU admission can predict patient outcomes. Early child plus spouse involvement was associated with LLST, suggesting that decisions about LLST were more likely to occur when the child and spouse were both involved compared to the involvement of only the spouse. More research is needed to further understand the involvement of different family members in ICU care and its association with patient outcomes.

(JMIR Med Inform 2022;10(6):e33921) doi:10.2196/33921

KEYWORDS

critical care; natural language processing; family; electronic health records; goals of care; intensive care unit; ICU

Introduction

Background

Mechanically ventilated critically ill patients often lack decisional capacity [1-3] and rely on family members for their care and medical decision-making [2-6]. In the critical care environment, where decisions about tests, procedures, and treatments must be made quickly [7,8], physicians turn to surrogate decision makers for guidance about goals of care and making decisions to limit life-sustaining treatment [1,6,7,9-11]. Critical care organizations have strongly encouraged a family-centered approach to care [12,13]; however, information about when, how, and which family members are engaged over the course of illness remains poorly understood [7].

Although clinicians often expect 1 family member to be the "voice" for the patient, several family members are often involved [14,15]. In the event that the patients no longer possess the requisite capacity to make their own health care decisions or are too ill, which is common in the intensive care unit (ICU) setting [16], the health care proxy is the most common way through which patients appoint a surrogate decision-maker to make decisions on their behalf [17]. Typically, the health care provider has a priority list of individuals to be designated for this role, and at the top of the hierarchy is often the patient's spouse followed by the adult child/children, parents, and adult sibling(s) [18,19]. In American families, the spouse is commonly the first in line to assume the role of a health care proxy [20] and is informed if he or she is aware of (1) the patient's personal definition of quality of life, (2) his or her specific plan if he or she cannot achieve this quality of life, and (3) desired location of death [21]. If no spouse is available to provide care, adult children often take on the role and sometimes share care tasks [22]. Although studies examining family members in the ICU have focused on family needs, communication, and satisfaction with care [23-27], to our knowledge, no studies have discerned the distinct involvement of spouses and children in care decisions and its impact on patient outcomes in the medical ICU (MICU) setting.

Objective

We sought to describe family member involvement in decision-making, by relationship status, for patients treated in the ICU. We also examined patient characteristics associated with child and spousal involvement. Using documentation of family interactions in clinical notes, we examined the association between child and spousal involvement in the first 48 hours of admission and ICU patient outcomes, including goals of care conversations (GOCCs), limitations in life-sustaining therapy (LLST), and mortality.

Methods

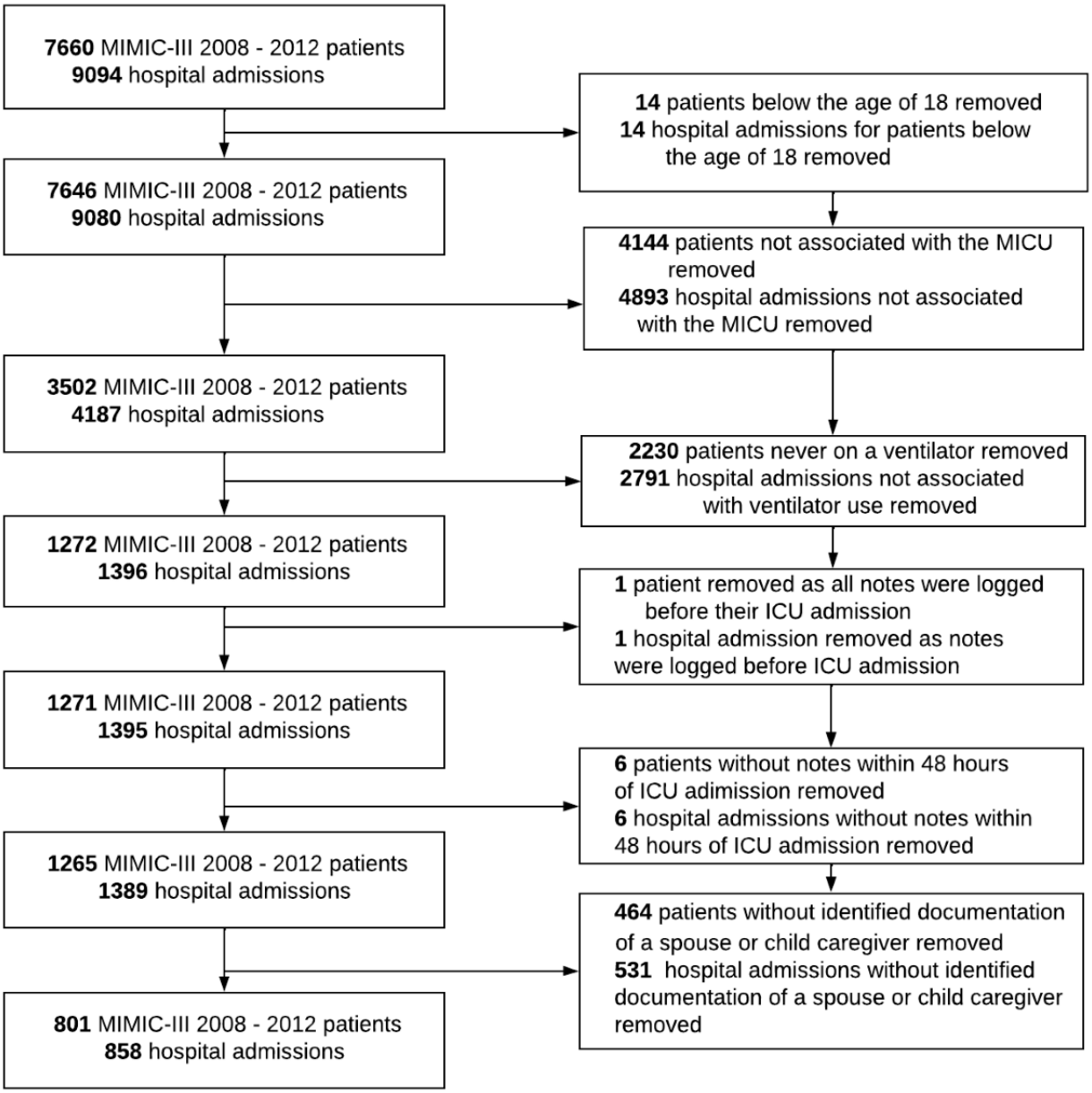
Data Source

Our data source was the Medical Information Mart for Intensive Care III (MIMIC-III) database, developed by the Massachusetts Institute of Technology (MIT) and Beth Israel Deaconess Medical Center (BIDMC), and it is a large, freely available database. The MIMIC database provided deidentified demographic, administrative, clinical, and survival outcome data for all adult ICU admissions at the BIDMC [28]. For our analysis, we used data between 2008 and 2012 to include clinical notes from a broad group of clinicians likely to document engagement with patients' families, including physicians, nursing staff, social workers, case managers, and physician assistants [29]. The Institutional Review Board of the BIDMC and MIT approved use of the MIMIC-III database by any investigator who fulfills data-user requirements [29]. This research was deemed exempt by the Dana-Farber Cancer Institute Institutional Review Board (approval number 18-192).

Study Population

The study population included patients at least 18 years of age who were treated in the MICU at the BIDMC in Boston between 2008 and 2012 (Figure 1). We focused exclusively on MICU patients commonly facing life-threatening conditions that may warrant family involvement in decision-making [30]. We excluded patients with an ICU length of stay (LOS) less than 48 hours and those lacking available clinical notes due to potential privacy disclosures (eg, VIPs). For patients with multiple ICU admissions during a single hospitalization, only the first admission was used for analysis.

Figure 1. Flow diagram showing patient selection in the study. ICU: intensive care unit; MIMIC-III: Medical Information Mart for Intensive Care III. It is a deidentified demographic, administrative, clinical, and survival outcome database for adult ICU admissions. MICU: medical intensive care unit.



Natural Language Processing (NLP)

Family communication is often recorded as free text in the clinical notes [31]. Manual abstraction of these data is time-consuming and prone to human error, thus benefiting from a structured approach using standard NLP methods [31]. The ability of NLP methods to identify electronic health record (EHR) documentation of family involvement in the ICU was evaluated using a multistep process. First, we constructed a keyword library to develop a standard structure, including typographical errors that might be present. We used the text annotation software, ClinicalRegex [32], to identify documentation of child and spousal or partner involvement in the EHR (referred to as "family involvement"). ClinicalRegex was developed by the Lindvall Lab at Dana-Farber Cancer Institute and has been applied in multiple studies [32-35] to identify defined keywords or phrases within clinical notes, accounting for varieties in language, spelling, and punctuation. Using a predefined ontology, the software displayed clinical notes that contain the highlighted keywords or phrases associated with family. Our ontology contained two domains of documentation regarding family involvement: (1) spouse or partner and (2) children. The keyword library was refined to prioritize sensitivity over specificity and validated by expert review of a random selection of notes identified by the library as well as manual review of notes not identified by the library. The final keyword library is provided in Multimedia Appendix 1.

Second, once the ontology was developed, independent coders (TFG, KMD, and SZ) reviewed a subset of 100 random samples of charts in ClinicalRegex using the keyword library to examine whether each clinical note contained keywords related to family

involvement. Human experts labeled notations using prespecified codes (eg, using "0" to label notations where keywords appeared out of context for exclusion or using "1" to label notations for inclusion), and the presence or absence of family-related documentation was determined at the hospital admission level. Interrater agreement was excellent (κ values of 0.83 and 0.82 for child and spouse, respectively).

Study Measures

Family Involvement

To identify family-related documentation in the EHR, we first conducted a literature search of relevant keywords related to spouse and child [22,36,37]. For our keyword library, we developed an extensive list to account for the wide variation in describing spouse and child. For example, spouse was described as husband, wife, fiancé, girlfriend, boyfriend, companion, partner, spouse, comate, etc. Child was described as son, daughter, grandchild, teenage, girl, boy, child, children, grandson, granddaughter, etc. Multimedia Appendix 1 presents the exact phrases used in the keyword bank. Multimedia Appendices 2 and 3 respectively describe examples of how the keywords found in the clinical notes were used in the relevant context as well as the keywords that were not used in the analysis because they were used in a nonrelevant context.

Sociodemographic and Clinical Factors

We collected demographic information (admission age, sex, race, ethnicity, and marital status) as well as clinical characteristics including the sequential organ failure assessment score (SOFA) and Elixhauser Comorbidity Index. The SOFA score described the time course of multiple organ dysfunction using a limited number of routinely measured variables [38], and the Elixhauser Comorbidity Index quantified the effect of comorbidities on patient outcomes [39]. The sociodemographic and clinical characteristics of the patients were ascertained by EHR data extraction.

Health Care Usage

For health care usage outcomes, the discharge location was included (eg, home, home health care, hospice, short-term hospital, long-term-care hospital, skilled nursing facility [SNF], "other facilities," and in-hospital death). The LOS for obtaining the hospitalization index and hospital readmission were also determined for each patient. For our analyses, home was defined as either home or home health care. Facility was defined as either hospice, short-term hospital, long-term care, SNF, or "other facilities."

Outcome Measures

GOCC Documentation

The National Quality Forum recommends that GOCCs be documented in the EHR within the first 48 hours of an ICU admission, especially among frail and seriously ill patients. For our study, we identified GOCCs using an operational definition previously described elsewhere [29]. GOCC documentation required both of the following details: (1) mention of a conversation with either the patient or a family member and (2) mention of a specific care preference pertaining to hospital care [29]. Ascertained by free-text data in the clinical notes, GOCC documentation included discussion about advance care planning activities (values, goals, and preferences considering future care), completion of advance directives or Physician Order for Life-Sustaining Treatment forms, or referral to hospice or subspecialty palliative care services [40].

LLST Conversations

Similar to our previous study [29] and other research [41], LLST included documentation from free-text data within clinical notes regarding a do-not-resuscitate or do-not-intubate (DNR/DNI) code status, LLST, acknowledgment of patient or family wishes to decline any interventional procedures (including central venous line, temp wire placement, etc) but agreement for medical management, preference for no heroic measures, no blood transfusions, no resuscitations, and no blood pressure interventions.

Mortality

To assess the 3-month mortality since hospital admission, we used a binary outcome of died and not died within 3 months since hospital admission based on EHR review.

Statistical Analysis

We used descriptive statistics to summarize the sample, including the sociodemographic and clinical characteristics of the patients as well as health care use and mortality. We performed univariate analyses to assess the relationships between the sociodemographic and clinical characteristics of the patients and family involvement, stratified by the type of family member (overall cohort, both child and spousal involvement, child only involvement, and spouse only involvement). To assess the independent associations between family involvement and GOCC, LLST, and 3-month mortality, we developed multivariable logistic regression models. For each dependent variable, separate models were fitted, adjusting for sex, marital status, race and ethnicity, age, SOFA, and Elixhauser scores identified a priori based on prior literature [22,38,42,43]. All statistical tests and CIs, as appropriate, were performed as 2-sided tests, and all reported P values <.05 were considered statistically significant. We performed all statistical analyses using Python version 3.7.6 and library statsmodels version 0.12.0.

Results

Patient Characteristics

Table 1 describes the sociodemographic and clinical characteristics of the patients at hospital admission (N=858). The median age was 68 (IQR: 57-79) years, most patients were non-Hispanic White (n=650, 75.8%), and approximately half were male (n=437, 50.9%) and married (n=479, 55.8%). The median SOFA and Elixhauser scores were 6 (IQR 4-9) and 5 (range 3-6), respectively. The median LOS was 9 (IQR 4.9-16.8) days. More than a quarter of these patients died in the ICU (n=253, 29.5%), whereas the majority were either discharged to a facility or home (n=379, 44.2% and n=223, 26%, respectively). When compared to child plus spouse involvement and spouse only involvement, patients with child only

involvement (n=352) were more likely to be female (235/352, 66.8%), not married or partnered (265/352, 75.3%), and older (median age of 76.7 [IQR 66-85] years) (Table 1). When both spouse and child were involved (n=202), patients were mostly male (123/202, 60.9%), married (170/202, 84.2%), and had a median age of 70 (range 61-77) years. In comparison with White patients, non-White patients had a high proportion of child only involvement (95/165, 57.6% vs. 242/650, 37.2%).

Table 1. Patient characteristics[a].

Characteristics	Overall (N=858)	Both (n=202)	Child (n=352)	Spouse (n=304)	*P* value
Sex, n (%)					<.001
Male	421 (49.1)	79 (39.1)	235 (66.8)	107 (35.2)	
Female	437 (50.9)	123 (60.9)	117 (33.2)	197 (64.8)	
Marital status, n (%)					<.001
Married	479 (55.8)	170 (84.2)	72 (20.5)	237 (78.0)	
Not married	354 (41.3)	27 (13.4)	265 (75.3)	62 (20.4)	
Unknown	25 (2.9)	5 (2.5)	15 (4.3)	5 (1.6)	
Ethnicity, n (%)					<.001
White (non-Hispanic)	650 (75.8)	163 (80.7)	242 (68.8)	245 (80.6)	
Other	165 (19.2)	29 (14.4)	95 (27.0)	41 (13.5)	
Unknown	43 (5.0)	10 (5.0)	15 (4.3)	18 (5.9)	
Admission age in years, median (IQR)	68.4 (56.5-79.4)	69.7 (61-77.4)	76.7 (66-85)	58.4 (48.4-67)	<.001
Hospital LOS[b] in days, median (IQR)	9 (4.9-16.8)	8.6 (4.7-16.1)	8 (4.7-14.7)	12.1 (6-21.1)	<.001
Discharge status, n (%)					<.001
Death	253 (29.5)	81 (40.1)	109 (31.0)	63 (20.7)	
Facility	379 (44.2)	85 (42.1)	158 (44.9)	136 (44.7)	
Home	223 (26.0)	36 (17.8)	84 (23.9)	103 (33.9)	
Unknown	3 (0.3)	0 (0)	1 (0.3)	2 (0.7)	
Mortality, n (%)					<.001
In-hospital mortality	253 (29.5)	81 (40.1)	109 (31.0)	63 (20.7)	
3 months from hospital admission	342 (39.9)	98 (48.5)	152 (43.2)	92 (30.3)	
1 year from hospital admission	442 (51.5)	118 (58.4)	198 (56.2)	126 (41.4)	
6 months from ICU[c] discharge	397 (46.3)	108 (53.5)	173 (49.1)	116 (38.2)	
Readmission, n (%)	196 (22.8)	43 (21.3)	90 (25.6)	63 (20.7)	.28
Documented goals of care conversation, n (%)	651 (75.9)	164 (81.2)	266 (75.6)	221 (72.7)	.09
Documented conversations about limitations in code status, n (%)	274 (31.9)	73 (36.1)	149 (42.3)	52 (17.1)	<.001
SOFA[d] score, median (IQR)	6 (4-9)	7 (5-10)	6 (4-9)	5.5 (3-8)	<.001
Elixhauser score, median (IQR)	5 (3-6)	5 (3-6)	5 (3-6)	4 (3-6)	.06

[a]Patient characteristics of study the cohort were stratified by documentation of family involvement. For discharge status, chi-square tests may not be valid due to a low number of examples in some categories.

[b]LOS: length of stay.

[c]ICU: intensive care unit.

[d]SOFA: sequential organ failure assessment score.

Association Between Family Involvement and GOCC

Overall, most patients had documented GOCC (651/858, 75.9%) (Table 1). Child involvement (odds ratio [OR] 0.81; 95% CI 0.49-1.34; *P*=.41) and involvement of child plus spouse (OR 1.28; 95% CI 0.8-2.03; *P*=.3) were not associated with GOCC when compared to spouse only involvement (Table 2).

Table 2. Goals of care conversations[a].

Variables	Odds ratio (95% CI)	*P* value
Sex (reference group: male)		
Female	1.18 (8.84-1.65)	.35
Marital status (reference group: married)		
Not married	1.19 (0.79-1.78)	.41
Unknown	1.09 (0.4-2.96)	.86
Ethnicity (reference group: White)		
Other	0.8 (0.54-1.2)	.28
Unknown	0.99 (0.46-2.13)	.97
Type of family member documentation identified (reference group: spouse only)		
Both child and spouse	1.28 (0.8-2.03)	.3
Child only	0.81 (0.49-1.34)	.41
Admission age	1.01 (1-1.03)	.05
Elixhauser score	0.99 (0.92-1.07)	.81
SOFA[b] score	1.09 (1.04-1.14)	<.001

[a]Exploratory analyses were conducted to investigate the association between documentation related to family member involvement and goals of care conversations.

[b]SOFA: sequential organ failure assessment score.

Association Between Family Involvement and LLST

More than a quarter of the patients (274/858, 31.9%) had documented LLST (Table 1). Child only involvement was not associated with LLST (OR 1.49; 95% CI 0.89-2.52; *P*=.13) compared to spouse only involvement. Child plus spouse involvement was associated with higher odds of LLST (OR 1.6; 95% CI 1.02-2.52; *P*=.04) compared to spouse only involvement (Table 3).

Table 3. Limitations in life-sustaining therapy conversations[a].

Variables	Odds ratio (95% CI)	*P* value
Sex (reference group: male)		
Female	0.98 (0.7-1.37)	.91
Marital status (reference group: married)		
Not married	1.51 (0.99-2.28)	.05
Unknown	1.16 (0.44-3.05)	.77
Ethnicity (reference group: White)		
Other	0.85 (0.57-1.28)	.44
Unknown	0.6 (0.27-1.36)	.22
Type of family member documentation identified (reference group: spouse only)		
Both child and spouse	1.6 (1.02-2.52)	.04
Child only	1.49 (0.89-2.52)	.13
Admission age	1.04 (1.03-1.06)	<.001
Elixhauser score	0.96 (0.89-1.03)	.24
SOFA[b] score	1.15 (1.11-1.2)	<.001

[a]Results of exploratory analyses to investigate the association between documentation related to family member involvement and limitations in life-sustaining therapy.

[b]SOFA: sequential organ failure assessment score

Association Between Family Involvement and Mortality

Over a third of the patients (342/858, 39.9%) died 3 months post hospital admission (Table 1). Compared to spouse only involvement, we found no significant differences in the 3-month mortality by family member type, including child plus spouse involvement (OR 1.38; 95% CI 0.91-2.09; *P*=.13) and child only involvement (OR 1.47; 95% CI 0.9-2.41; *P*=.12) (Table 4).

Table 4. Mortality at 3 months following admission[a].

Variables	Odds ratio (95% CI)	*P* value
Sex (reference group: male)		
Female	0.76 (0.56-1.05)	.09
Marital status (reference group: married)		
Not married	0.71 (0.47-1.05)	.09
Unknown	1.01 (0.41-2.51)	.98
Ethnicity (reference group: White)		
Other	0.82 (0.56-1.22)	.33
Unknown	1.28 (0.62-2.65)	.51
Type of family member documentation identified (reference group: spouse only)		
Both child and spouse	1.38 (0.91-2.09)	.13
Child only	1.47 (0.9-2.41)	.12
Admission age	1.03 (1.02-1.04)	<.001
Elixhauser score	1.01 (0.95-1.09)	.7
$SOFA_b$ score	1.2 (1.15-1.25)	<.001

[a]Results of exploratory analyses to investigate the association between documentation related to family involvement and 3-month mortality since hospital admission.

[b]SOFA: sequential organ failure assessment score.

Discussion

Principal Results

This study demonstrated that child plus spouse involvement in decision-making within the first 48 hours of an ICU stay was associated with LLST for mechanically ventilated patients when compared to spouse involvement only. To our knowledge, this is the first study to demonstrate an association between spouse plus child involvement and LLST in mechanically ventilated patients in the ICU. Family members may find it easier to make complex decisions in a group with other family members, and this approach may help in reaching a consensus in the context of a poor prognosis. Prior research has shown that family members take on the end-of-life (EOL) decision-maker role together as a unit and collaborate, and even designated surrogate decision makers prefer to structure the interaction around collaborative group decision-making rather than take on the role individually [14].

Unlike the association found between LLST and family involvement, there was no association between family member involvement and documentation of GOCC. One possible explanation is that a GOCC is defined as a palliative and end-of-life care process measure [40,44], meaning that such conversations are part of evidence-based guidelines and will occur regardless of which family member is present [45]. Meanwhile, LLST is the next step after a GOCC occurs and is important to establish when actually making decisions about life-limiting therapies, which may collectively involve the patients, their family members, and clinicians.

Comparison With Prior Work

Research has demonstrated that the type of family involvement often varies across racial and ethnic groups and there is a growing number of studies exploring the role of race, ethnicity, and culture in caregiving [36,46,47]. Compared to White patients, we observed that non-White patients had a high proportion of child only involvement. Similarly, previous studies have found that African American patients are more likely to receive assistance from adult children rather than spouses [47-49]. Williams and Dilworth-Anderson examined connections of social support for 187 community-dwelling African American elders and demonstrated that the adult child was the most common type of relationship to the care recipient (62%), surpassing spouse (6%), friend (3%), and other kin (29%) [50]. Similarly, Miller and Guo demonstrated that African American caregivers for persons with dementia were found to be younger, less educated, having lower income, and married for fewer years than White caregivers [51]. Though this study included participants from a single site, which may impact generalizability, the findings demonstrate potential racial and ethnic differences regarding the type of family members involved in care within the ICU setting, but further research is warranted.

Given the rising number of individuals facing serious illness, receiving critical care, and living longer, our study adds to the growing body of knowledge that calls for the need to develop approaches that are tailored to the specific subpopulations of family members who are involved in ICU patient care and decision-making.

Limitations

This study has several limitations. First, we examined data from 2008 to 2012, so our findings may not be generalizable to the more recent years. Second, the cross-sectional nature of the study did not enable us to assess causality or temporality between family involvement and patient outcomes. Third, because our sample was limited to clinical notes from a single tertiary care hospital in northeastern United States and lacked racial diversity, our algorithm may not be generalizable to other hospitals, ICU populations, or geographic areas. Fourth, as noted in other studies [34,44,52], our methods were dependent on the quantity and quality of documentation that exist in the EHR, so it is possible that some family-related documentation or actual interaction with and involvement of families may have been missed. Moreover, our models may not fully account for all possible confounders, and we were unable to capture other factors that may impact the relationship between family involvement and patient outcomes. Fifth, we focused on documentation generated within the first 48 hours by nurses, case managers, social workers, physician assistants, and physicians, but critical care is a broad, interdisciplinary specialty. The role of other clinicians' documentations in describing outcomes in the ICU setting is not known. Future work should examine documentation of family involvement generated by other clinical disciplines and other ICU settings. Finally, we used rule-based NLP models, which only detect phrases in notes if they match the specified keywords.

Conclusions

This study fills an important gap in our understanding of family involvement in patient care and decision-making early in ICU stays. Findings suggest that better decisions about LLSTs will be made if additional family members are engaged, and clinicians should seek out everyone who may want to or need to participate.

Acknowledgments

This study was funded by the Cambia Health Foundation Sojourns Scholars Leadership Program and the Robert Wood Johnson Foundation Harold Amos Medical Faculty Development Program.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Keyword library. This appendix presents the exact phrases used in the keyword bank. It includes an extensive list to account for the wide variations in describing spouse and child. For example, spouse is described as husband, wife, fiancé, girlfriend, boyfriend, companion, partner, spouse, comate, etc. Child is described as son, daughter, grandchild, teenage, girl, boy, child, children, grandson, granddaughter, etc.
[PDF File (Adobe PDF File), 155 KB - medinform_v10i6e33921_app1.pdf]

Multimedia Appendix 2
Keywords used in relevant context. This Multimedia Appendix provides examples of how the keywords found in the clinical notes were used in relevant context in the analysis.
[PDF File (Adobe PDF File), 137 KB - medinform_v10i6e33921_app2.pdf]

Multimedia Appendix 3
Keywords used in nonrelevant context. This Multimedia Appendix provides examples of how the keywords found in the clinical notes were used in a nonrelevant context and eliminated during the analysis.
[PDF File (Adobe PDF File), 31 KB - medinform_v10i6e33921_app3.pdf]

References

1. Happ MB, Tate JA. Family caregiving in critical illness: research opportunities and considerations. West J Nurs Res 2017 Sep;39(9):1219-1221. [doi: 10.1177/0193945917714760] [Medline: 28791936]
2. Curtis JR, White DB. Practical guidance for evidence-based ICU family conferences. Chest 2008 Oct;134(4):835-843 [FREE Full text] [doi: 10.1378/chest.08-0235] [Medline: 18842916]
3. Suen AO, Butler RA, Arnold R, Myers B, Witteman HO, Cox CE, et al. Developing the family support tool: an interactive, web-based tool to help families navigate the complexities of surrogate decision making in ICUs. J Crit Care 2020 Apr;56:132-139 [FREE Full text] [doi: 10.1016/j.jcrc.2019.12.002] [Medline: 31896447]
4. Silveira MJ, Kim SYH, Langa KM. Advance directives and outcomes of surrogate decision making before death. N Engl J Med 2010 Apr;362(13):1211-1218 [FREE Full text] [doi: 10.1056/NEJMsa0907901] [Medline: 20357283]

5. Bibas L, Peretz-Larochelle M, Adhikari NK, Goldfarb MJ, Luk A, Englesakis M, et al. Association of surrogate decision-making interventions for critically ill adults with patient, family, and resource use outcomes: a systematic review and meta-analysis. JAMA Netw Open 2019 Jul;2(7):e197229 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.7229] [Medline: 31322688]
6. McAdam JL, Arai S, Puntillo KA. Unrecognized contributions of families in the intensive care unit. Intensive Care Med 2008 Jun;34(6):1097-1101. [doi: 10.1007/s00134-008-1066-z] [Medline: 18369593]
7. Kruser JM, Benjamin BT, Gordon EJ, Michelson KN, Wunderink RG, Holl JL, et al. Patient and family engagement during treatment decisions in an ICU: a discourse analysis of the electronic health record. Crit Care Med 2019 Jun;47(6):784-791 [FREE Full text] [doi: 10.1097/CCM.0000000000003711] [Medline: 30896465]
8. Bruce CR, Fetter JE, Blumenthal-Barby JS. Cascade effects in critical care medicine: a call for practice changes. Am J Respir Crit Care Med 2013 Dec;188(12):1384-1385. [doi: 10.1164/rccm.201309-1606ED] [Medline: 24328766]
9. Apatira L, Boyd EA, Malvar G, Evans LR, Luce JM, Lo B, et al. Hope, truth, and preparing for death: perspectives of surrogate decision makers. Ann Intern Med 2008 Dec;149(12):861-868 [FREE Full text] [doi: 10.7326/0003-4819-149-12-200812160-00005] [Medline: 19075205]
10. Mackie BR, Mitchell M, Schults J. Application of the READY framework supports effective communication between health care providers and family members in intensive care. Aust Crit Care 2021 May;34(3):296-299 [FREE Full text] [doi: 10.1016/j.aucc.2020.07.010] [Medline: 33069591]
11. Liput SA, Kane-Gill SL, Seybert AL, Smithburger PL. A review of the perceptions of healthcare providers and family members toward family involvement in active adult patient care in the ICU. Crit Care Med 2016 Jun;44(6):1191-1197. [doi: 10.1097/CCM.0000000000001641] [Medline: 26958747]
12. Davidson JE, Powers K, Hedayat KM, Tieszen M, Kon AA, Shepard E, American College of Critical Care Medicine Task Force 2004-2005, Society of Critical Care Medicine. Clinical practice guidelines for support of the family in the patient-centered intensive care unit: American College of Critical Care Medicine Task Force 2004-2005. Crit Care Med 2007 Feb;35(2):605-622. [doi: 10.1097/01.CCM.0000254067.14607.EB] [Medline: 17205007]
13. Truog RD, Campbell ML, Curtis JR, Haas CE, Luce JM, Rubenfeld GD, American Academy of Critical Care Medicine. Recommendations for end-of-life care in the intensive care unit: a consensus statement by the American College [corrected] of Critical Care Medicine. Crit Care Med 2008 Mar;36(3):953-963. [doi: 10.1097/CCM.0B013E3181659096] [Medline: 18431285]
14. Trees AR, Ohs JE, Murray MC. Family communication about end-of-life decisions and the enactment of the decision-maker role. Behav Sci (Basel) 2017 Jun;7(2):36 [FREE Full text] [doi: 10.3390/bs7020036] [Medline: 28590407]
15. Quinn JR, Schmitt M, Baggs JG, Norton SA, Dombeck MT, Sellers CR. Family members' informal roles in end-of-life decision making in adult intensive care units. Am J Crit Care 2012 Jan;21(1):43-51 [FREE Full text] [doi: 10.4037/ajcc2012520] [Medline: 22210699]
16. Lautrette A, Peigne V, Watts J, Souweine B, Azoulay E. Surrogate decision makers for incompetent ICU patients: a European perspective. Curr Opin Crit Care 2008 Dec;14(6):714-719. [doi: 10.1097/MCC.0b013e3283196319] [Medline: 19005315]
17. Moye J, Sabatino CP, Brendel RW. Evaluation of the capacity to appoint a healthcare proxy. Am J Geriatr Psychiatry 2013 Apr;21(4):326-336 [FREE Full text] [doi: 10.1016/j.jagp.2012.09.001] [Medline: 23498379]
18. Pope TM. Comparing the FHCDA to surrogate decision making laws in other states. NYSBA Health Law Journal.: Widener Law School Legal Studies Research Paper; 2011 Mar. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1797930 [accessed 2022-05-25]
19. Pope TM. Legal fundamentals of surrogate decision making. Chest 2012 Apr;141(4):1074-1081. [doi: 10.1378/chest.11-2336] [Medline: 22474149]
20. Brody EM. "Women in the middle" and family help to older people. Gerontologist 1981 Oct;21(5):471-480. [doi: 10.1093/geront/21.5.471] [Medline: 7338304]
21. Ma JD, Benn M, Nelson SH, Campillo A, Heavey SF, Cramer A, et al. Exploring the definition of an informed health care proxy. J Palliat Med 2016 Mar;19(3):250-251. [doi: 10.1089/jpm.2015.0439] [Medline: 26836962]
22. Pinquart M, Sörensen S. Spouses, adult children, and children-in-law as caregivers of older adults: a meta-analytic comparison. Psychol Aging 2011 Mar;26(1):1-14 [FREE Full text] [doi: 10.1037/a0021863] [Medline: 21417538]
23. Heyland D, Rocker GM, Dodek PM, Kutsogiannis DJ, Konopad E, Cook DJ, et al. Family satisfaction with care in the intensive care unit: results of a multiple center study. Crit Care Med 2002 Jul;30(7):1413-1418. [doi: 10.1097/00003246-200207000-00002] [Medline: 12130954]
24. Johnson D, Wilson M, Cavanaugh B, Bryden C, Gudmundson D, Moodley O. Measuring the ability to meet family needs in an intensive care unit. Crit Care Med 1998 Feb;26(2):266-271. [doi: 10.1097/00003246-199802000-00023] [Medline: 9468163]
25. Auerbach SM, Kiesler DJ, Wartella J, Rausch S, Ward KR, Ivatury R. Optimism, satisfaction with needs met, interpersonal perceptions of the healthcare team, and emotional distress in patients' family members during critical care hospitalization. Am J Crit Care 2005 May;14(3):202-210. [Medline: 15840894]

26. Hwang DY, Yagoda D, Perrey HM, Tehan TM, Guanci M, Ananian L, et al. Assessment of satisfaction with care among family members of survivors in a neuroscience intensive care unit. J Neurosci Nurs 2014 Apr;46(2):106-116 [FREE Full text] [doi: 10.1097/JNN.0000000000000038] [Medline: 24556658]
27. Khalaila R. Patients' family satisfaction with needs met at the medical intensive care unit. J Adv Nurs 2013 May;69(5):1172-1182. [doi: 10.1111/j.1365-2648.2012.06109.x] [Medline: 22931366]
28. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
29. Chan A, Chien I, Moseley E, Salman S, Bourland SK, Lamas D, et al. Deep learning algorithms to identify documentation of serious illness conversations during intensive care unit admissions. Palliat Med 2019 Feb;33(2):187-196. [doi: 10.1177/0269216318810421] [Medline: 30427267]
30. Auriemma CL, Lyon SM, Strelec LE, Kent S, Barg FK, Halpern SD. Defining the medical intensive care unit in the words of patients and their family members: a freelisting analysis. Am J Crit Care 2015 Jul;24(4):e47-e55 [FREE Full text] [doi: 10.4037/ajcc2015717] [Medline: 26134339]
31. Lucini FR, Krewulak KD, Fiest KM, Bagshaw SM, Zuege DJ, Lee J, et al. Natural language processing to measure the frequency and mode of communication between healthcare professionals and family members of critically ill patients. J Am Med Inform Assoc 2021 Mar;28(3):541-548 [FREE Full text] [doi: 10.1093/jamia/ocaa263] [Medline: 33201981]
32. Lindvall C, Lilley EJ, Zupanc SN, Chien I, Udelsman BV, Walling A, et al. Natural language processing to assess end-of-life quality indicators in cancer patients receiving palliative surgery. J Palliat Med 2019 Feb;22(2):183-187. [doi: 10.1089/jpm.2018.0326] [Medline: 30328764]
33. Udelsman BV, Lilley EJ, Qadan M, Chang DC, Lillemoe KD, Lindvall C, et al. Deficits in the palliative care process measures in patients with advanced pancreatic cancer undergoing operative and invasive nonoperative palliative procedures. Ann Surg Oncol 2019 Dec;26(13):4204-4212. [doi: 10.1245/s10434-019-07757-2] [Medline: 31463695]
34. Udelsman BV, Moseley ET, Sudore RL, Keating NL, Lindvall C. Deep natural language processing identifies variation in care preference documentation. J Pain Symptom Manage 2020 Jun;59(6):1186-1194.e3. [doi: 10.1016/j.jpainsymman.2019.12.374] [Medline: 31926970]
35. Poort H, Zupanc SN, Leiter RE, Wright AA, Lindvall C. Documentation of palliative and end-of-life care process measures among young adults who died of cancer: a natural language processing approach. J Adolesc Young Adult Oncol 2020 Feb;9(1):100-104. [doi: 10.1089/jayao.2019.0040] [Medline: 31411524]
36. Cohen SA, Cook SK, Sando TA, Brown MJ, Longo DR. Socioeconomic and demographic disparities in caregiving intensity and quality of life in informal caregivers: a first look at the National Study of Caregiving. J Gerontol Nurs 2017 Jun;43(6):17-24. [doi: 10.3928/00989134-20170224-01] [Medline: 28253411]
37. Obringer K, Hilgenberg C, Booker K. Needs of adult family members of intensive care unit patients. J Clin Nurs 2012 Jun;21(11-12):1651-1658. [doi: 10.1111/j.1365-2702.2011.03989.x] [Medline: 22404287]
38. 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. on behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med 1996 Jul;22(7):707-710. [doi: 10.1007/BF01709751] [Medline: 8844239]
39. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998 Jan;36(1):8-27. [doi: 10.1097/00005650-199801000-00004] [Medline: 9431328]
40. Lee RY, Brumback LC, Lober WB, Sibley J, Nielsen EL, Treece PD, et al. Identifying goals of care conversations in the electronic health record using natural language processing and machine learning. J Pain Symptom Manage 2021 Jan;61(1):136-142.e2 [FREE Full text] [doi: 10.1016/j.jpainsymman.2020.08.024] [Medline: 32858164]
41. Efstathiou N, Vanderspank-Wright B, Vandyk A, Al-Janabi M, Daham Z, Sarti A, et al. Terminal withdrawal of mechanical ventilation in adult intensive care units: a systematic review and narrative synthesis of perceptions, experiences and practices. Palliat Med 2020 Oct;34(9):1140-1164. [doi: 10.1177/0269216320935002] [Medline: 32597309]
42. Cook SK, Snellings L, Cohen SA. Socioeconomic and demographic factors modify observed relationship between caregiving intensity and three dimensions of quality of life in informal adult children caregivers. Health Qual Life Outcomes 2018 Aug;16(1):169 [FREE Full text] [doi: 10.1186/s12955-018-0996-6] [Medline: 30157852]
43. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med 2015 Jan;3(1):42-52 [FREE Full text] [doi: 10.1016/S2213-2600(14)70239-5] [Medline: 25466337]
44. Udelsman B, Chien I, Ouchi K, Brizzi K, Tulsky JA, Lindvall C. Needle in a haystack: natural language processing to identify serious illness. J Palliat Med 2019 Feb;22(2):179-182 [FREE Full text] [doi: 10.1089/jpm.2018.0294] [Medline: 30251922]
45. Lin JJ, Smith CB, Feder S, Bickell NA, Schulman-Green D. Patients' and oncologists' views on family involvement in goals of care conversations. Psychooncology 2018 Mar;27(3):1035-1041. [doi: 10.1002/pon.4630] [Medline: 29315989]
46. Dilworth-Anderson P, Williams IC, Gibson BE. Issues of race, ethnicity, and culture in caregiving research: a 20-year review (1980-2000). Gerontologist 2002 Apr;42(2):237-272. [doi: 10.1093/geront/42.2.237] [Medline: 11914467]

47. Fingerman KL, VanderDrift LE, Dotterer AM, Birditt KS, Zarit SH. Support to aging parents and grown children in Black and White families. Gerontologist 2011 Aug;51(4):441-452 [FREE Full text] [doi: 10.1093/geront/gnq114] [Medline: 21199862]
48. Taylor RJ, Chatters LM. Patterns of informal support to elderly black adults: Family, friends, and church members. Soc Work 1986 Nov;31(6):432-438. [doi: 10.1093/sw/31.6.432]
49. Bullock K, Crawford SL, Tennstedt SL. Employment and caregiving: exploration of African American caregivers. Soc Work 2003 Apr;48(2):150-162. [doi: 10.1093/sw/48.2.150] [Medline: 12718411]
50. Williams SW, Dilworth-Anderson P. Systems of social support in families who care for dependent African American elders. Gerontologist 2002 Apr;42(2):224-236. [doi: 10.1093/geront/42.2.224] [Medline: 11914466]
51. Miller B, Guo S. Social support for spouse caregivers of persons with dementia. J Gerontol B Psychol Sci Soc Sci 2000 May;55(3):S163-S172. [doi: 10.1093/geronb/55.3.s163] [Medline: 11833984]
52. Chien I, Shi A, Chan A, Lindvall C. Identification of serious illness conversations in unstructured clinical notes using deep neural networks. In: Artificial Intelligence in Health. Cham: Springer; 2019 Feb Presented at: International Workshop on Artificial Intelligence in Health; July 13-14, 2018; Stockholm, Sweden p. 199-212 URL: https://doi.org/10.1007/978-3-030-12738-1_15 [doi: 10.1007/978-3-030-12738-1_15]

Abbreviations

BIDMC: Beth Israel Deaconess Medical Center
EHR: electronic health record
GOCC: goals of care conversation
ICU: intensive care unit
LLST: limitations in life-sustaining therapy
LOS: length of stay
MICU: medical intensive care unit
MIMIC: Medical Information Mart for Intensive Care
MIT: Massachusetts Institute of Technology
NLP: natural language processing
SNF: skilled nursing facility
SOFA: sequential organ failure assessment score

Edited by C Lovis; submitted 29.09.21; peer-reviewed by M Ahmed Kamal, A Azizi, M Johansson, M Tomey; comments to author 20.02.22; revised version received 01.04.22; accepted 21.04.22; published 15.06.22.

Please cite as:
Gray TF, Kwok A, Do KM, Zeng S, Moseley ET, Dbeis YM, Umeton R, Tulsky JA, El-Jawahri A, Lindvall C
Associations Between Family Member Involvement and Outcomes of Patients Admitted to the Intensive Care Unit: Retrospective Cohort Study
JMIR Med Inform 2022;10(6):e33921
URL: https://medinform.jmir.org/2022/6/e33921
doi:10.2196/33921
PMID:35704362

Original Paper

Automatic International Classification of Diseases Coding System: Deep Contextualized Language Model With Rule-Based Approaches

Pei-Fu Chen[1,2*], MD; Kuan-Chih Chen[1,3*], MD, MSc; Wei-Chih Liao[1], MSc; Feipei Lai[1,4,5], PhD; Tai-Liang He[4], BSc; Sheng-Che Lin[4], BSc; Wei-Jen Chen[1], BSc; Chi-Yu Yang[6,7], MD; Yu-Cheng Lin[8,9], MD, PhD; I-Chang Tsai[10], PhD; Chi-Hao Chiu[11], MS; Shu-Chih Chang[12], MA; Fang-Ming Hung[13,14], MD

[1]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan
[2]Department of Anesthesiology, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[3]Department of Internal Medicine, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[4]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[5]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
[6]Department of Information Technology, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[7]Section of Cardiovascular Medicine, Cardiovascular Center, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[8]Department of Pediatrics, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[9]Department of Healthcare Administration, Oriental Institute of Technology, New Taipei City, Taiwan
[10]Artificial Intelligence Center, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[11]Section of Health Insurance, Department of Medical Affairs, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[12]Medical Records Department, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[13]Department of Medical Affairs, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[14]Department of Surgical Intensive Care Unit, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[*]these authors contributed equally

Corresponding Author:
Fang-Ming Hung, MD
Department of Medical Affairs
Far Eastern Memorial Hospital
No. 21, Sec. 2, Nanya S. Rd., Banciao Dist.
New Taipei City, 220216
Taiwan
Phone: 886 2 8966 7000
Fax: 886 2 8966 5567
Email: drphilip101@gmail.com

Abstract

Background: The tenth revision of the International Classification of Diseases (ICD-10) is widely used for epidemiological research and health management. The clinical modification (CM) and procedure coding system (PCS) of ICD-10 were developed to describe more clinical details with increasing diagnosis and procedure codes and applied in disease-related groups for reimbursement. The expansion of codes made the coding time-consuming and less accurate. The state-of-the-art model using deep contextual word embeddings was used for automatic multilabel text classification of ICD-10. In addition to input discharge diagnoses (DD), the performance can be improved by appropriate preprocessing methods for the text from other document types, such as medical history, comorbidity and complication, surgical method, and special examination.

Objective: This study aims to establish a contextual language model with rule-based preprocessing methods to develop the model for ICD-10 multilabel classification.

Methods: We retrieved electronic health records from a medical center. We first compared different word embedding methods. Second, we compared the preprocessing methods using the best-performing embeddings. We compared biomedical bidirectional encoder representations from transformers (BioBERT), clinical generalized autoregressive pretraining for language understanding (Clinical XLNet), label tree-based attention-aware deep model for high-performance extreme multilabel text classification

(AttentionXLM), and word-to-vector (Word2Vec) to predict ICD-10-CM. To compare different preprocessing methods for ICD-10-CM, we included DD, medical history, and comorbidity and complication as inputs. We compared the performance of ICD-10-CM prediction using different preprocesses, including definition training, external cause code removal, number conversion, and combination code filtering. For the ICD-10 PCS, the model was trained using different combinations of DD, surgical method, and key words of special examination. The micro F_1 score and the micro area under the receiver operating characteristic curve were used to compare the model's performance with that of different preprocessing methods.

Results: BioBERT had an F_1 score of 0.701 and outperformed other models such as Clinical XLNet, AttentionXLM, and Word2Vec. For the ICD-10-CM, the model had an F_1 score that significantly increased from 0.749 (95% CI 0.744-0.753) to 0.769 (95% CI 0.764-0.773) with the ICD-10 definition training, external cause code removal, number conversion, and combination code filter. For the ICD-10-PCS, the model had an F_1 score that significantly increased from 0.670 (95% CI 0.663-0.678) to 0.726 (95% CI 0.719-0.732) with a combination of discharge diagnoses, surgical methods, and key words of special examination. With our preprocessing methods, the model had the highest area under the receiver operating characteristic curve of 0.853 (95% CI 0.849-0.855) and 0.831 (95% CI 0.827-0.834) for ICD-10-CM and ICD-10-PCS, respectively.

Conclusions: The performance of our model with the pretrained contextualized language model and rule-based preprocessing method is better than that of the state-of-the-art model for ICD-10-CM or ICD-10-PCS. This study highlights the importance of rule-based preprocessing methods based on coder coding rules.

(JMIR Med Inform 2022;10(6):e37557) doi:10.2196/37557

KEYWORDS

deep learning; International Classification of Diseases; medical records; multilabel text classification; natural language processing; coding system; algorithm; electronic health record; data mining

Introduction

Background

The International Classification of Diseases (ICD) aims to systematically record, analyze, interpret, and compare mortality and morbidity data collected in different areas. ICD transforms the diagnosis of diseases and other health problems from text to alphanumeric codes, which are mixed with English letters and numbers [1]. ICD has become an internationally accepted diagnostic classification system for epidemiological research and health management.

The World Health Organization (WHO) introduced the tenth revision of the International Classification of Diseases (ICD-10) in the 1990s to accommodate the increasing number of diagnoses and related health problems [1]. The clinical modification (CM) and procedure coding system (PCS) of ICD-10 (ICD-10-CM and ICD-10-PCS) have been developed to describe more clinical details with increasing diagnosis and procedure codes and applied in payment methodologies, such as disease-related groups in the United States [2,3]. The transition from ICD-9 to ICD-10-CM or ICD-10-PCS expanded the number of codes. There are only approximately 14,000 diagnosis codes and 3800 procedure codes in ICD-9, but approximately 69,000 in ICD-10-CM and 72,000 in ICD-10-PCS [3]. The expanded codes suppress productivity and increase the cost of disease coding [4]. In practice, the disease coder spent more time interpreting the text of the medical records to ensure the correctness of the disease [4].

The speed and correctness of the classification of the disease coder will be affected by incomplete medical records, orders of diagnosis, undetailed surgical findings, and fragmented exam reports. In addition, hospitals must increase their accuracy in terms of reimbursement. The research found that income can be increased by approximately 5% with a clinician-auditor review in patients discharged following an emergency admission [5].

Related Work

In recent years, text classification from electronic health records (EHR) data has been widely studied in natural language processing [6], which is a subdiscipline in the fields of artificial intelligence and linguistics. This field explores how to process and use natural language by computers into meaningful representations and maintain the relationships of meanings according to the purpose [7]. Text classification can be divided into the 3 categories of binary, multiclass, and multilabel. Among these, multilabel text classification outputs multiple labels with one or more classes. The multilabel classification task is more challenging because the number of possible combinations of results is greater if the label set is larger.

Teng et al [8] recently proposed a model predicting ICD-10-CM using a medical topic mining method and a cross-textual attentional neural network. It had an F_1 score of 0.96 in a single label of "atrial fibrillation." However, even with the same methods proposed to predict the top 50 most frequent ICD-10-CM codes, their model had an F_1 score of 0.68. This shows that multilabel classification is more complicated than single-label classification. Multilabel classification for ICD-10-PCS is even more challenging owing to its sparsity. Subotin et al [9] proposed a model with code co-occurrence propensity, which improved the prediction of ICD-10-PCS with an F_1 score from 0.50 to 0.56.

Previous Work

To facilitate the laborious and time-consuming work process, we have shown that the ICD-10 autocoding system achieved an F_1 score of 0.67 and 0.58 in CM and PCS by applying

word-to-vector (Word2Vec) [10]. Furthermore, we achieved a better F_1 score of 0.72 and 0.62 in CM and PCS through bidirectional encoder representations from transformers (BERT). In addition, an attention mechanism was used in this classification model to visualize the importance of words used to train new disease coders [11].

In our previous work, some problems were encountered, such as handling the following issues. Some meaningful numbers used in medical terms were removed from the data sets in the preprocessing stage. The combination codes comprising 2 diagnoses in 1 code were hard to be predicted. Other than discharge diagnoses, information from the discharge records was not efficiently included, such as medical history, comorbidity, and complication. In addition, because the writing of medical records was different from the original ICD-10-CM code definition, training our model with the ICD-10-CM definition may be helpful.

Surgical method records and special examination reports are helpful for disease coders to determine the ICD-10-PCS. However, information from special examination reports is challenging to be extracted because it is mixed with uninformative content, such as ultrasound, radiology, endoscopy, and electroencephalography. Furthermore, information from surgical method records is also essential, but the combination algorithm for these types of documents should be studied.

Objective

This study focuses on interpreting medical records to tackle the problems mentioned above because we found that the accuracy is limited without a rule-based approach. We propose that we can make our model more accurate by adopting coding rules from experienced disease coders in our preprocess. Therefore, this study aims to establish a contextual language model with rule-based preprocessing methods to develop a more accurate and explainable ICD-10 autocoding system.

Methods

Ethical Considerations

This retrospective study was approved by the institutional review board of the Far Eastern Memorial Hospital (109086-F and 110028-F), which waived the requirement for informed consent.

Data Collection

Data were acquired from the electronic medical records of the Far Eastern Memorial Hospital, a medical center in Taiwan, from January 2018 to December 2020. The collected data included admission date, discharge date, discharge summary, ICD-10-CM codes, and ICD-10-PCS codes. The ground-truth ICD-10-CM or ICD-10-PCS codes were labeled by the disease coders.

Data Description

We obtained 101,974 documents for ICD-10-CM codes and 105,466 documents for ICD-10-PCS codes. Our discharge summary contains 5 types of documents. The discharge diagnoses (DD) listed the main diagnoses related to this hospitalization. The surgical method (SM) includes a description of the surgical procedures and findings. The special examination (SE) includes ultrasound, radiological, endoscopic, and electroencephalography reports. Medical history (MH) contains the process of developing the present illness and the past medical history. Comorbidity and complications (CC) included complications noted during hospitalization.

Most of these studies included CC and MH (Figure 1). The count of the 3 types of documents in each chapter of the ICD-10-CM and ICD-10-PCS are shown in Multimedia Appendix 1. The chapters were determined by the first 3 codes of the ICD-10 labels annotated by disease coders. The maximal word count was up to 2342 in SE, and the mean word count was up to 149 in MH (Table 1).

Figure 1. Data counts of 5 types of documents.

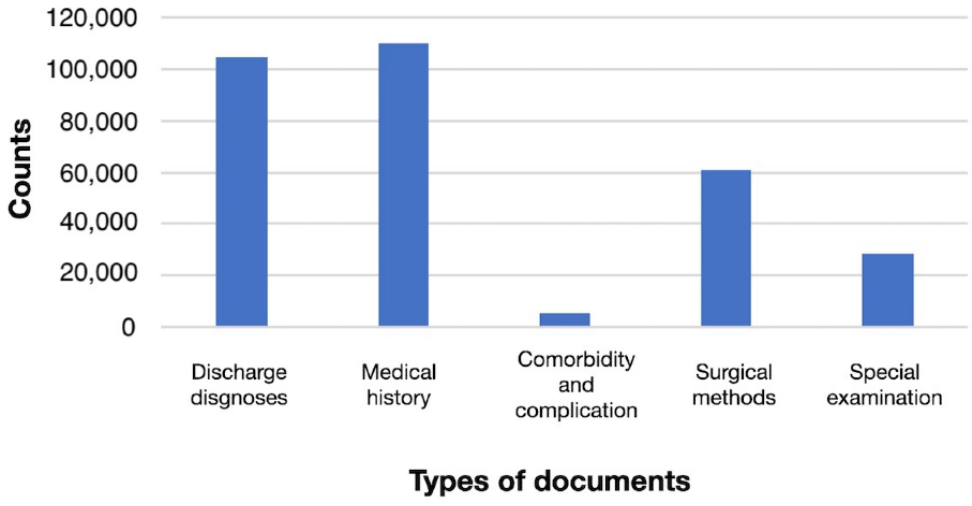


Table 1. Word counts of 5 types of documents.

Document type	Maximal word count	Mean word count
Discharge diagnoses	480	31
Surgical method	487	11
Special examination	2342	86
Medical history	586	149
Comorbidity and complication	338	5

Common Text Preprocessing

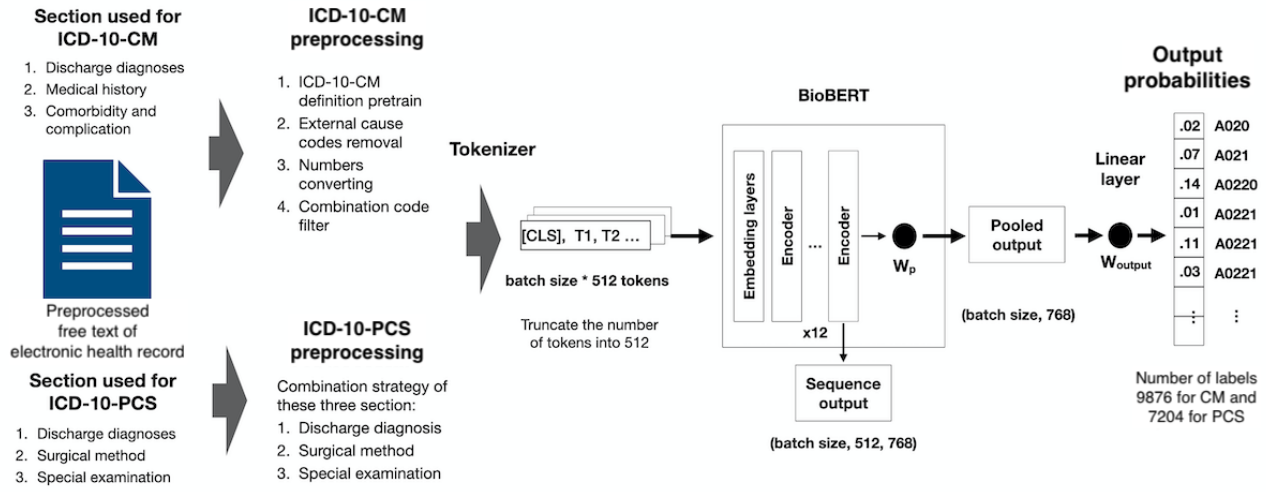
Null or duplicate data sets and punctuation were removed using the Natural Language Toolkit [12]. Non-English characters were removed before further preprocessing. The text in our EHR was written in mixed English and Chinese. The Chinese part contains the names of the people, places, special customs, and transferred hospital, and is irrelevant to the diagnosis.

Study Design

We first compared different word embedding methods. Second, we compared the preprocessing methods using the best-performing word embedding methods. To choose the best-performing embeddings, we compared the performance of Word2Vec [13], label tree-based attention-aware deep model for high-performance extreme multilabel text classification (AttentionXLM) [14], biomedical BERT (BioBERT) [15], and clinical generalized autoregressive pretraining for language understanding (clinical XLNet) [16] to predict ICD-10-CM with DD as input. BioBERT had the highest F_1 score and was chosen to compare the following preprocessing methods for ICD-10-CM or ICD-10-PCS (Multimedia Appendix 2).

The sections used for predicting ICD-10-CM were DD, MH, and CC; the sections used for predicting ICD-10-PCS were DD, SM, and SE. The concatenated input text from these sections was long and contained fewer informative components. A proper preprocessing method should be designed to extract helpful information from text. We randomly split the data in a 9:1 ratio into training and validation sets. After the model was trained with the training set, the validation set was used to compare the effects of the following preprocessing methods: the change in the model performance of the trained definition, external cause code removal, number conversion, and combination code filter, which are shown for ICD-10-CM stepwise. The model performance of inputting different document section combinations was compared for ICD-10-PCS, including DD, SM, and SE (Figure 2).

Figure 2. Data processing flow chart and the model architecture. BioBERT: bidirectional encoder representations from transformers for biomedical text mining. CLS: classification; CM: clinical modification; ICD: International Classification of Diseases; PCS: procedure coding system; T: token; Woutput: output weight; Wp: pooled weight.



Model Architecture

After preprocessing, the text was tokenized using the BERT tokenizer. The tokens for BioBERT were truncated to 512 in length because of the model limit [15]. Tokens are then inputted into the BioBERT. A linear layer was connected to the pooled output of BioBERT with labels. The labels are one-hot encodings of all individual ICD-10-CM or ICD-10-PCS codes in our data set, which are 9876 for CM and 7204 for PCS (Figure 2). We calculated the loss by cross entropy. We trained the model using the Adam optimizer and a learning rate of 0.00005 until 100 epochs or met the early stop criteria (less than 0.0001 changes for 10 epochs).

Data Preprocessing for ICD-10-CM

We included DD, MH, and CC to train the model for ICD-10-CM. We designed a process to include helpful information and remove less informative content. This process

contains several components, including the following: MH extraction, CC combining, ICD-10-CM definition training, external cause code removal, number conversion, and combination code filter. The effects of adding the ICD-10-CM definition, external cause code removal, number conversion, and combination code filter on the model performance were compared with the performance before adding these processes.

Medical History

We included the MH to extract chronic diseases not mentioned in the DD because we found that some chronic diseases, such as hypertension or chronic kidney disease, were not recorded in approximately 15% of DD in our data. Because the mean length of MH is 5 times that of DD (Table 1), we only extracted key words from MH instead of directly merging DD and MH. We listed these key words and their ICD-10-CM codes in Multimedia Appendix 3. These key words were produced after discussions with disease coders. Only key words found in the text in the MH will be retained for combination after the key word extractor is used.

Comorbidity and Complication Combining

Although CC is null in smoothly discharged patients, it affects the ICD-10-CM code if it is not null. ICD-10-CM codes that are frequently inferred from CC include nausea, vomiting, diarrhea, fatigue, and pneumonia. The mean length of the CC was only one-sixth of the DD (Table 1), and thus we combined DD with CC directly.

ICD-10-CM Definition Trained

We initiated our model with weights from BioBERT and trained the model on the official ICD-10-CM definition by the WHO as the input and the respective ICD-10-CM code as the output [1]. The model was trained for 100 epochs with early stop criteria (less than 0.0001 changes for 10 epochs). For example, if the output ICD-10-CM code is N39.0, the input text is “urinary tract infection, site not specified.”.

External Cause Codes Removal

External cause codes (V01-Y98) define environmental events, circumstances, and conditions, such as the cause of injury, poisoning, and other adverse effects related to an injury. However, it is challenging for a model to predict external cause codes because relevant information is seldom recorded. Because external cause codes do not affect the final disease-related group payment, we removed them from our labels.

Number Converting

There are numbers in our EHR, such as the date of the MH, the report’s physiological value, and the header of each line. They were removed because most of them were not informative for our classification task. However, we found that some numbers may affect the ICD-10-CM or ICD-10-PCS prediction, such as pregnancy weeks (“36 weeks gestation of pregnancy”), stage of chronic diseases (“stage 4 chronic kidney disease”), type of disease (“type 2 diabetes mellitus”), and grade of disease (“follicular lymphoma grade 1” and “modified Rankin scale 0”). Thus, we converted all the known essential numbers back to alphabets, such as “stage four chronic kidney disease,” “type two diabetes mellitus,” and “thirty-six weeks gestation of pregnancy,” before removing all numbers.

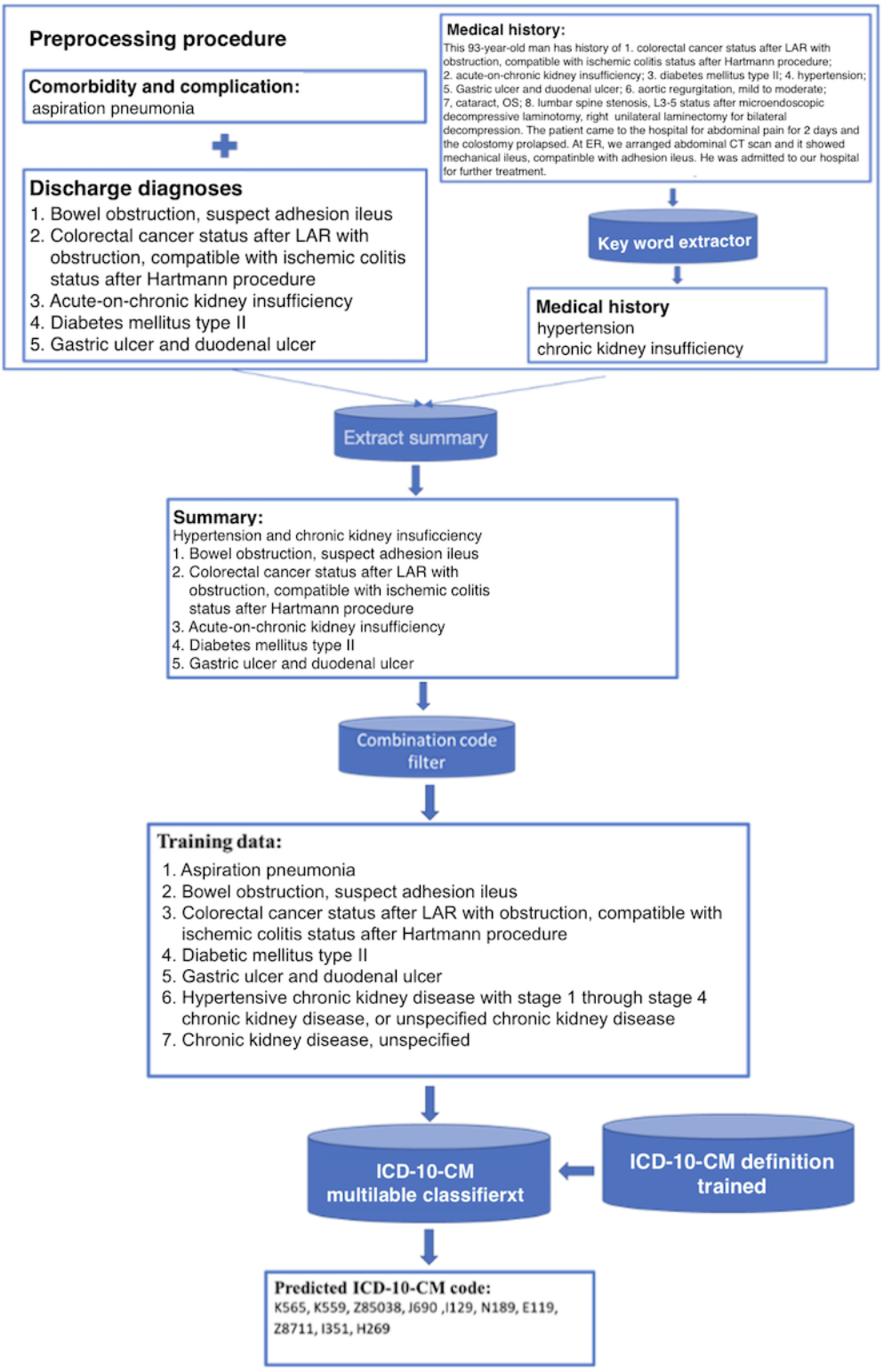
Combination Code Filter

A combination code represents the diagnosis of one or more comorbidities. For example, hypertension with various comorbidities refers to different combinations of codes. To solve these problems, we designed a combination code filter (Multimedia Appendix 4). If the input text contains “hypertension,” it will check whether this case has chronic kidney disease and heart failure. If yes, the combination code filter replaces the original text with the definition of the combination code. In this manner, we prevented the model from providing 2 codes instead of using combination codes.

Illustrating Preprocessing for Models Predicting ICD-10-CM

An example of preprocessing the input data for the models predicting ICD-10-CM is shown in Figure 3. After number conversion, we combined DD with extracted key words from MH, such as “hypertension” and “chronic kidney insufficiency,” into the extract summary. We then transformed the summary using a combination code filter into the training data. We first trained our model using the ICD-10-CM definition and then trained it on the training data.

Figure 3. Data preprocessing framework of ICD-10-CM classification model. CM: clinical modification; CT: computed tomography; ER: emergency room; ICD: International Classification of Diseases; L: lumbar; LAR: low anterior resection; OS: oculus sinister.



Data Preprocessing for ICD-10-PCS

We included DD, SM, and SE to train the model for the ICD-10-PCS. In addition to DD, SM and SE provide helpful information for determining ICD-10-PCS. We trained the model with DD alone, SM alone, and 3 strategies for combining DD with SM and SE, and then compared their performances.

Surgical Method

The mean length of SM was one-third of that of DD (Table 1). SM was recorded only if the patient underwent major procedures. To extract the most helpful information for training our model, we proposed a combination of DD and SM.

Special Examination

The mean length of SE was 3 times that of DD (Table 1). In an SE report, not all examinations will have the corresponding ICD-10-PCS codes, such as radiological examination or electroencephalography. Therefore, these components should be removed accordingly.

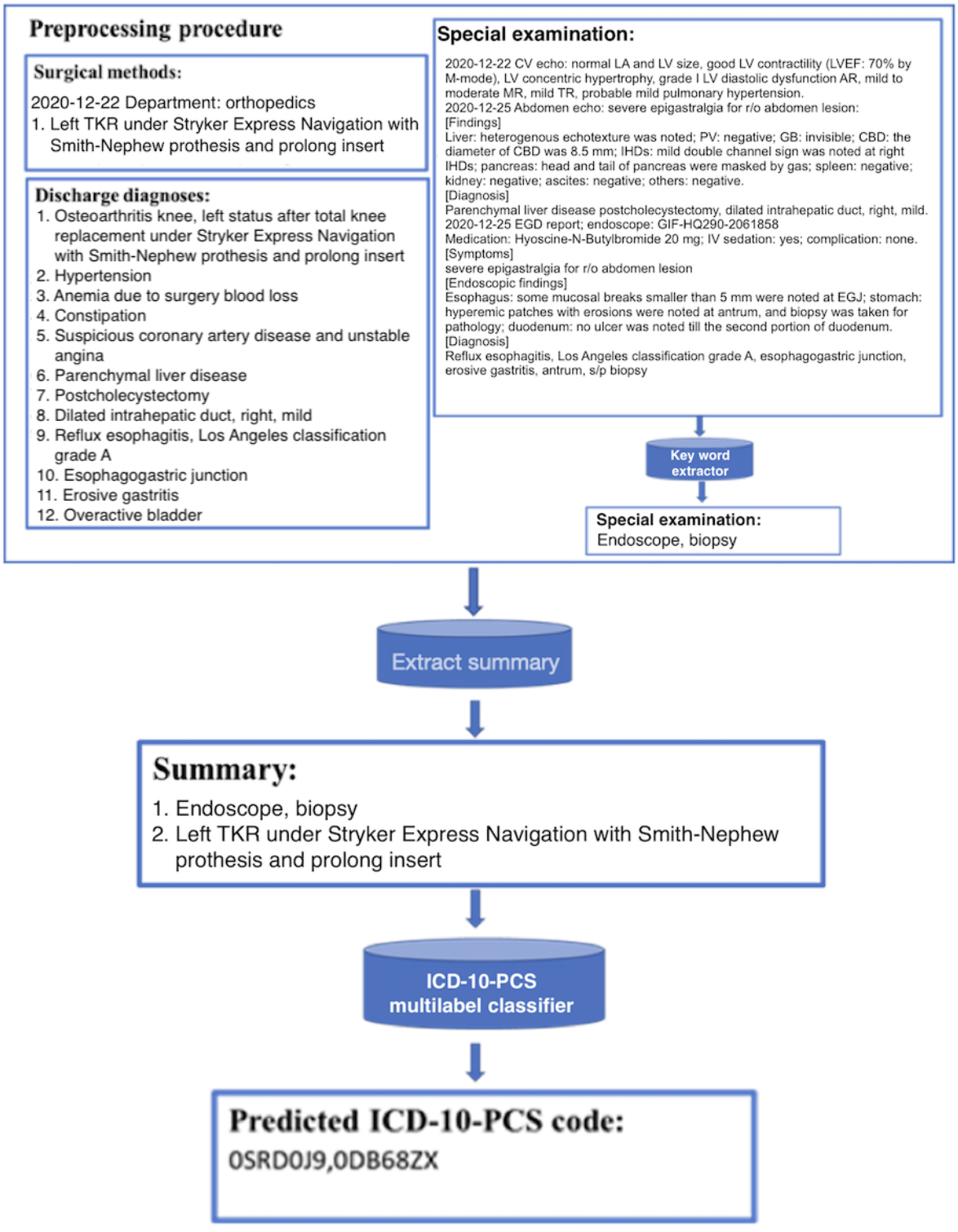
We designed a key word extractor to extract helpful information from SE and to avoid excessive text length. We listed these key words and their ICD-10-PCS codes from high to low frequency in Multimedia Appendix 5. These key words were produced by a discussion with the disease coders. Only key words found in the text in the SE were retained after the key word extractor was used.

After extracting the key words from the SE, we used 2 different combination strategies. First, we input the DD only if the patient has no SM or SE. In the second method, we input the DD if the patient had no SM and added key words from the SE.

Illustrating Preprocessing for Models Predicting ICD-10-PCS

An example of preprocessing the input data for models predicting ICD-10-PCS is shown in Figure 4. We first combined DD with extracted key words from SE, such as “endoscope” and “biopsy,” into the extract summary. We then trained our model on these data to predict ICD-10-PCS.

Figure 4. Data preprocessing framework of ICD-10-PCS classification model. AR: aortic regurgitation; CAD: coronary arterial disease; CBD: common bile duct; CV: cardiovascular; EGD: esophagogastroduodenoscopy; EGJ: esophago-gastric junction; GB: gall bladder; ICD: International Classification of Diseases; IHD: intrahepatic duct; IV, intravenous; LA: left atrium; LV: left ventricle; LVEF: left ventricular ejection fraction; MR: mitral regurgitation; PCS: procedure coding system; PV: portal vein; R/O: rule out; s/p: status post; TKR: total knee replacement; TR: tricuspid regurgitation.



Preprocessing for ICD-10-CM Label Classification

To compare different preprocessing methods for ICD-10-CM, we included DD, MH, and CC as inputs. We compared the performance of ICD-10-CM prediction using different preprocesses, including definition training, external cause code removal, number conversion, and combination code filtering.

Preprocessing for ICD-10-PCS Label Classification

In the ICD-10-PCS part of this study, DD, SM, and SE were included as inputs. We compared the prediction performance of the input text, including only DD, SM, and the 3 combination strategies. Combination strategy 1, “SM or DD”—we input the DD only if the case has no SM. Combination strategy 2,

"(SM+SE) or DD"—we input the DD only if the case has no SM or SE. Combination strategy 3, "(SM+SE) or (CD+SE)"—we only input DD if the case has no SM and add key words of SE.

Evaluation Metrics

Microprecision is the summation of true positives divided by the summation of all predicted positive cases (Formula 1). Microrecall is the summation of true positives divided by the summation of all actual positive cases (Formula 2). The micro F_1 score is the harmonic mean of the microrecall and microprecision, and it is an overall measure of the quality of a classifier's predictions (Formula 3). The area under the receiver operating characteristic curve (AUROC) was calculated by taking the true-positive rate against the false-positive rate. The micro-average calculates the metrics globally by considering each element of the label indicator matrix as a label. We chose the micro F_1 score and micro-AUROC to compare the model performance. The F_1 score, precision, recall, and AUROC are bootstrapped 100 times to calculate the 95% confidence interval.

Results

ICD-10-CM Label Classification

In our ICD-10-CM multilabel text classification task, each case contained approximately 1 to 20 codes from A00 to Z99. The label set was 9876 in the CM. In the comparison of different embedding models, BioBERT, Clinical XLNet, AttentionXLM, and Word2Vec had the F_1 score of 0.701, 0.685, 0.654, and 0.651, respectively. The BioBERT model had the highest F_1 score and was selected for the following experiment. Table 2 shows a comparison of the different preprocessing methods for the ICD-10-CM. The baseline model had a micro F_1 score of 0.749 (95% CI 0.744-0.753). After the model was trained with the definition, it had an F_1 score of 0.759 (95% CI 0.754-0.763). After removing the external cause codes, converting the number to the alphabet, and applying a combination code filter, the model had an F_1 score of 0.763 (95% CI 0.759-0.767), 0.767 (95% CI 0.761-0.772), and 0.769 (95% CI 0.764-0.773), respectively. The baseline model had the AUROC of 0.839 (95% CI 0.835-0.842). With all the preprocessing methods used, the model had an AUROC of 0.858 (95% CI 0.849-0.855).

Table 2. Comparison of different preprocessing methods for BioBERT[a] model on ICD[b]-10-CM[c]. Preprocessing methods are added one by one and 95% CIs are calculated by bootstrapping.

Preprocessing method	Micro F_1 score (95% CI)	Microprecision (95% CI)	Microrecall (95% CI)	AUROC[d] (95% CI)
Baseline	0.749 (0.744-0.753)	0.836 (0.832-0.840)	0.678 (0.672-0.684)	0.839 (0.835-0.842)
+Trained with definition	0.759 (0.754-0.763)	0.833 (0.829-0.838)	0.696 (0.690-0.702)	0.848 (0.845-0.851)
+External cause codes removal	0.763 (0.759-0.767)	0.843 (0.840-0.846)	0.697 (0.691-0.702)	0.849 (0.846-0.851)
+Number converting	0.767 (0.761-0.772)	0.845 (0.840-0.849)	0.702 (0.695-0.708)	0.851 (0.847-0.854)
+Combination code filter	0.769 (0.764-0.773)	0.845 (0.841-0.850)	0.706 (0.699-0.711)	0.853 (0.849-0.855)

[a]BioBERT: bidirectional encoder representations from transformers for biomedical text mining.

[b]ICD: International Classification of Diseases.

[c]CM: clinical modification.

[d]AUROC: area under the receiver operating characteristic curve.

ICD-10-PCS Label Classification

In our ICD-10-PCS multilabel text classification task, each case contained approximately 1-20 codes. The label set was 7204 in the PCS. Table 3 shows a comparison of different input document combinations for the ICD-10-PCS. The models trained with only DD and SM had an F_1 score of 0.670 (95% CI 0.663-0.678) and 0.618 (95% CI 0.607-0.627), respectively. The model trained with combination strategies 1 (SM or DD), 2 ([SM+SE] or DD), and 3 ([SM+SE] or [DD+SE]) had an F_1 score of 0.714 (95% CI 0.708-0.721), 0.724 (95% CI 0.718-0.730), and 0.726 (95% CI 0.719-0.732), respectively. The models trained with only DD had the AUROC of 0.800 (95% CI 0.796-0.805). With combination strategy 3, the model had the highest AUROC of 0.831 (95% CI 0.827-0.834).

Table 3. Comparison of different preprocessing methods for BioBERT[a] model on ICD[b]-10-PCS[c]. The 95% CIs are calculated by bootstrapping.

Preprocessing method	Micro F_1 score (95% CI)	Microprecision (95% CI)	Microrecall (95% CI)	AUROC[d] (95% CI)
DD[e]	0.670 (0.663-0.678)	0.756 (0.750-0.761)	0.601 (0.593-0.610)	0.800 (0.796-0.805)
SM[f]	0.618 (0.607-0.627)	0.750 (0.741-0.762)	0.524 (0.512-0.534)	0.762 (0.756-0.767)
SM or DD	0.714 (0.708-0.721)	0.790 (0.784-0.791)	0.651 (0.644-0.660)	0.826 (0.822-0.830)
(SM+SE[g]) or DD	0.724 (0.718-0.730)	0.801 (0.794-0.808)	0.661 (0.654-0.668)	0.830 (0.827-0.834)
(SM+SE) or (DD+SE)	0.726 (0.719-0.732)	0.803 (0.797-0.810)	0.661 (0.654-0.669)	0.831 (0.827-0.834)

[a]BioBERT: bidirectional encoder representations from transformers for biomedical text mining.

[b]ICD: International Classification of Diseases.

[c]PCS: procedure coding system.

[d]AUROC: area under the receiver operating characteristic curve.

[e]DD: discharge diagnoses.

[f]SM: surgical method.

[g]SE: special examination.

Discussion

Principal Findings

In our study of the multilabel text classification of ICD-10-CM or ICD-10-PCS, each case contained 1-20 codes, and the label set contained up to 9876 and 7204 in CM and PCS, respectively. In our previous study, the model had an F_1 score of 0.71 and 0.62 in ICD-10-CM and ICD-10-PCS [11]. In this study, we proposed preprocessing methods for ICD-10-CM and ICD-10-PCS, respectively. For the ICD-10-CM, the model had a significant F_1 score increase from 0.749 (95% CI 0.744-0.753) to 0.769 (95% CI 0.764-0.773) and a significant AUROC increase from 0.839 (95% CI 0.835-0.842) to 0.853 (95% CI 0.849-0.855). For the ICD-10-PCS, the model had an F_1 score that significantly increased from 0.670 (95% CI 0.663-0.678) to 0.726 (95% CI 0.719-0.732) and an AUROC that significantly increased from 0.800 (95% CI 0.796-0.805) to 0.831 (95% CI 0.827-0.834).

In our comparison of different word embedding methods for ICD-10-CM classification, BioBERT achieved the highest F_1 score of 0.701 among all embedding methods. This result is consistent with previous research that contextualized representations (BERT and XLNet) showing consistent improvement over noncontextualized models (Word2Vec and AttentionXLM) in multilabel text classification tasks [17]. BioBERT was pretrained on PubMed abstracts and PubMed Central full-text articles to improve the performance of biomedical text-mining tasks [15]. Previous studies confirmed that BioBERT outperformed other embedding methods in classifying ICD-10-CM [11,18].

Training the model with the ICD-10-CM definition increased its F_1 score from 0.749 to 0.759 (1.3%). Each ICD-10-CM code has a textual description of the definition on the WHO website [1]. Although the text in medical records is different from the WHO's definition, its semantics should approximate that definition. The results showed that training with definition increased the model performance for the multilabel classification of clinical text. External cause code removal increases the model's F_1 score from 0.759 to 0.763 (0.5%). The improvement is limited because external cause codes only accounted for 2.73% (2787/101,974) of our cases.

The number conversion increased the model's F_1 score from 0.763 to 0.767 (0.5%). Number converting affected 33.3% (33,978/101,974) of our cases. Retaining informative numbers such as disease type, grade, stages, and pregnancy weeks helps the model learn the relation of these numbers to the different codes. For example, there were differences between type 1 diabetes mellitus (E10) and type 2 diabetes mellitus (E11), follicular lymphoma grades I (C82.0) and II (C82.1), chronic kidney disease stages 1 (N18.1) and 4 (N18.4), and full-term uncomplicated delivery (O80) and preterm delivery (060). The combination code filter increases the model's F_1 score from 0.767 to 0.769 (0.2%). The rules of the combination code are challenging to learn through machine learning because this text may be linked to 2 different codes instead of 1 combination code. With all preprocessing methods, the F_1 score increased from 0.749 to 0.769 (2.6%). Our result is better than the state-of-the-art model of ICD-10-CM with an F_1 score of 0.68 [8] because we designed a key word extractor and trained our model with ICD-10-CM definition, external cause code removal, number conversion, and combination code filter.

The trained model had the F_1 score of 0.670 and 0.618 for DD and SM, respectively. DD is more informative for predicting ICD-10-PCS than SM when used alone. However, the model trained using combination strategy 1 (SM or DD) had an F_1 score of 0.714. The F_1 score was 6.6% and 15.5% higher than that of DD alone and SM alone, respectively. The F_1 score of the model trained with SM alone was lower than that of the model trained with DD alone because only 58% (60,558/104,411) of the cases had SM compared to cases with DD. If a patient underwent surgery, the ICD-10-PCS codes were coded according to the SM records. The model trained with combination strategies 2 ([SM+SE] or DD) and 3 ([SM+SE] or [DD+SE]) had an F_1 score of 0.724 and 0.726, respectively. Their F_1 scores were 1.4% and 1.7% higher than those of Strategy 1. Adding SE to SM or DD is effective in improving

the model performance because several ICD-10-PCS codes are coded according to ultrasound or endoscopic reports in SM. Our result is better than the state-of-the-art model of ICD-10-PCS with an F_1 score of 0.56 [9] because we designed a key word extractor and combined DD with SM and SE.

Limitations

Our study had some limitations. First, the data were obtained from a single medical center. Writing habits and disease prevalence may vary between hospitals. Different purposes of coding in different areas may also affect the labels. External validation should be conducted in future studies. Second, although we attempted to include most of the content from the health record, other parts may also contribute to the prediction, such as problem lists and progress notes. Further studies are required to manage these issues.

Conclusions

ICD-10-CM and ICD-10-PCS codes are widely applied in surveillance, clinical research, and reimbursement. Because of the complexity of ICD-10-CM and ICD-10-PCS, it takes approximately 40.4 min for a record to be coded into ICD-10-CM or ICD-10-PCS manually [2]. This study proposed a model with a combination of a pretrained contextualized language model and rule-based preprocessing methods that outperformed the state-of-the-art models in predicting ICD-10-CM or ICD-10-PCS. This study highlights the importance of rule-based preprocessing methods based on coder coding rules. In EHR, other documents are read manually to determine ICD-10-CM or ICD-10-PCS codes, such as radiology reports, laboratory data, and the problem list. An effective preprocessing method to include documents can be studied in the future.

Acknowledgments

This study was supported by grants from the Ministry of Science and Technology, Taiwan (Grant MOST 110-2634-F-002-032-) and the Far Eastern Memorial Hospital, Taiwan (Grant FEMH-2021-C-056). The sponsors had no role in the study design, data collection and analysis, publication decisions, or manuscript drafting.

Authors' Contributions

FL and WCL designed the study. WCL, TLH, and SCL designed and developed the system. PFC, KCC, CYY, YCL, ICT, CHC, SCC, and FMH collected the data. PFC and KCC conducted the experiments. WCL and PFC conducted the statistical analyses. PFC, KCC, and WCL drafted the manuscript. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Counts of types of documents in each chapter of ICD-10-CM and procedure coding system (PCS). CM: clinical modification; ICD: International Classification of Diseases.
[DOCX File , 185 KB - medinform_v10i6e37557_app1.docx]

Multimedia Appendix 2
Comparing performance and hyperparameters of different embedding models.
[DOCX File , 124 KB - medinform_v10i6e37557_app2.docx]

Multimedia Appendix 3
ICD-10-CM codes with key words in medical history. CM: clinical modification; ICD: International Classification of Diseases.
[DOCX File , 25 KB - medinform_v10i6e37557_app3.docx]

Multimedia Appendix 4
Hypertension-related combination code and amount.
[DOCX File , 16 KB - medinform_v10i6e37557_app4.docx]

Multimedia Appendix 5
ICD-10-PCS codes with key words in special examination. ICD: International Classification of Diseases; PCS: procedure coding system.
[DOCX File , 18 KB - medinform_v10i6e37557_app5.docx]

References

1. International Classification of Diseases, 10th Revision. World Health Organization. 2015. URL: https://icd.who.int/browse10/2015/en [accessed 2021-08-04]

2. Steindel SJ. International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. J Am Med Inform Assoc 2010 May 01;17(3):274-282. [doi: 10.1136/jamia.2009.001230] [Medline: 20442144]
3. Mills R, Butler R, McCullough E, Bao M, Averill R. Impact of the transition to ICD-10 on Medicare inpatient hospital payments. Medicare Medicaid Res Rev 2011 Jun 06;1(2):E1-E13. [doi: 10.5600/mmrr.001.02.a02] [Medline: 22340773]
4. Kusnoor, Blasingame MN, Williams AM, DesAutels SJ, Su J, Giuse NB. A narrative review of the impact of the transition to ICD-10 and ICD-10-CM/PCS. JAMIA Open 2020 Apr;3(1):126-131. [doi: 10.1093/jamiaopen/ooz066] [Medline: 32607494]
5. Nouraei SAR, Virk JS, Hudovsky A, Wathen C, Darzi A, Parsons D. Accuracy of clinician-clinical coder information handover following acute medical admissions: implication for using administrative datasets in clinical outcomes management. J Public Health (Oxf) 2016 Jun 23;38(2):352-362. [doi: 10.1093/pubmed/fdv041] [Medline: 25907271]
6. Pivovarov, Elhadad N. Automated methods for the summarization of electronic health records. J Am Med Inform Assoc 2015 Sep;22(5):938-947. [doi: 10.1093/jamia/ocv032] [Medline: 25882031]
7. Chowdhury GG. Natural language processing. Ann. Rev. Info. Sci. Tech 2005 Jan 31;37(1):51-89. [doi: 10.1002/aris.1440370103]
8. Teng F, Ma Z, Chen J, Xiao M, Huang L. Automatic Medical Code Assignment via Deep Learning Approach for Intelligent Healthcare. IEEE J. Biomed. Health Inform 2020 Sep;24(9):2506-2515. [doi: 10.1109/jbhi.2020.2996937]
9. Subotin, Davis AR. A method for modeling co-occurrence propensity of clinical codes with application to ICD-10-PCS auto-coding. J Am Med Inform Assoc 2016 Sep;23(5):866-871. [doi: 10.1093/jamia/ocv201] [Medline: 26911826]
10. Wang S, Chang Y, Kuo L, Lai F, Chen Y, Yu F, et al. Using Deep Learning for Automatic Icd-10 Classification from FreeText Data. Eur J Biomed Inform 2020;16(1):1-10.
11. Chen P, Wang S, Liao W, Kuo L, Chen K, Lin Y, et al. Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning. JMIR Med Inform 2021 Aug 31;9(8):e23230 [FREE Full text] [doi: 10.2196/23230] [Medline: 34463639]
12. Loper E, Bird S. NLTK: The Natural Language Toolkit. ArXiv Preprint posted online May 5, 2002.
13. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. ArXiv Preprint posted online September 7, 2013.
14. You, Ronghui. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. Advances in Neural Information Processing Systems 2019.
15. Lee, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
16. Huang K, Singh A, Chen S, Moseley E, Deng C, George N, et al. Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation. ArXiv Preprint posted online November 2020. [doi: 10.18653/v1/2020.clinicalnlp-1.11]
17. Schumacher E, Dredze M. Learning unsupervised contextual representations for medical synonym discovery. JAMIA Open 2019 Dec;2(4):538-546. [doi: 10.1093/jamiaopen/ooz057] [Medline: 32025651]
18. Blanco A, Perez-de-Viñaspre O, Pérez A, Casillas A. Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity. Comput Methods Programs Biomed 2020 May;188:105264. [doi: 10.1016/j.cmpb.2019.105264] [Medline: 31851906]

Abbreviations

AttentionXLM: label tree-based attention-aware deep model for high-performance extreme multi-label text classification
AUROC: area under the receiver operating characteristic curve
BERT: bidirectional encoder representations from transformers
BioBERT: bidirectional encoder representations from transformers for biomedical text mining
CC: comorbidity and complications
CM: clinical modification
DD: discharge diagnoses
EHR: electronic health records
ICD: International Classification of Diseases
MH: medical history
PCS: procedure coding system
SE: special examination
SM: surgical method
WHO: World Health Organization
Word2Vec: word to vector

XLNet: generalized autoregressive pretraining for language

Edited by C Lovis; submitted 25.02.22; peer-reviewed by HJ Dai, C Gaudet-Blavignac, A Hasan; comments to author 18.03.22; revised version received 13.05.22; accepted 12.06.22; published 29.06.22.

Please cite as:
Chen PF, Chen KC, Liao WC, Lai F, He TL, Lin SC, Chen WJ, Yang CY, Lin YC, Tsai IC, Chiu CH, Chang SC, Hung FM
Automatic International Classification of Diseases Coding System: Deep Contextualized Language Model With Rule-Based Approaches
JMIR Med Inform 2022;10(6):e37557
URL: https://medinform.jmir.org/2022/6/e37557
doi:10.2196/37557
PMID:35767353

Original Paper

A Clinical Decision Support System for Assessing the Risk of Cervical Cancer: Development and Evaluation Study

Nasrin Chekin[1*], MSc; Haleh Ayatollahi[2*], PhD; Mojgan Karimi Zarchi[3,4*], MD

[1]Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran

[2]Health Management and Economics Research Center, Health Management Research Institute, Iran University of Medical Sciences, Tehran, Iran

[3]Department of Obstetrics and Gynecology, School of Medicine, Iran University of Medical Sciences, Tehran, Iran

[4]Endometriosis Research Center, Iran University of Medical Sciences, Tehran, Iran

[*]all authors contributed equally

Corresponding Author:
Haleh Ayatollahi, PhD
Health Management and Economics Research Center, Health Management Research Institute
Iran University of Medical Sciences
No 4, Yasemi St, Vali-e-Asr St
Tehran, 1996713883
Iran
Phone: 98 2188794301
Email: ayatollahi.h@iums.ac.ir

Abstract

Background: Cervical cancer has been recognized as a preventable type of cancer. As the assessment of all the risk factors of a disease is challenging for physicians, information technology and risk assessment models have been used to estimate the degree of risk.

Objective: The aim of this study was to develop a clinical decision support system to assess the risk of cervical cancer.

Methods: This study was conducted in 2 phases in 2021. In the first phase of the study, 20 gynecologists completed a questionnaire to determine the essential parameters for assessing the risk of cervical cancer, and the data were analyzed using descriptive statistics. In the second phase of the study, the prototype of the clinical decision support system was developed and evaluated.

Results: The findings revealed that the most important parameters for assessing the risk of cervical cancer consisted of general and specific parameters. In total, the 8 parameters that had the greatest impact on the risk of cervical cancer were selected. After developing the clinical decision support system, it was evaluated and the mean values of sensitivity, specificity, and accuracy were 85.81%, 93.82%, and 91.39%, respectively.

Conclusions: The clinical decision support system developed in this study can facilitate the process of identifying people who are at risk of developing cervical cancer. In addition, it can help to increase the quality of health care and reduce the costs associated with the treatment of cervical cancer.

(JMIR Med Inform 2022;10(6):e34753) doi:10.2196/34753

KEYWORDS

cervical cancer; clinical decision support system; risk assessment; medical informatics; cancer; oncology; decision support; risk; CDSS; cervical; prototype; evaluation; testing

Introduction

Cervical cancer is one of the most common and deadliest cancers after breast cancer in women [1]. Approximately 85% of cervical cancer deaths occur in transitional countries, and the rate of cervical cancer death in low- to middle-income countries is 18 times higher than that of high-income countries [2]. Among the causes of cervical cancer, human papillomavirus (HPV) types 16 and 18 are associated with more than 70% of cervical cancers. Other risk factors include early marriage, sexual intercourse before the age of 16, multiple sex partners, smoking, and some genital infections, such as HIV or chlamydia, that can be transmitted through sexual contact [3].

Cervical cancer has been recognized as a preventable type of cancer, as it has a long journey before tissue invasion, and can be prevented by proper screening plans and treating primary lesions [4]. However, risky cases are not diagnosed at an early stage in most transitional countries mainly due to the shortage of obstetricians and gynecologists or patients' fear of and objection to invasive procedures. Therefore, most women with this disease, as compared with other diseases, die at a younger age [5]. To solve this problem, cervical cancer screening and the risk assessment of this disease are among the most common actions that should be taken with the aim of prevention, diagnosis, and treatment of lesions at the primary stage [4]. Current statistical risk assessment models estimate the likelihood of cancer development by examining the association between genetic, environmental, and behavioral risk factors [6]. These models classify women as high- and low-risk patients using clinical data. As a result, invasive procedures are not required for all patients and are only recommended for high-risk patients [7].

As mentioned before, the shortage of different physician specialties, including obstetricians and gynecologists, is among the substantial barriers to providing health care services for women in many low- and middle-income countries [8]. Therefore, a team-based care model along with using digital tools has been suggested to increase the accessibility and quality of health care services [9]. Currently, the use of information technology, and in particular, the use of clinical decision support systems (CDSSs) in the field of medicine has supported other traditional approaches to solve complex medical issues and make more appropriate decisions [10]. Simply, a CDSS is an interactive and flexible information system that is developed specifically to support solving nonstructural problems and improve the decision-making process [11]. These systems can be used by different health care professionals including general practitioners (GPs) and nurses and help them make the right decision at the point of need. The applications of CDSSs include screening different diseases, providing clinicians with reliable information for decision-making, presenting a variety of treatment strategies, and predicting drug interactions to improve patient care and reduce medical and nursing errors [12].

It is also expected that using health information technologies such as CDSS helps improve equity by providing health care services for different groups of patients in a variety of geographical locations [13]. However, there might be some shortfalls in using CDSSs. For example, human decision-makers may directly adopt computer recommendations mainly due to reasons such as increasing efficiency, the higher objectivity of computer conclusions, or having difficulty justifying any deviation from the computer recommendations. Moreover, low-quality data may cause the system to make incorrect decisions, and problems may arise when contextual factors that are relevant but not represented in the data sets are ignored in decision-making. This may also cause errors in identifying high-risk patients. Other risks of automated decisions include the shifting of responsibility, potential manipulation, and the lack of traceability by patients [13].

In the field of oncology, CDSSs can help assess the risk of cancer development by using clinical data and quantifying the impact of cancer risk factors [14]. These systems can also support early disease detection and allow GPs to provide a care plan when specialists are not available [15,16]. Although some similar systems have been previously developed for cervical cancer risk assessment, the number and types of input and output variables and the types of rules and algorithms used are different. According to the literature, machine learning algorithms to predict cervical cancer [17], artificial neural networks (ANNs) to combine the cytology and biomarker results [18], and ANNs to classify the normal and abnormal cells in the cervix region of the uterus [19] have been applied in previous studies. However, in these studies, the cytology results were the main input variables. Given the limited number of research conducted on the application of information technology to assess the risk of cervical cancer, the aim of this study was to develop and implement a CDSS to assess the risk of this disease by considering more simple variables to help patients and clinicians avoid unnecessary invasive procedures, save time, reduce costs, and increase the quality of care.

Methods

This study was conducted in 2 phases in 2021.

Phase 1

The first phase of the study included determining the essential parameters for assessing the risk of cervical cancer. Initially, a list of these parameters was provided based on literature reviews [4,15,20-24]. Subsequently, 20 gynecologists completed a 5-point Likert scale questionnaire (very important=5, important=4, moderately important=3, slightly important=2, and unimportant=1) to determine the most important parameters included on the list. The questionnaire consisted of 2 sections. The first section collected the participants' personal information, such as age and work experience, and the second section consisted of 50 parameters and risk factors related to cervical cancer. The face and content validity of the questionnaire was assessed by 5 gynecologists. The reliability of the questionnaire was calculated using the test-retest method, and 15 gynecologists out of the research sample were asked to complete the questionnaire twice within 2 weeks. Afterward, the correlation coefficient was calculated for the questionnaire (r=.87).

To analyze the data, descriptive statistics and SPSS software (version 24; IBM Corp) were used. Initially, the mean values and SDs were calculated for each parameter. All parameters with a mean value of 4 or more were selected to focus on the main parameters and facilitate the process of writing the rules [10]. Subsequently, one of the gynecologists was consulted, and 8 important parameters were selected to be included in the system.

Phase 2

In the second phase of the study, the system rules were written based on the findings of the first phase of the research and by using MATLAB software (version 9.5; MathWorks Inc). The graphical user interface of the system was designed, and the sensitivity, specificity, and accuracy of the system were evaluated. In this phase, the required data were collected from the outpatient medical records of patients referred to gynecology

clinics (n=93). The gynecologists were requested to complete a data collection form including the 8 selected parameters for each patient and determine the patient's risk of cervical cancer based on their own knowledge and experience. Finally, the level of the risk suggested by system was compared to the gynecologists' opinions (gold standard) using the Cohen κ coefficient. A κ value greater than 0.75 indicates a very good agreement, a κ value less than 0.4 indicates a weak agreement, and a κ value between 0.4 and 0.75 indicates a relatively good agreement [10]. The receiver operating characteristic (ROC) curve of the system was also drawn. The greater the diagnostic power of the system, the ROC curve will be above the square diameter and closer to the ideal condition of an area under curve of 1 [10].

Ethics Approval

Ethics approval was obtained from the National Committee of Ethics in Biomedical Research (IR.IUMS.REC.1400.940).

Results

The findings of the first phase of the study indicated that of the 20 gynecologists, those in the age range of 41-45 years (n=8, 40%) and with work experiences of 5-10 years (n=12, 60%) were the most frequent. According to the participants' perspectives, a number of general and specific parameters were more important than others for assessing the risk of cervical cancer (Table 1).

As previously noted, one of the gynecologists was consulted, and 8 important parameters were selected among all items with a mean value of 4 or more to be included in the system. These parameters were the history of high-risk HPV (16, 18), number of patient's sexual partners, history of various sexually transmitted infections, smoking status, Papanicolaou (Pap smear) test results, number of husband's legal sexual partners, age of the first sexual intercourse, and history of cervical and vaginal diseases.

After writing the If-Then rules, the graphical user interface of the system was designed using MATLAB software (Figure 1). The interface of the CDSS consisted of input and output variables. The input variables included the data for the 8 important parameters mentioned above, and the output variable was the risk assessment result that showed 4 different levels: safe, low risk, moderate risk, and high risk.

The system was evaluated using data collected from patients who were referred to gynecological clinics. In total, 100 patients visited the gynecological clinics in 1 month; however, 7 patients were excluded due to definite cervical cancer diagnoses, and the number of patients was reduced to 93. The patients' data were entered into the system and the results were compared to the gynecologists' opinions. Table 2 shows the values of sensitivity, specificity, and accuracy for the different risk groups.

The Cohen κ coefficient was also calculated to compare the risk level assessed by the CDSS and the gynecologists' opinions. The results revealed that the κ value was 0.89 for the low-risk group, 0.73 for the moderate-risk group, 0.74 for the high-risk group, and 0.79 for the whole system. As the κ values were greater than or close to 0.75, it can be concluded that there was a good agreement between the system performance and gynecologists' opinions. The ROC curve of the system was also drawn (Figure 2). The results showed that the ROC curve was above the square diameter and close to the ideal condition of an area under curve of 1. This indicated high diagnostic power by the system.

Table 1. Important general and specific parameters for assessing the risk of cervical cancer.

Parameter	Degree of importance[a]					Mean (SD)
	Very important, n (%)	Important, n (%)	Moderately important, n (%)	Slightly important, n (%)	Unimportant, n (%)	
General parameters						
Patient's age	9 (45)	8 (40)	3 (15)	0 (0)	0 (0)	4.30 (0.73)
Smoking status	11 (55)	5 (25)	4 (20)	0 (0)	0 (0)	4.40 (0.75)
History of exposure to smoke	10 (5)	5 (25)	4 (20)	1 (5)	0 (0)	4.20 (0.95)
Patient's social status	8 (40)	9 (45)	3 (15)	0 (0)	0 (0)	4.25 (0.71)
Marital status	8 (40)	9 (45)	3 (15)	0 (0)	0 (0)	4.25 (0.71)
History of high-risk HPV[b] (16, 18)	20 (100)	0 (0)	0 (0)	0 (0)	0 (0)	5 (0)
History of HPV vaccination	13 (65)	5 (25)	2 (10)	0 (0)	0 (0)	4.55 (0.68)
Family history of cervical cancer	12 (60)	4 (20)	3 (15)	0 (0)	1 (5)	4.30 (1.08)
Genetic factors	12 (60)	3 (15)	5 (25)	0 (0)	0 (0)	4.35 (0.87)
Number of sexual partners	16 (80)	3 (15)	1 (5)	0 (0)	0 (0)	4.75 (0.55)
Number of husband's legal sexual partners	16 (80)	4 (20)	0 (0)	0 (0)	0 (0)	4.80 (0.41)
Age of marriage	7 (35)	10 (50)	3 (15)	0 (0)	0 (0)	4.20 (0.69)
Age of the first sexual intercourse	9 (45)	7 (35)	4 (20)	0 (0)	0 (0)	4.25 (0.78)
Number of sexual intercourse per month	7 (35)	6 (30)	7 (35)	0 (0)	0 (0)	4 (0.85)
Specific parameters						
Sexual health status	10 (50)	10 (50)	0 (0)	0 (0)	0 (0)	4.5 (0.51)
Papanicolaou test results	17 (85)	2 (10)	0 (0)	1 (5)	0 (0)	4.75 (0.71)
History of immune deficiency diseases	14 (70)	6 (30)	0 (0)	0 (0)	0 (0)	4.70 (0.47)
History of cervical and vaginal diseases	16 (80)	3 (15)	1 (5)	0 (0)	0 (0)	4.75 (0.55)
History of ovarian and fallopian tube diseases	7 (35)	8 (40)	4 (20)	1 (5)	0 (0)	4.05 (0.88)
History of hysterectomy	8 (40)	6 (30)	4 (20)	2 (10)	0 (0)	4 (1.02)
History of sexually transmitted infections	14 (70)	6 (30)	0 (0)	0 (0)	0 (0)	4.70 (0.47)

[a]Very important=5, important=4, moderately important=3, slightly important=2, and unimportant=1.

[b]HPV: human papillomavirus.

Figure 1. User interface of the clinical decision support system (CDSS) to assess the risk of cervical cancer. HPV: human papillomavirus.

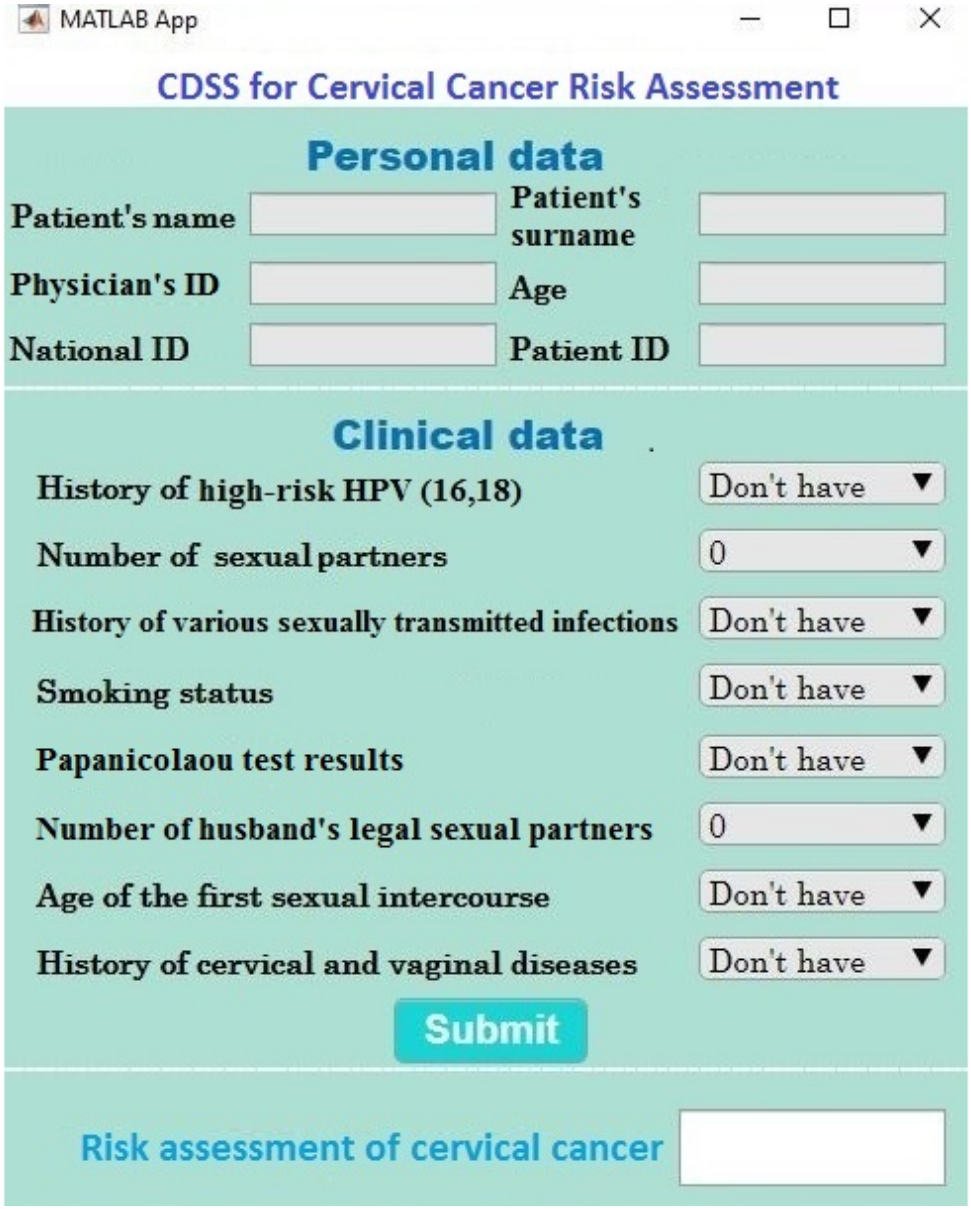
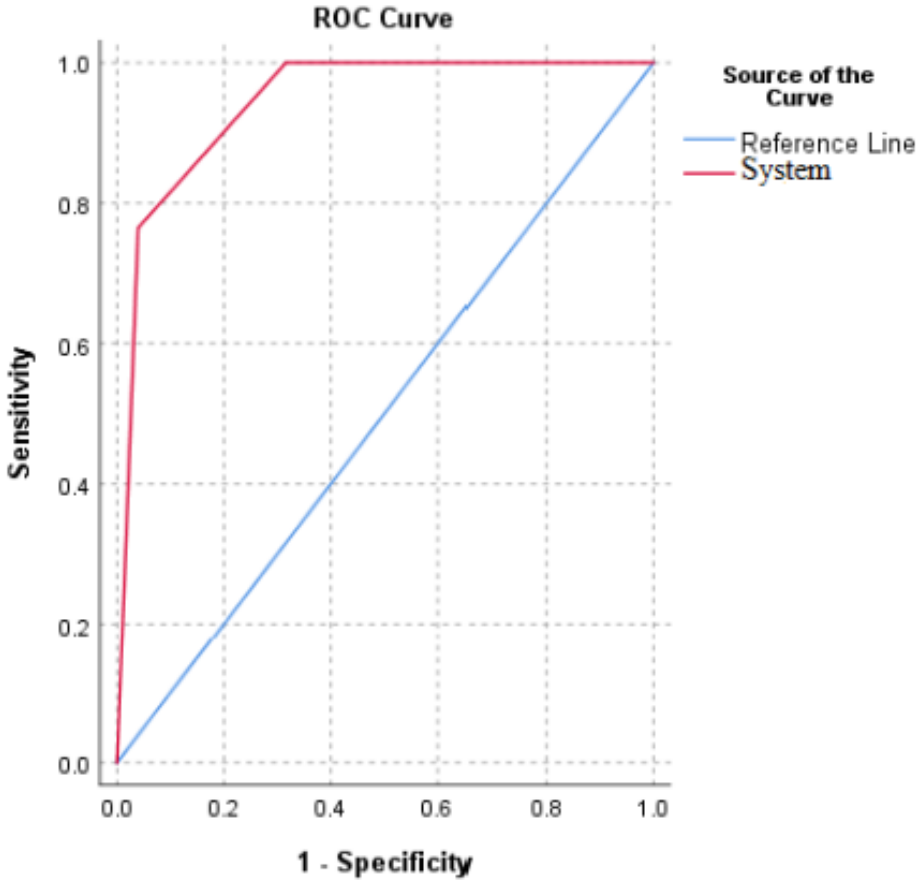


Table 2. Sensitivity, specificity, and accuracy of the system for different risk groups.

Risk group	Evaluation criteria		
	Sensitivity, %	Specificity, %	Accuracy, %
Low risk	93.70	95.50	94.62
Moderate risk	78.26	90	87.09
High risk	76.47	96.05	92.47
Mean values for the system	82.81	93.85	91.39

Figure 2. Receiver operating characteristic (ROC) curve.



Discussion

Principal Findings

In this study, the essential parameters for cervical cancer risk assessment were identified and divided into 2 categories of general and specific parameters. To design a CDSS, the most important risk factors were selected based on the gynecologists' opinions and consultation with a specialist. The results of the evaluation study showed that the developed system had a high level of sensitivity, specificity, and accuracy and, in most cases, was able to identify at-risk patients similar to the specialists.

Assessing the risk of a disease is one of the greatest challenges in medical sciences. Most clinical decisions are made based on the physicians' personal understanding and experience; however, their expertise may not be adequate for assessing the risk of all diseases or disorders. Therefore, the risk assessment of diseases has been the focus of many research studies in recent years [10]. As there are different risk factors for a disease, information technology and risk assessment models are used to quantify the risk level [21,25]. Regarding cervical cancer, it is possible to identify at-risk women by determining the risk factors and measuring the effect of these factors on the risk of cancer. In addition, prevention or intervention in the early stages of the disease can be made possible by early detection in patients and then carrying out further examinations [16,26].

Comparison With Prior Studies

In this study, the patient's age, smoking status, social status, and marital status were the general parameters and the history of high-risk HPV (16, 18), history of HPV vaccination, family history of cervical cancer, genetic factors, and number of patient's sexual partners were the specific parameters that had the highest mean values of importance. Similarly, in a study conducted by Vaisy et al [27], the patient's age, age of the first delivery, history of abortion and curettage, number of pregnancies, and economic and social status were identified as risk factors of cervical cancer. Vaisy et al also showed that marital status, the number of marriages, marriage under the age of 16 years, and taking birth control pills can increase the risk of cervical cancer.

Another study conducted by Nojomi et al [28] indicated that demographic variables such as marital status, occupation, literacy, the duration of using birth control pills, the history of abortion, the family history of cervical cancer, smoking status, age at marriage, and mother's age at the birth of her first child are among the cervical cancer risk factors. The researchers also indicated that a positive family history of cervical cancer, low age of marriage, high number of pregnancies, low age at the birth of the first child, and long-term use of birth control pills were the most significant risk factors. Similarly, Nkfusai et al [29] examined the role of smoking status, the number of sexual partners, the family history of cervical cancer, the history of HIV infection, and having more than 5 deliveries as cervical cancer risk factors.

Therefore, the essential parameters for assessing the risk of cervical cancer found in the first phase of the study were consistent with the findings of other similar studies. It should be noted that although 2 groups of general and specific parameters were considered in this study, 8 parameters were selected based on consulting with a gynecologist to facilitate the process of rule writing, developing, and implementing the CDSS. These 8 parameters were the history of high-risk HPV (16, 18), number of sexual partners, history of various sexually transmitted infections, smoking status, Papanicolaou test results, number of husband's legal sexual partners, age of the first sexual intercourse, and history of cervical and vaginal diseases. These parameters have also been mentioned in other similar studies [27-29]. After determining the essential parameters in assessing the risk of cervical cancer, a prototype of the CDSS was developed using MATLAB software. The rules of the system were determined after consulting a gynecologist, and the graphical user interface was developed using MATLAB software. The users could enter data into the system, and the result of the cervical cancer risk assessment would be displayed as safe, low risk, moderate risk, or high risk.

Similarly, Omololu and Adeoluo [30] extracted a number of cervical cancer risk factors from patient records. These risk factors included HPV infection, the number of sexual partners, the age of the first sexual intercourse, extramarital affairs of spouses, economic and social status, the use of oral birth control pills, and genetic history. In their study, cervical cancer diagnosis was considered as the system output and adaptive neuro-fuzzy inference was used. However, in this study, the system was able to assess the risk of cervical cancer by using If-Then rules.

After developing the system, the data collected from the outpatient medical records were used to evaluate the system performance. Among the low-risk, moderate-risk, and high-risk groups, the highest sensitivity (93.70%) and accuracy (94.62%) belonged to the low-risk group, the highest specificity (96.05%) and lowest sensitivity (76.47%) belonged to the high-risk group, and the lowest specificity (90%) and accuracy (87.09%) belonged to the moderate-risk group. In general, the sensitivity, specificity, and accuracy of the system were calculated to be 82.81%, 93.85%, and 91.39%, respectively.

Similarly, Hu et al [20] used a regression model and an ANN to assess the risk of cervical cancer. After evaluating the model, the sensitivity and specificity of the model were 95.2% and 99%, respectively. In another study, Lee et al [24] validated a risk scoring system. They used patient medical records to collect the data and the Cox risk model to determine the risk score. The results indicated that the sensitivity and specificity in the group under Papanicolaou screening with a follow-up of less than 3 years were 75% and 94.1%, respectively. The sensitivity and specificity in the similar group with a follow-up of less than 5 years were 66.7% and 93.5%, respectively, and in the screening group using cytological tests, the sensitivity and specificity were 88.2% and 87.7%, respectively. Bountaries et al [31] used the retrospective data of patients who underwent colposcopy. Their CDSS classified cancer lesions using a hybrid genetic algorithm and Bayesian classification. To evaluate the system, they compared the sensitivity and specificity of their CDSS in diagnosing cancerous lesions with the Papanicolaou test and HPV detection results. The sensitivity and specificity of the developed system were 83.4% and 88.1%, respectively.

It is notable that the sensitivities, specificities, and accuracies cannot be compared between the different systems mainly due to the differences in the input and output variables and algorithms used. Although neural networks and other algorithms that might have higher precision in detecting at-risk patients were not used in this study, the results of this study showed that the developed system had a high level of sensitivity, specificity, and accuracy similar to other systems and could be used to screen and identify women at risk of developing cervical cancer. The designed system can be used by different health care providers including nurses, GPs, and gynecologists, as it had been developed based on basic clinical data. It can help regular screenings and prevent invasive tests for all patients. Moreover, identifying at-risk women at the early stages of the disease can help treat primary lesions and reduces malignancy and death [16]. In addition, a better allocation of heath care resources and improving the quality of care are expected by classifying patients into different risk groups.

Research Limitations

There are various parameters to assess the risk of cervical cancer; however, it is difficult to gather and consider all of these parameters in a single CDSS. Therefore, in this study, the essential parameters were selected and considered for developing the system based on the gynecologists' opinions. Including other parameters in future systems and using more sophisticated methods for system design may help assess the risk of cervical cancer more precisely. Future researchers can use parameters that were not included in the current system, or they can use new parameters that might be introduced by other researchers.

Conclusion

The aim of this study was to develop a CDSS to assess the risk of cervical cancer. In this study, 8 essential parameters were selected and considered as input variables. The output of the system showed the risk of cervical cancer in 4 levels: safe, low risk, moderate risk, and high risk. The findings of this study revealed that the system performance was very similar to the gynecologists' opinions. Such a system could be used for cervical cancer screening or in regions where access to gynecologists is limited. The use of this system can help improve the quality of care and manage patients more effectively. Moreover, the reduction of the mortality rate of cervical cancer through continuous and timely patient screening would be another benefit of using this system.

Acknowledgments

This research was supported by a grant (1400-1-48-21549) from the Health Management and Economics Research Center affiliated with the Iran University of Medical Sciences.

Conflicts of Interest

None declared.

References

1. Hanifi M, Jalili Z, Tavakoli R. The effects of an educational intervention based on the BASENEF model on promoting cervical cancer prevention behaviors among women. Payesh 2018;17(1):67-73 [FREE Full text]
2. Ding T, Ma H, Feng J. A three-gene novel predictor for improving the prognosis of cervical cancer. Oncol Lett 2019 Nov;18(5):4907-4915 [FREE Full text] [doi: 10.3892/ol.2019.10815] [Medline: 31612001]
3. Shrestha AD, Neupane D, Vedsted P, Kallestrup P. Cervical cancer prevalence, incidence and mortality in low and middle income countries: a systematic review. Asian Pac J Cancer Prev 2018 Feb 26;19(2):319-324 [FREE Full text] [doi: 10.22034/APJCP.2018.19.2.319] [Medline: 29479954]
4. Ghaoomi M, Aminimoghadam S, Safari H, Mahmoudzadeh A. Awareness and practice of cervical cancer and Pap smear testing in a teaching hospital in Tehran. Tehran Univ Med J 2016 Jun;74(3):183-189 [FREE Full text]
5. Nahvijou A. Decision analysis of cervical cancer screenings strategies in Iran based on simulation models. PhD Thesis. Tehran: Iran University of Medical Sciences; 2014.
6. Walker JG, Licqurish S, Chiang PPC, Pirotta M, Emery JD. Cancer risk assessment tools in primary care: a systematic review of randomized controlled trials. Ann Fam Med 2015 Sep;13(5):480-489 [FREE Full text] [doi: 10.1370/afm.1837] [Medline: 26371271]
7. Buijsse B, Simmons RK, Griffin SJ, Schulze MB. Risk assessment tools for identifying individuals at risk of developing type 2 diabetes. Epidemiol Rev 2011;33:46-62 [FREE Full text] [doi: 10.1093/epirev/mxq019] [Medline: 21622851]
8. Hoyler M, Finlayson SRG, McClain CD, Meara JG, Hagander L. Shortage of doctors, shortage of data: a review of the global surgery, obstetrics, and anesthesia workforce literature. World J Surg 2014 Feb 12;38(2):269-280 [FREE Full text] [doi: 10.1007/s00268-013-2324-y] [Medline: 24218153]
9. Rayburn W, Tracy E. Changes in the practice of obstetrics and gynecology. Obstet Gynecol Surv 2016 Jan;71(1):43-50. [doi: 10.1097/OGX.0000000000000264] [Medline: 26819135]
10. Boni TTA, Ayatollahi H, Langarizadeh M. A clinical decision support system for assessing the risk of cardiovascular diseases in diabetic hemodialysis patients. Curr Diabetes Rev 2020 Mar 20;16(3):262-269. [doi: 10.2174/1573399815666190531100012] [Medline: 31146666]
11. Safdari R, Mirzaie M. Clinical decision support systems: concepts, technical considerations and barriers. 2011 Presented at: Proceedings of the 1st congress on the application of information technology in health; October 19, 2011; Sari, Iran.
12. Sutton R, Pincock D, Baumgart D, Sadowski D, Fedorak R, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020 Feb 06;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
13. Orwat C. Risks of discrimination through the use of algorithms. Federal Anti-Discrimination Agency. 2019 Sep. URL: https://www.antidiskriminierungsstelle.de/EN/homepage/_documents/download_diskr_risiken_verwendung_von_algorithmen.pdf?__blob=publicationFile&v=1 [accessed 2022-05-16]
14. Katki HA, Wacholder S, Solomon D, Castle PE, Schiffman M. Risk estimation for the next generation of prevention programmes for cervical cancer. Lancet Oncol 2009 Nov;10(11):1022-1023 [FREE Full text] [doi: 10.1016/S1470-2045(09)70253-0] [Medline: 19767237]
15. Ravikumar K, MacLaughlin K, Scheitel M, Kessler M, Wagholikar K, Liu H, et al. Improving the accuracy of a clinical decision support system for cervical cancer screening and surveillance. Appl Clin Inform 2018 Jan;9(1):62-71 [FREE Full text] [doi: 10.1055/s-0037-1617451] [Medline: 29365341]
16. Castle PE, Sideri M, Jeronimo J, Solomon D, Schiffman M. Risk assessment to guide the prevention of cervical cancer. Am J Obstet Gynecol 2007 Oct;197(4):356.e1-356.e6 [FREE Full text] [doi: 10.1016/j.ajog.2007.07.049] [Medline: 17904958]
17. Curia F. Cervical cancer risk prediction with robust ensemble and explainable black boxes method. Health Technol 2021 May 14;11(4):875-885 [FREE Full text] [doi: 10.1007/s12553-021-00554-6]
18. Kyrgiou M, Pouliakis A, Panayiotides JG, Margari N, Bountris P, Valasoulis G, et al. Personalised management of women with cervical abnormalities using a clinical decision support scoring system. Gynecol Oncol 2016 Apr;141(1):29-35. [doi: 10.1016/j.ygyno.2015.12.032] [Medline: 27016226]
19. Devi MA, Ravi S, Vaishnavi J, Punitha S. Classification of cervical cancer using artificial neural networks. Procedia Computer Science 2016;89:465-472. [doi: 10.1016/j.procs.2016.06.105]

20. Hu B, Tao N, Zeng F, Zhao M, Qiu L, Chen W, et al. A risk evaluation model of cervical cancer based on etiology and human leukocyte antigen allele susceptibility. Int J Infect Dis 2014 Nov;28:8-12 [FREE Full text] [doi: 10.1016/j.ijid.2014.05.015] [Medline: 25223804]
21. Al-Madani W, Ahmed AE, Arabi H, Al Khodairy S, Al Mutairi N, Jazieh AR. Modelling risk assessment for cervical cancer in symptomatic Saudi women. Saudi Med J 2019 May;40(5):447-451 [FREE Full text] [doi: 10.15537/smj.2019.5.24085] [Medline: 31056620]
22. Vanakankovit N, Taneepanichskul S. Effect of oral contraceptives on risk of cervical cancer. J Med Assoc Thai 2008 Jan;91(1):7-12. [Medline: 18386537]
23. Vaisy A, Lotfinejad S, Zhian F. Risk of cancer with combined oral contraceptive use among Iranian women. Asian Pac J Cancer Prev 2014;15(14):5517-5522 [FREE Full text] [doi: 10.7314/apjcp.2014.15.14.5517] [Medline: 25081657]
24. Lee C, Peng C, Li R, Chen Y, Tsai H, Hung Y, et al. Risk evaluation for the development of cervical intraepithelial neoplasia: development and validation of risk-scoring schemes. Int J Cancer 2015 Jan 15;136(2):340-349 [FREE Full text] [doi: 10.1002/ijc.28982] [Medline: 24841989]
25. Gorthi A, Firtion C, Vepa J. Automated risk assessment tool for pregnancy care. 2009 Nov 13 Presented at: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society; September 3-6, 2009; Minneapolis, MN p. 6222-6225. [doi: 10.1109/iembs.2009.5334644]
26. Khodakarami N, Farzaneh F, Yavari P, Khayamzadeh M, Taheripanah R, Akbari M. The new guideline for cervical cancer screening in low risk Iranian women. Iranian Journal of Obestetric, Gynecology and Infertility 2014 Apr;17(95):8-17 [FREE Full text]
27. Vaisy A, Lotfinejad S, Zhian F. Risk factors for cervical cancer among women referred to health services center of Tehran University of Medical Sciences. J Ardabil Univ Med Sci 2013;13(3):327-336 [FREE Full text]
28. Nojomi M, Modares Gilani M, Erfani A, Mozafari N, Motaghi A. The study of frequent of risk factors of cervical cancer among women attending general hospital in Tehran 2005-2006. Razi Journal of Medical Sciences 2007;14(56):189-195 [FREE Full text]
29. Nkfusai NC, Cumber SN, Anchang-Kimbi JK, Nji KE, Shirinde J, Anong ND. Assessment of the current state of knowledge and risk factors of cervical cancer among women in the Buea Health District, Cameroon. Pan Afr Med J 2019;33:38 [FREE Full text] [doi: 10.11604/pamj.2019.33.38.16767] [Medline: 31384353]
30. Omololu A, Adeoluo M. Modeling and diagnosis of cervical cancer using adaptive neuro fuzzy inferences system. World J Res Rev 2018 May;6(5):1-3 [FREE Full text]
31. Bountris P, Topaka E, Pouliakis A, Haritou M, Karakitsos P, Koutsouris D. Development of a clinical decision support system using genetic algorithms and Bayesian classification for improving the personalised management of women attending a colposcopy room. Healthc Technol Lett 2016 Jun 14;3(2):143-149 [FREE Full text] [doi: 10.1049/htl.2015.0051] [Medline: 27382484]

Abbreviations

ANN: artificial neural network
CDSS: clinical decision support system
GP: general practitioner
HPV: human papillomavirus
ROC: receiver operating characteristic

Edited by C Lovis; submitted 06.11.21; peer-reviewed by G McLeod, G Vyshka; comments to author 02.01.22; revised version received 05.02.22; accepted 25.02.22; published 22.06.22.

Please cite as:
Chekin N, Ayatollahi H, Karimi Zarchi M
A Clinical Decision Support System for Assessing the Risk of Cervical Cancer: Development and Evaluation Study
JMIR Med Inform 2022;10(6):e34753
URL: https://medinform.jmir.org/2022/6/e34753
doi:10.2196/34753
PMID:35731549

Original Paper

Personalized Recommendations for Physical Activity e-Coaching (OntoRecoModel): Ontological Modeling

Ayan Chatterjee[1], MEng; Andreas Prinz[1], PhD

Department of Information and Communication Technology, Center for eHealth, University of Agder, Grimstad, Norway

Corresponding Author:
Ayan Chatterjee, MEng
Department of Information and Communication Technology
Center for eHealth
University of Agder
Jon Lilletuns Vei 9
Grimstad, 4879
Norway
Phone: 47 94719372
Email: ayan.chatterjee@uia.no

Abstract

Background: Automatic e-coaching may motivate individuals to lead a healthy lifestyle with early health risk prediction, personalized recommendation generation, and goal evaluation. Multiple studies have reported on uninterrupted and automatic monitoring of behavioral aspects (such as sedentary time, amount, and type of physical activity); however, e-coaching and personalized feedback techniques are still in a nascent stage. Current intelligent coaching strategies are mostly based on the handcrafted string messages that rarely individualize to each user's needs, context, and preferences. Therefore, more realistic, flexible, practical, sophisticated, and engaging strategies are needed to model personalized recommendations.

Objective: This study aims to design and develop an ontology to model personalized recommendation message intent, components (such as suggestion, feedback, argument, and follow-ups), and contents (such as spatial and temporal content and objects relevant to perform the recommended activities). A reasoning technique will help to discover implied knowledge from the proposed ontology. Furthermore, recommendation messages can be classified into different categories in the proposed ontology.

Methods: The ontology was created using Protégé (version 5.5.0) open-source software. We used the Java-based Jena Framework (version 3.16) to build a semantic web application as a proof of concept, which included Resource Description Framework application programming interface, World Wide Web Consortium Web Ontology Language application programming interface, native tuple database, and SPARQL Protocol and Resource Description Framework Query Language query engine. The HermiT (version 1.4.3.x) ontology reasoner available in Protégé 5.x implemented the logical and structural consistency of the proposed ontology. To verify the proposed ontology model, we simulated data for 8 test cases. The personalized recommendation messages were generated based on the processing of personal activity data in combination with contextual weather data and personal preference data. The developed ontology was processed using a query engine against a rule base to generate personalized recommendations.

Results: The proposed ontology was implemented in automatic activity coaching to generate and deliver meaningful, personalized lifestyle recommendations. The ontology can be visualized using OWLViz and OntoGraf. In addition, we developed an ontology verification module that behaves similar to a rule-based decision support system to analyze the generation and delivery of personalized recommendation messages following a logical structure.

Conclusions: This study led to the creation of a meaningful ontology to generate and model personalized recommendation messages for physical activity coaching.

(JMIR Med Inform 2022;10(6):e33847) doi:10.2196/33847

KEYWORDS

descriptive logic; ontology; e-coach; reasoning; recommendation generation

Introduction

Overview

Currently, risk factors associated with unhealthy lifestyle have been recognized as the foremost contributors to chronic illness and mortality in developed countries [1-6]. An e-coach system can guide people and convey the appropriate recommendations in context with sufficient time to prevent and improve living with chronic conditions. It is a set of computerized components that constitute an artificial entity that can observe, reason about, learn from, and predict a user's behaviors, in context and over time, and engages proactively in an ongoing collaborative conversation with the user to aid planning and promote effective goal striving using persuasive techniques [7-10]. Motivating people toward a healthy lifestyle has been challenging without appropriate and continuous support and correct intervention planning [7-10]. Personalized recommendation technology in health care may be helpful to address such challenges. It requires the proper collection of personal health and wellness data and the right recommendation generation and delivery in a meaningful way. Our previous study [11] focused on creating a meaningful, context-specific holistic ontology to model raw and unstructured observations of personal health and wellness data collected from heterogeneous sources (eg, sensors, interviews, and questionnaires) with semantic metadata and create a compact and logical abstraction for health risk prediction. However, this comprehensive study concentrated on rule-based recommendation generation and semantic modeling of recommendation messages for physical activity coaching.

Motivation

Generation of motivational messages is essential in e-coaching. Motivational messages provide quick information on time in a more natural and meaningful manner to translate behavioral observations into inspiring, easy-to-follow, and achievable actions. Moreover, these messages must be diverse to make the e-coach system more reasonable and reliable. In activity coaching, personalized motivational messages can offer inspiration for a day, week, or month based on the activity goals. It helps to regain motivation when the individual has lost motivation to attain activity goals. The medium of recommendation delivery can be diverse and depends on personal interaction choices (eg, graphical visualization, pop-up textual notification, and audio-visual material). In existing studies, motivational messages have textual forms that follow a static predefined format; therefore, they are difficult to individualize. Existing ontologies do not include model recommendation message intent, components, and contents important to automatically select accurate messages in e-coaching. Personalized recommendation generation for a healthy lifestyle is closely related to personal preferences. Thus, personal preferences can be of 3 types: activity goal setting (eg, nature of goals—direct vs motivational goals and generic vs personalized goals), response type (eg, mode to communicate extended health state, health state prediction, and customized recommendations for activity coaching), and nature of interaction with the e-coach system (eg, mode, frequency, and medium). In this study, we have gone one step ahead to perform semantic (ontological) modeling of preference data and recommendation messages beyond static textual form to describe its characteristics, metadata, and content information.

The use of ontologies has certain benefits while modeling recommendation messages. It helps to interpret which recommendation message is to be generated using a binary tree-like structure (if-then or if-then-else conditional statement). Interpretability makes identifying the cause-and-effect relationships between data input and data output easy. In ontology, the logical and structural representation of knowledge, hierarchical model structuring (eg, class and subclass model), and inferred knowledge generation with reasoners can solve interpretability problems in decision-making. Furthermore, benefits such as extensibility, flexibility, generality, and decoupling of knowledge help ontology to develop an appropriate solution to model recommendation messages in automatic coaching.

Aim of the Study

This study proposes a Web Ontology Language (OWL)-based ontology (OntoRecoModel) to deal with personal preferences and recommendation messages and annotate them with semantic metadata information. The *OntoRecoModel* will not only support a logical representation of data and messages but also encourage rule-based decision-making to generate personalized recommendation messages using SPARQL Protocol and Resource Description Framework (RDF) Query Language (SPARQL) as a verification study against different test cases with simulated data. Moreover, we assessed the performance of the ontology against mean reasoning time and query execution time. In *OntoRecoModel*, we annotated the participant's data with Semantic Web Rule Language (SWRL) and stored the resultant OWL file in a triple-store format for better readability. The *OntoRecoModel* allows automatic knowledge inferencing and efficient knowledge representation to balance a trade-off between complexity, persuasiveness, and reasoning about formal knowledge. The entire study was divided into the following two sections: (1) *OntoRecoModel* design and implementation for semantic annotation and (2) its verification with simulated data. The main contributions of this study were the following:

1. Annotation of personal preferences data (activity goal setting, response type, and interaction type) and recommendation messages in the *OntoRecoModel*.
2. Preparation of semantic rules to execute SPARQL queries for different test cases.
3. Use of the prepared rules to generate personalized activity recommendations.

For this set of semantic data, it will be regarded as an assertion of true facts. The main goal of this paper was to trigger a logical rule of shape (A IMPLIES B) in a logically equivalent manner (NOT [A] or B). If some specific variables are inferred to be true, some suggestions should be provided to the participants of the semantic data source.

Related Work

This section offers existing knowledge relevant to current research and a qualitative comparison between our proposed ontology and the existing ontologies based on selected categories

in Table 1. An ontology is a formal description of knowledge as concepts within a domain and their relationships. It uses existing technologies to develop new ideas through conceptual modeling or proof-of-concept studies to solve general real-world or project-specific semantic modeling problems. There are other approaches to knowledge representation that use formal specifications, such as vocabularies, taxonomies, thesaurus, topic maps, and logical models. However, unlike taxonomy or relational database schemas, ontologies express relationships and allow users to bring together or link multiple concepts in novel ways. Furthermore, all the related ontologies are not available in open source. Therefore, it is not straightforward to make quantitative comparisons between different related studies.

Table 1. A qualitative comparison between our proposed study and the existing studies.

Study	Used technologies	Annotation of sensor data	Annotation of personal and health data or health management data	Rule-based recommendation generation	Annotation of preference data	Annotation of recommendation messages
Our study	OWL[a], HermiT, RDF[b], SPARQL[c], TDB[d], OWLViz, OntoGraf, and Java	Yes	No	Yes	Yes	Yes
Chatterjee et al [11]	OWL, HermiT, RDF, SPARQL, TDB, OWLViz, SSN[e], SNOMED-CT[f], OntoGraf, and Java	Yes	Yes	Yes	No	No
Kim et al [12]	OWL	No	Yes	No	No	No
Sojic et al [13]	OWL and SWRL[g]	No	Yes	No	No	No
Kim et al [14]	OWL and FaCT++	No	Yes	No	No	No
Lasierra et al [15]	OWL, RDF, and SPARQL	No	Yes	Yes	No	No
Yao and Kumar [16]	OWL and SWRL	No	Yes	Yes	No	No
Chi et al [17]	OWL and SWRL	No	Yes	Yes	No	No
Rhayem et al [18]	OWL and SWRL	Yes	No	Yes	No	No
Galopin et al [19]	OWL and SWRL	No	Yes	Yes	No	No
Sherimon and Krishnan [20]	OWL and SWRL	No	Yes	Yes	No	No
Hristoskova et al [21]	SOA[h], Amigo, OWL, and SWRL	No	Yes	Yes	No	No
Riano et al [22]	OWL	No	No	Yes	No	No
Jin and Kim [23]	SSN and IETF YANG	Yes	No	No	No	No
Ganguly et al [24]	OWL	No	No	Yes	No	No
Bouza et al [25]	OWL, Decision Tree, and Java	No	No	Yes	No	No
Villalonga et al [26]	OWL and SPARQL	No	No	Yes	No	Yes

[a]OWL: Web Ontology Language.

[b]RDF: Resource Description Framework.

[c]SPARQL: SPARQL Protocol and RDF Query Language.

[d]TDB: tuple database.

[e]SSN: semantic sensor network.

[f]SNOMED-CT: Systematized Nomenclature of Medicine–Clinical Terms.

[g]SWRL: Semantic Web Rule Language.

[h]SOA: service-oriented architecture.

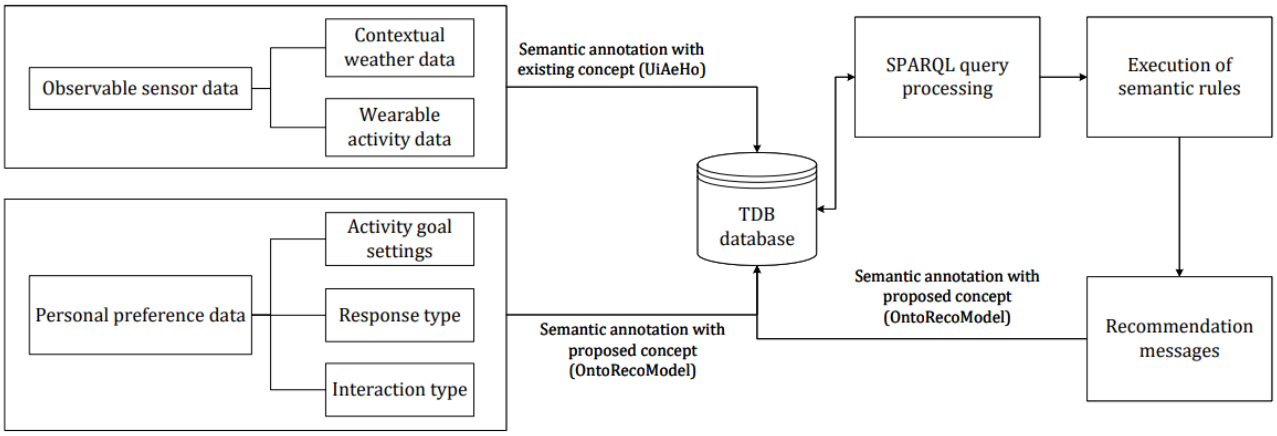
Kim et al [12] developed an ontology model for obesity management, which realizes spontaneous participation of participants and continuous weight monitoring through the nursing process in the field of mobile devices. The scope of obesity management includes behavioral intervention, dietary advice, and physical activity. Similarly, the study includes

evaluation data (BMI, gender, and hip circumference), inferred data to express diagnostic results, evaluation (causes of obesity), success or failure in behavior change, and implementation (education, advice, and intervention). Sojic et al [13] used OWL to model a specific ontology in the obesity field to design reasoning models to personalize health status assessments to be age-specific and gender-specific. The ontology helps to classify personal files according to changes in personal behavior or characteristics over time and automatically infer personal health status, which is of great significance for obesity assessment and prevention. They used SWRL to write the ontology rules. Kim et al [14] proposed a physical activity ontology model to support the interoperability of physical activity data. The ontology was developed in Protégé (version 4.x), and the FaCT++ reasoner verified its structural consistency. On the basis of the automatic calculation paradigm, Monitoring, Analysis, Planning, and Execution, an automatic ontology-based method was developed by Lasierra et al [15] to manage information in the home-based remote monitoring service scenario. Furthermore, they proposed the following three stages [27] for ontology-driven home-based personalized care for the patients with chronic illnesses: stage 1—ontology design and implementation, stage 2—the application of ontology to study the personalization problem, and stage 3—software prototype implementation. The proposed ontology was designed in the Protégé-OWL (version 4.0.2) ontology editor using OWL–Description Logic (OWL-DL) language and verified using the FaCT++ reasoner. Ontology development involves data from heterogeneous sources, such as clinical knowledge, data from medical devices, and patient's contextual data. Yao and Kumar [16] proposed a new *flexible workflow based on clinical context* method, which used ontology modeling to incorporate flexible and adaptive clinical pathways into clinical decision support system (CDSS). They developed 18 SWRL rules to explain practical knowledge of heart failure. The model was verified using the Pellet Reasoner plug-in for Protégé 3.4. In addition, they developed a proof-of-concept prototype of the proposed method using the Drools framework. Chi et al [17] used OWL and SWRL to construct a dietary consultation system. The knowledge base (KB) involves the interaction of heterogeneous data sources and factors such as patient's disease stage, physical condition, activity level, food intake, and key nutritional restrictions. Rhayem et al [18] proposed an ontology (HealthIoT)–based system for patient monitoring using sensors, radio frequency identification, and actuators. They claim that the data obtained from medically connected devices are huge, and therefore, lack restraint and comprehensibility and are manipulated by other systems and devices. Therefore, they proposed an ontology model that represents connected medical devices and their data according to semantic rules and, then, used the proposed Internet of Things medical insurance system for model evaluation, which supports decision-making after analyzing the patient's vital signs. Galopin et al [19] proposed an ontology-based prototype CDSS to manage patients with multiple chronic diseases in accordance with clinical practice guidelines. They prepared a KB based on the clinical practice guidelines and patient observation data. The KB decision rule is based on the *if-then* rule. Sherimon and Krishnan [20] proposed an ontology system (OntoDiabetic) using OWL2 language to support CDSS for patients with cardiovascular disease, diabetic nephropathy, and hypertension to follow clinical guidelines and *if-then* decision rules. Hristoskova et al [21] proposed another ontology-driven environmental intelligence (AmI) framework to support personalized medical detection and alert generation based on the analysis of vital signs collected from patients diagnosed with congestive heart failure. The CDSS system can classify individual congestive heart failure risk stages and notify patients through AmI's reasoning engine. Riano et al [22] proposed an ontology-based CDSS to monitor and intervene in patients with chronic diseases to prevent critical situations, such as misdiagnosis, undetected comorbidities, lack of information, unobserved related diseases, or prevention. An eHealth system was designed and implemented by Jin and Kim [23] using the IETF YANG ontology based on the semantic sensor network (SSN). This method helped to automatically configure eHealth sensors (responsible for collecting body temperature, blood pressure, electromyography, and galvanic skin response) with the help of information and communication technology and supported querying the sensor network through semantic interoperability for the planned eHealth system. The proposed eHealth system consisted of 3 main components—SSN (eHealth sensor, patient, and URI), internet (eHealth server and KB), and eHealth client (patients and professionals). The proposed semantic model used *YANG to JavaScript Object Notation converter* to convert YANG semantic model data into JavaScript Object Notation semantic model data to achieve semantic interoperability, and then, stored it in a database or KB. Ganguly et al [24] proposed an ontology-based model for managing semantic interoperability issues in diabetic diet management. The development of the framework includes dialogue game rules, DSS with KB (rule library and database), dialogue model based on decision-making mechanism, dialogue game grammar, decision-making mechanism, and translation rules. Bouza et al [25] proposed a domain ontology-based decision tree algorithm and a reasoner to separate instances with more general features for recommender system (*SemTree*) that outperformed comparable approaches in recommendation generation. Chatterjee et al [11] focused on the creation of a meaningful, context-specific ontology (*University of Agder eHealth Ontology [UiAeHo]*) to model unintuitive, raw, and unstructured observations of health and wellness data (eg, sensors, interviews, and questionnaires) with semantic metadata and create a compact and logical abstraction for health risk prediction. Villalonga et al [26] proposed a holistic ontology model to annotate and classify motivational messages for physical activity coaching.

Most studies have developed ontologies that use OWL to solve data interoperability and knowledge representation problems. However, integrating personal health and wellness data, sensor observations, preference settings, semantic rules, semantic annotations, clinical guidelines, health risk prediction, and personalized recommendation generation remains as a problem in eHealth. We gathered ideas from existing studies to conceptualize our ontology design and implementation. In our previous study [11], we developed *UiAeHo* ontology to annotate personal and person-generated health and wellness data, sensor observations, health status in OWL format, combining SSN and Systematized Nomenclature of Medicine–Clinical Terms. Here,

we extended the study to annotate preference settings and activity status and tailored recommendation messages for activity e-coaching. The design and development of *UiAeHo* were focused more on obesity and overweight case studies However, this study focuses strictly on activity coaching and recommendation modeling. In addition, our proposed ontology was verified with semantic rules to generate different categories of recommendation messages for different cases. The high-level graphical representation of the proposed approach has been depicted in Figure 1 to show a distinction between *OntoRecoModel* and *UiAeHo* ontologies. *OntoRecoModel* annotates the following 3 types of data: sensor data (activity and weather), personal preference data, and personalized recommendations. Annotation of the sensor data in *OntoRecoModel* was based on the existing *UiAeHo* ontology following a semantic structure. Sensor data (activity data and contextual weather data) were included in this ontology design to exhibit that our *OntoRecoModel* can generate contextual and personalized recommendations in combination with personal preference data and semantic rules.

Figure 1. High-level representation of the proposed approach. SPARQL: SPARQL Protocol and Resource Description Framework Query Language; TDB: tuple database; UiAeHo: University of Agder eHealth Ontology.



Methods

Domain Ontology

Ontology supports flexibility in its design to solve real-world modeling and knowledge representation problems. It is a formal model of a specific domain, with the following essential elements: individuals or objects, classes, attributes, relationships, and axioms. The class diagram of a program written using object-oriented programming [28,29] visually depicts an ontology. The concept of ontology was created thousands of years ago in the philosophical domain, and it has the design flexibility of using existing ontology [29,30].

The open-world assumption knowledge representation style uses OWL, RDF, and RDF schema syntax. It can be optimized using the ontology model, and the consistency of its logic and structure can be verified using the ontology reasoning machine. An ontology O is defined as a tuple $\Omega=(\acute{C}, R)$, where $\acute{C}$ is the set of concepts and R is a set of relations. An ontology has a tree-like hierarchical structure (O_h) with the following properties [31,32]:

1. L=levels (O_h)=total number of levels in the ontology hierarchy, $0 \leq n \leq L$, where $n \in Z^+$ and n=0 represent the root node
2. $C_{n,j}$=a model classifying O at a level n; where, $j \in (0, 1,....., |C_n|)$
3. |C|=number of instances classified as class C
4. E=edge ($C_{n,j}$, $C_{n-1,k}$)=edge between node $C_{n,j}$ and its parent node $C_{n-1,k}$

Ontology Design Approach

An ontology can be designed in 5 ways: inspirational, inductive, synthetic, deductive, and collaborative [33]. We used a mixed method in our ontology design after combining the inspirational and deductive approaches. The inspirational approach helped us to identify the need for the ontology design, and the deductive approach focused more on the development of the *OntoRecoModel* model in Protégé. Moreover, the deductive approach helped us to adapt and adjust general principles to develop an anticipatory ontology of personalized activity recommendations as a study case. It includes general concepts that are filtered and refined to personalize specific domain subsets. The overall approaches were distributed in the following phases:

1. Literature search: we identified the necessary ontology components in healthy lifestyle management through a literature review, as described in the *Related Work* section. This study aimed to integrate ideas from the related ontology development in our proposed work.
2. Ideation: we discussed with 12 experts in the domain of information and communication technologies with research background in health care to design the concept of the ontology to fit in an activity e-coaching.
3. Annotation: we designed and developed the *OntoRecoModel* ontology to annotate personal preference data and motivational recommendation messages.
4. Rule base: we created a rule base for SPARQL query engine for query execution and personalized recommendation message generation (rule-based inference).
5. Verification: we verified our proposed *OntoRecoModel* ontology using simulated data against different test cases.

The feasibility study of the proposed *OntoRecoModel* consists of the following steps—(1) designing the ontology to fit in activity e-coaching concept; (2) modeling the ontology in the Protégé open-source platform and reasoning with HermiT reasoner; (3) integrating the concepts, such as annotation of personal preference data and motivational recommendation messages in *OntoRecoModel*; (4) implementing *OntoRecoModel* with logical axioms, declaration axioms, classes, instances, object properties, and data properties; and (5) setting up the rule base for ontology verification with SPARQL queries. We further discussed how interpretation can be associated with rule-based activity recommendation generation.

The specifications related to this study, as maintained by World Wide Web Consortium, are XML, URI, RDF, Turtle, RDF schema, OWL, SPARQL, and SWRL. The following terms are related to *OntoRecoModel* representation and processing:

1. Propositional variables (the atomic name of the truth value can be changed from one model to another)
2. Constants (the only propositional variables are TRUE and FALSE; thus, their truth values cannot be changed)
3. Operators (a set of logical connectors in each logic)

Here, we used operators, such as NOT, AND, OR, IMPLIES, EQUIV, and quantifiers (a set of logical quantifiers in a given logic). In this study, we used FORALL as the universal quantifier, EXISTS as the existential quantifier, quantification clause (a set of propositional variables connected by operators and quantifiers), clause (a quantification clause without any quantifier), formulas (a collection of clauses and quantified clauses linked together by logical operators), and process models (a collection of assignments for each propositional variable, so that when simplified, the process will lead to the constant TRUE).

Different open-access ontology editors are available in the market, such as NeOn Toolkit, Protégé, FOAF editor, TopBraid Composer, WebOnto Ontolingua Server, OntoEdit, WebODE, and Ontosaurus. The editors support the development of OWL-based ontologies. In addition, these editors support reasoning. The reasoner is a crucial component for using OWL ontology [11]. It derives new truths about the concepts that are modeled using OWL ontology. All queries on OWL ontology (and its imported closures) can be performed using reasoners [11,34,35]. Therefore, the knowledge in the ontology may not be explicit, and a reasoner is needed to infer the implicit knowledge to obtain the correct query results. If reasoner implementation is needed, the reasoner must be accessed through application programming interface (API). The OWL API includes various interfaces for accessing OWL reasoners. Reasoners can be categorized into 3 groups—OWL-DL, OWL–expression language, and OWL–query language [11,34-42]. This study considered Protégé (version 5.x) as an ontology editor for ontology design and development, OWLViz for ontology visualization, and HermiT (version 1.4.x; ∈ OWL-DL) reasoner for validating the ontology structure. In addition, we used an open-source Apache Jena Fuseki server [39] for SPARQL processing [43,44] with a tuple database (TDB). TDB supports Jena APIs [45,46] and can be used as a stand-alone high-performance RDF storage.

Ontology Modeling

Ontology modeling in Protégé can be classified into the following 2 categories: OWL-based and frame-based categories. We have used Protégé-OWL editor to model *OntoRecoModel* following the open KB connectivity protocol using classes, instances (objects), properties (object properties and data properties), and relationships. The steps of *OntoRecoModel* modeling in Protégé are described in Textbox 1.

Textbox 1. OntoRecoModel modeling steps in Protégé.

Step 1

Creation of a new Web Ontology Language project in Protégé and save it as a Turtle Resource Description Framework (RDF) format (OntoRecoModel.ttl)

Step 2

Create named classes under the superclass *owl:Thing*, maintaining consistency

- Create a group of classes (G=[C_1, C_2,....., C_n])
- Define disjoint classes (Cx ∩ Cy=[ø], where Cx and Cy G)
- Define subclasses
- Define disjoint subclasses

Step 3

Creation of Web Ontology Language properties after identifying classes and their properties

- Object properties (association between objects)
- Data properties (relates objects to XML schema datatype or rdf:literal)
- Annotation properties to annotate classes, objects, and properties

Step 4

Define nature of the properties

- Subproperties (A ⊆ B, where A and B are two nonempty sets)
- Inverse properties (x×y=I, where x, y A; I=identity element)
- Functional properties (X=A×X, where X is the set of all sequences <a1, a2,..., an> for a1, a2,.., an A)
- Inverse functional properties (for a function f: X → Y, its inverse f^{-1}: Y → X, where X, Y R)
- Transitive properties (S ⊆ S or if x=y and y=z, then x=z, where x, y, z ⊆ S set)
- Symmetric properties (if x=y, then y=x, where x, y ⊆ S set)
- Reflexive properties (x=x, where x R)

Step 5

Addition of existing ontology classes (eg, semantic sensor network ontology classes to annotate sensor observations)

Step 6

Define property domain (D) and range (R) for both object properties and data properties as axioms in reasoning

Step 7

Define property restrictions

- Qualifier restrictions (existential and universal)
- Cardinality restrictions (≥1)
- hasValue restrictions (datatype)

Step 8

Ontology processing with reasoner to check structural and logical consistency and compute the inferred ontology class hierarchy

- Blue color class in inferred hierarchy for reclassification
- Red color class in inferred hierarchy for inconsistent class

Step 9

Remove inconsistencies from the ontology tree using pruning method

Step 10

Query processing with SPARQL Protocol and RDF Query Language and storing the *Terse RDF Triple Language* file into tuple database for persistence

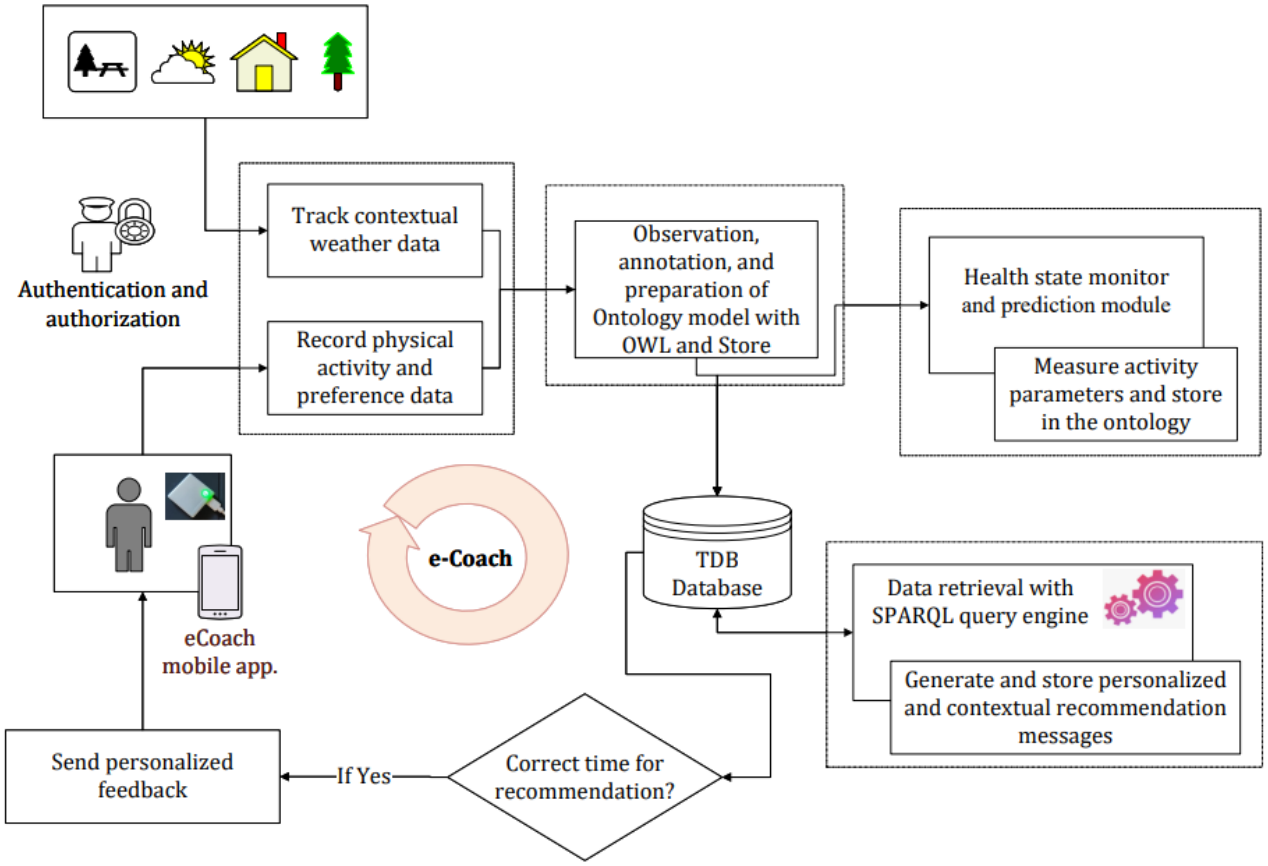
Ontology Implementation

Scope

We have planned to integrate the proposed *OntoRecoModel* model into an automatic activity coaching system for the semantic representation of activity sensor data, weather sensor data, personal preference data, and recommendation messages. The annotation of sensor data was pre-existing, and we used the concept from our previous study [11]. Furthermore, we showed a direction to use the proposed ontology model for automatic rule-based tailored activity recommendation generation with SPARQL queries to motivate individuals to maintain a healthy lifestyle. *OntoRecoModel* has gone one step forward to represent motivational recommendation messages beyond the *string* representation. Furthermore, the rule base helped to interpret the logic behind recommendation generation with logical AND and OR operations. We verified the ontology against a few test cases, which consisted of simulated data.

The targeted activity e-coach system has three modules, as depicted in Figure 2—(1) data collection and annotation module, (2) health state monitor and prediction module, and (3) recommendation generation module. In the data collection and annotation module, we showed a direction to annotate personal preference data essential for personalized recommendation generation. Health state monitor and prediction models periodically load individual activity data and analyze them using a data-driven machine learning (ML) approach or a rule-driven binary conditional approach. We considered a rule-driven approach for monitoring individual activity data using SPARQL queries. It determines whether a participant is sedentary or active over a day based on the recorded activity data. The annotated query processing results are stored in the database. Then, the personalized recommendation generation module combines the annotated SPARQL query results with the annotated preference data to generate tailored recommendation messages for motivation, which may help individuals to achieve their activity goals.

Figure 2. The modules of the e-coach prototype system. OWL: Web Ontology Language; SPARQL: SPARQL Protocol and Resource Description Framework Query Language; TDB: tuple database.



Annotation of Sensor Data

As shown in our previous study, this study achieved annotation of activity sensor data and contextual weather sensor data using pre-existing SSN ontology [11]. We used a similar logic; however, we annotated them more realistically. We examined the recorded activity parameters of different wearable activity sensors, such as Fitbit Versa, MOX2-5, and Garmin, and discovered that the following parameters are essential and common across these activity sensors: sedentary time, low physical activity (LPA) time, medium physical activity (MPA) time, vigorous physical activity (VPA) time, and total number of steps. Therefore, in this ontology, we annotated these activity parameters. Similarly, we analyzed data from different weather APIs, such as AccuWeather, Yr.no, and OpenWeather API. We found that the following observable weather parameters are common across these APIs: city, country, weather code, status, description, temperature, real feel, air pressure, humidity,

visibility, and wind speed. Thus, it may help *OntoRecoModel* to be functional, irrespective of the choice of standard activity sensor and weather APIs.

Annotation of Personal Preference Data

Personal preferences reflect individual expectations from an e-coach system. We planned to collect personal preference data at the beginning of the individual e-coaching session. We classified preference data into three categories: (1) activity goal settings, (2) response type for coaching, and (3) interaction type. Activity goals were categorized into 2 groups: personalized versus generic and direct versus motivational. The generic goals in activity coaching are the general activity guidelines set by the World Health Organization [47]. Personalized activity goals can be of multiple types (eg, weight reduction, staying active, body fat level, and proper sleeping). Direct goals tell the participant to perform direct activities (such as walking 2 km tomorrow).

In contrast, motivational goals inspire the participants to perform some tasks through persuasion (eg, If you walk 1 km further, you can watch an excellent soccer game). Response type for e-coaching can be either direct (eg, a pop-up message or notification to receive activity progression alert) or indirect (eg, graphical representation of activity progression). Individuals can be encouraged with personalized, evidence-based, and contextual response generation and its purposeful presentation (eg, graphic illustration, selection of colors, contrasts, visual aspects of movements, and menus, which are adjustable with device type). Interaction is an action that occurs owing to the mutual effect of ≥2 objects. The concept of 2-way effects is essential in interaction, not 1-way causal effects. The interaction types can be the mode (eg, style and graph), medium (eg, audio, voice, and text), and frequency (eg, hourly, daily, weekly, and monthly). Notification generation is a subcategory of interaction and may be persistent or nonpersistent.

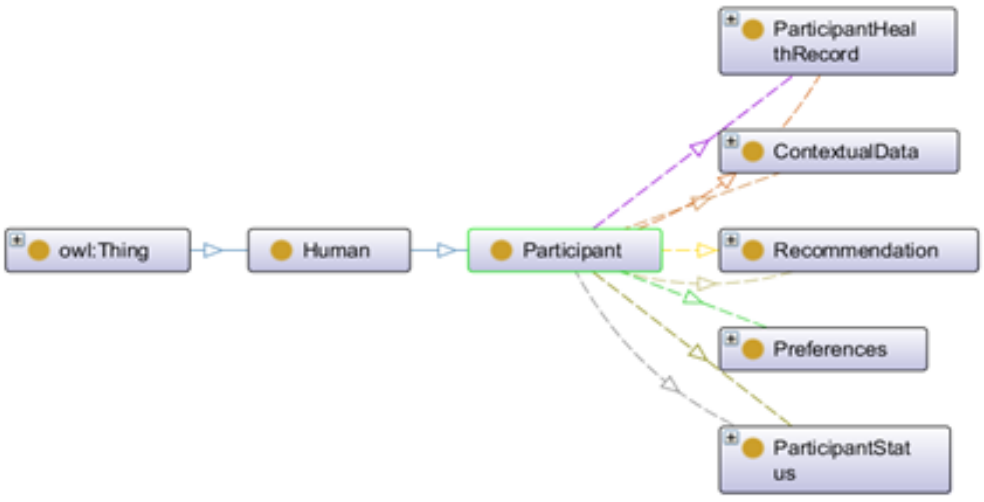
Annotation of Recommendation Messages

The recommender module generates personalized and contextual recommendations based on the prediction status. The recommendations can be direct (eg, pop-up notifications as alerts) or indirect (eg, visual representation). Direct or immediate notifications can contain 2 types of messages: to-do or formal (eg, You need to complete 1500 more steps in the next 2 hours to reach your daily goal) and informal (eg, Good work, keep it up! You have achieved the targeted steps). Therefore, we broke down the recommendation message concepts into intents and components. Intent defines the message's intention (eg, formal or informal). Message components define time, element (eg, data types in XML schema definition language), action (eg, pop-up and graphical visualization), and subject. An individual can receive >1 meaningful recommendation message based on the one-to-many relationship.

Ontology Classes and Properties

Figure 3 to 6 describe *OntoRecoModel* with mandatory classes to annotate the sensor, preference, and recommendation data.

Figure 3. High-level graphical representation of participant using OntoGraf in Protégé. OWL: Web Ontology Language.



Participant is the subclass of the human class (Figure 3). They have dedicated role and credentials (objectProperties: hasRole, hasPassword, and hasUniqueUserId) to authorize and authenticate themselves in the system. Participants are adults (both men and women), digitally literate, and clinically fit individuals. They are associated with the data properties such as hasAge, hasDesignation, hasEmail, hasFirstName, hasLastName, hasGender, and hasMobile. Each participant has their health record (hasHealthRecord), such as activity data; status (hasStatus), such as *active* or *inactive*; context information, such as weather status; preferences (hasPreferences); and recommendations (hasReceivedRecommendation).

Sensor data are ObservableEntity (Figure 4). Observation value is the subclass of ObservableEntity. ActivityDataValue and ExternalWeatherValue are the subclass of Observation value class. ActivityData and ActivityDataValue are linked to represent individual activity data. ActivityData class is a subclass of ParticipantHealthRecord and has objectProperty—hasBeenCollectedBy to represent associated activity data values (class: ActivityDataValue) as an observable entity. We have planned to collect activity data (such as steps, LPA, MPA, VPA, sleep time, and sedentary bouts) with a

wearable MOX2-5 activity sensor. In contrast, contextual data are observable weather-related data (city, country, weather code, status, description, temperature, real feel, air pressure, humidity, visibility, and wind speed), which are planned to be collected through the OpenWeather web interfaces. ContextData class is the subclass of ContextualData class and linked with ExternalWeatherValue to represent contextual weather data. TemporalEntity class represents the time stamp when the observational data have been captured and personalized recommendations have been generated (data property: hasDateTime).

Recommendation is a broad area, and we considered only activity recommendations in this study. ActivityRecommendation is a subclass of Recommendation class and parent to the MessageIntent and MessageComponent with the following objectProperties: hasMessageIntent and hasMessageComponent. MessageIntent class is the parent to ToDo and Informal classes with the following objectProperties: hasRecoInformal and hasRecoToDo (Figure 5). MessageComponent is the parent of Time, Element, Action, and Subject classes with the following objectProperties: hasTime, hasElement, hasAction, and hasSubject. Preferences is a subclass of the Qualifier class and related to the Goal, Interaction, and ResponseType (subclasses of the Preference class) with the following objectProperties: hasInteractionType, hasResponseType, and hasGoal. Preference class is a questionnaire-based method to receive participant's choices on goal setting, response type for e-coaching, and nature of interaction with the e-coach system.

Preference class has 3 subclasses: ResponseType, Goal, and Interaction. Goal class has 2 subclasses: Daily and Weekly (Figure 6). Each activity recommendations are either *generic* or *personalized*. Thus, recommendation generation depends on the assessment of the health status of the participants, regarding activity measurement and contextual information. Contextual data help recommend participants to plan indoor or outdoor activities based on external weather conditions. Table S1 in Multimedia Appendix 1 [48-51] summarizes the set of identified recommendation messages used for the test setup (ontology verification) and prepared based on positive psychology [52] and the concept of persuasion [48]. Recommendations generated on day *n* will reflect daily activity and contemplate what to perform on the day *n+1* to achieve the weekly goal. Preference data are personalized and customizable. All the necessary data for this study and their nature are summarized in Table S2 in Multimedia Appendix 2.

Description logic is the formal knowledge representation of ontology language, which provides a good trade-off between the expressiveness, complexity, and efficiency of knowledge representation and structured knowledge reasoning. We have the following proposition variables and recommended messages with their links to ensure that the paper is fully understood. Now, we need a set of clauses so that specific models can assign these variables to true, which triggers the sending of recommendations. SROIQ Description Logic [53] is the logic that provides the formal basis for OWL2 and has been used as the formal logic for reasoning in this study (Table S3 in Multimedia Appendix 3).

Figure 4. High-level graphical representation of observable data using OntoGraf in Protégé.

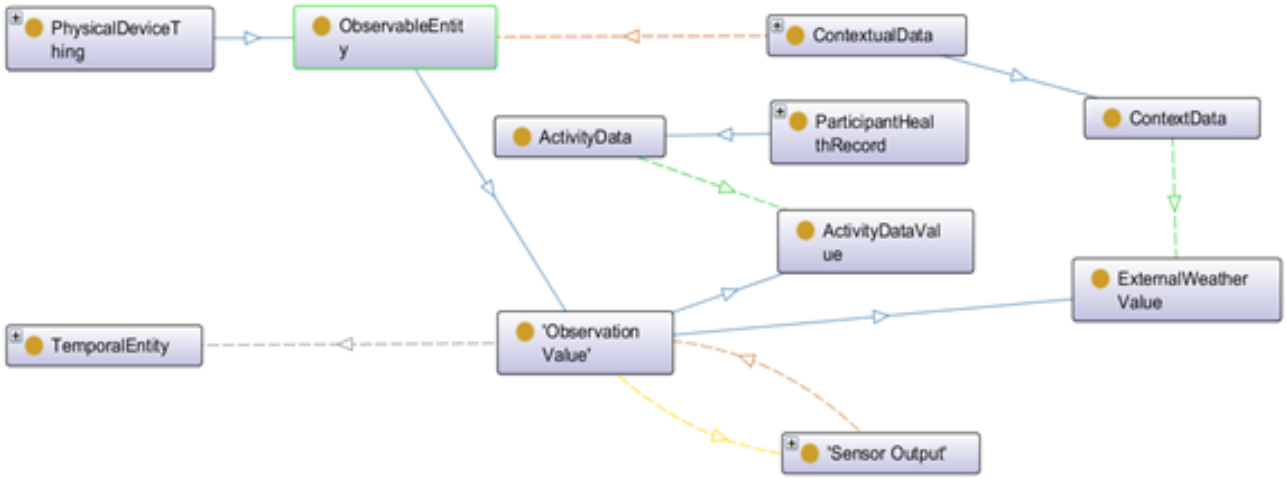


Figure 5. High-level graphical representation of recommendation using OntoGraf in Protégé.

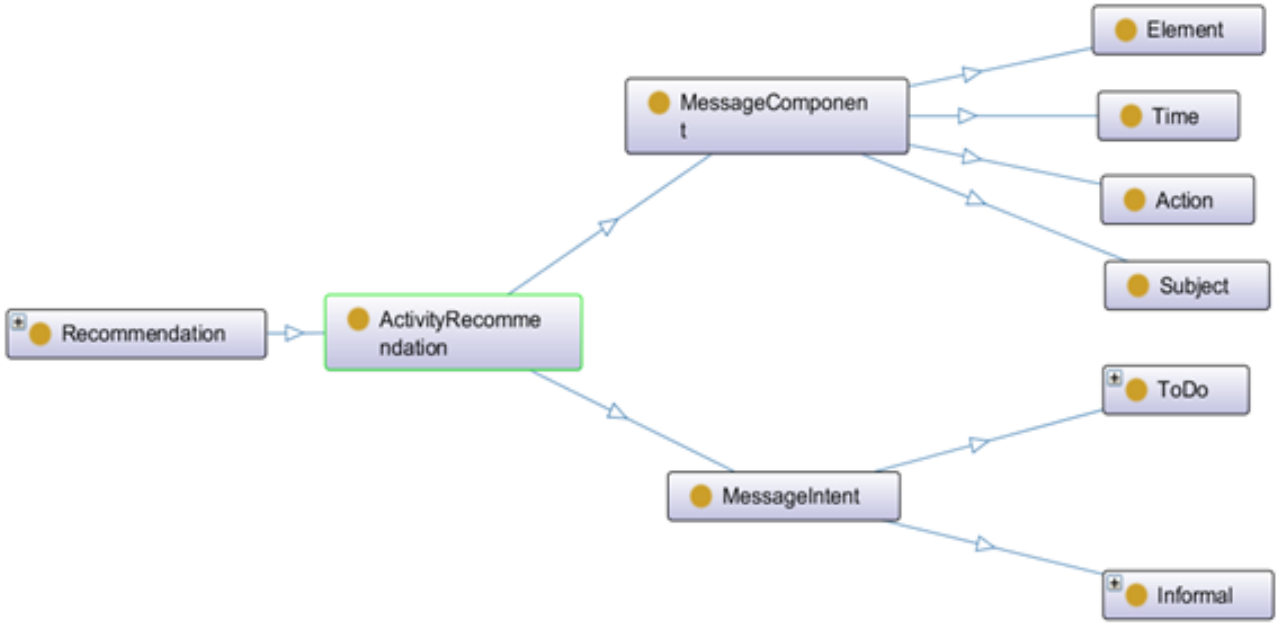
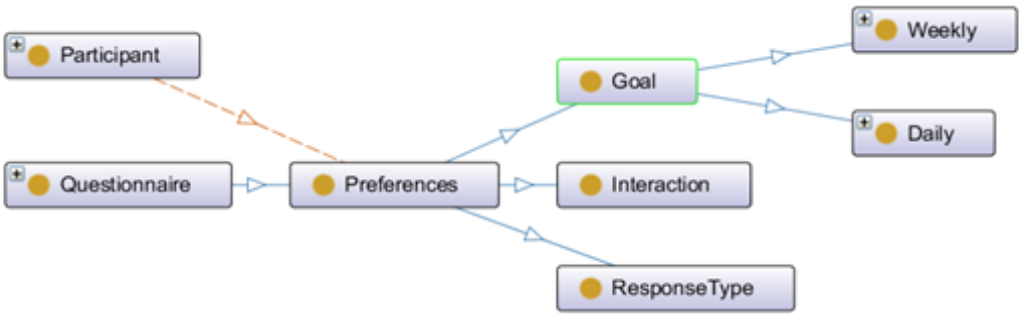


Figure 6. High-level graphical representation of preferences using OntoGraf in Protégé.



Ontology Verification

Test Cases With Simulated Data

We considered 8 test cases, as described in Table S4 in Multimedia Appendix 4, with simulated data for the proposed ontology verification. In the table, all the data are simulated. Therefore, no ethical approval was required. Cases 1 to 4 were associated with goal type—*generic* (World Health Organization standard guidelines to stay active for an entire week). Cases 5 to 8 were associated with goal type—*personalized*. More detailed description of different cases is provided in Textbox 2. The primary objective of the test cases was to check whether the daily step goal and daily sleep goal were achieved. The sedentary time and total time of VPA, MPA, and LPA were evaluated as a part of the secondary goal achievement. Daily goal achievement consisted of both primary objective and secondary objectives.

For all the test cases, the contextual weather data were considered constant (Table S5 in Multimedia Appendix 5). These test cases were added to the proposed ontology as individuals. SPARQL query processor engine processed the simulated data against certain test cases.

Textbox 2. Different test cases and their description.

Goal type: *Generic*

- Case 1 (11): Daily step goal and sleep goal are achieved.
- Case 2 (10): Daily step goal is achieved; however, sleep goal is not achieved.
- Case 3 (01): Daily step goal is not achieved; however, sleep goal is achieved.
- Case 4 (00): Daily step goal and sleep goal are not achieved.

Goal type: *Personalized*

- Case 5 (11): Daily step goal and sleep goal are achieved.
- Case 6 (10): Daily step goal is achieved; however, sleep goal is not achieved.
- Case 7 (01): Daily step goal is not achieved; however, sleep goal is achieved.
- Case 8 (00): Daily step goal and sleep goal are not achieved.

Note:

- 1 and 0 are two binary numbers and represent an on-off switch.
- 0 indicates that certain feature is false and 1 indicates that certain feature is true.
- Their combination (00, 01, 10, and 11) represents the following 2 combined features: daily step goal and daily sleep goal.
- The combination produces a total of 2^n possible test cases (00, 01, 10, and 11) for each goal type.

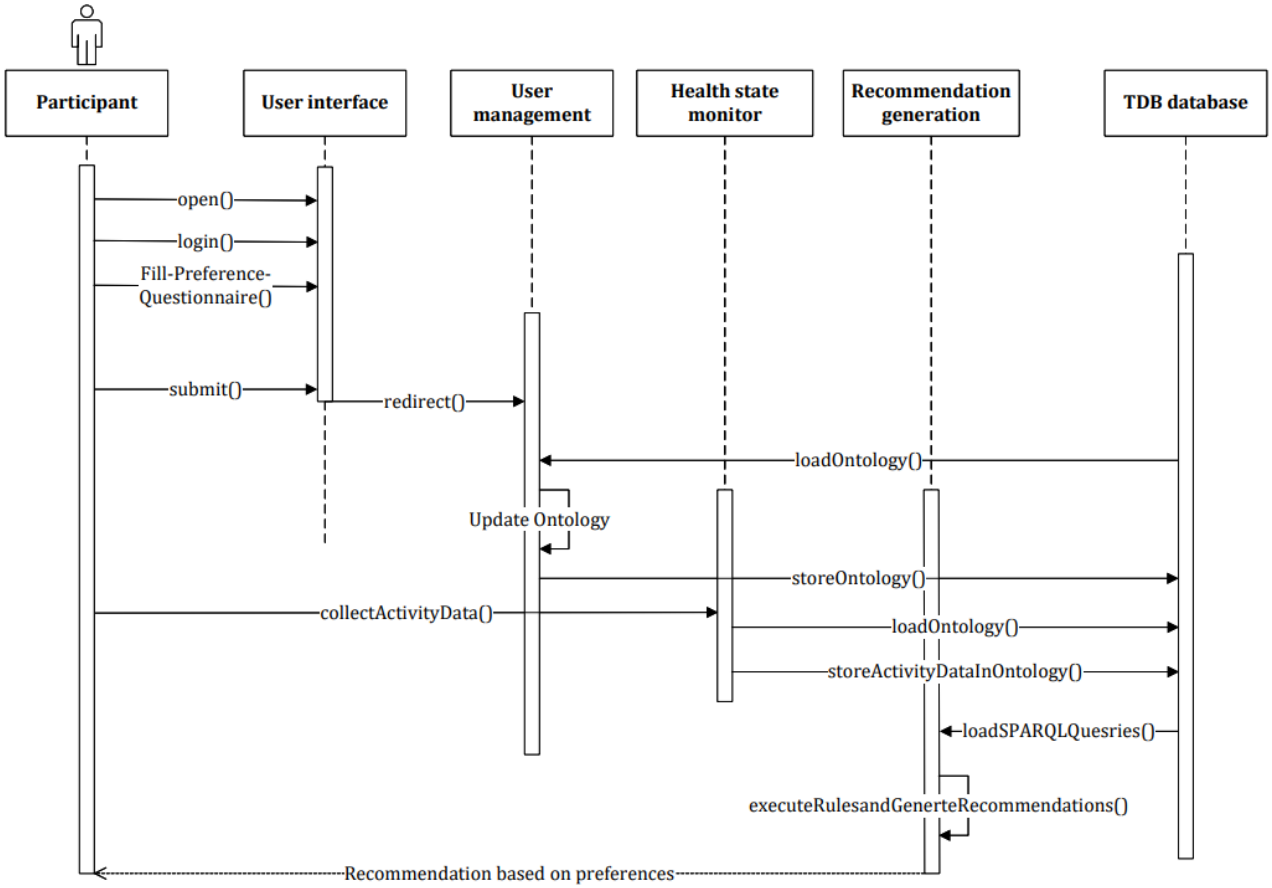
Rule Creation for SPARQL and Rule Execution

Rules were composed of cause (A) and effect (B) to imply A → B. For each of the conditions mentioned in Table S3 in Multimedia Appendix 3, the recommendation module performed a SPARQL query every day to determine the type of recommended message to be delivered to each participant, as shown in the Unified Modeling Language sequence diagram (Figure 7). The execution of each of the predefined semantic rules specified in Table S3 in Multimedia Appendix 3 depended on the performance of the SPARQL queries, and the rules were created according to clinical guidelines [48-50]. This study subdivided 12 semantic rules into activity-level classification (n=10, 83%), weather classification (n=1, 8%), and satisfiability (n=1, 8%). The added concepts and rules were relatively easy to follow and use.

Observable and measurable parameters related to the activities and context of the individual participants on the time stamp were obtained based on SPARQL queries at preference-based intervals. The rules 1 to 8 in Table S3 in Multimedia Appendix 3 assigned truth values to variables to ensure consistency. We confirmed with HermiT that the correct recommendation message was triggered for specific situations. However, it was necessary to ensure that no variable combination makes the entire formula unsatisfiable; that is, no model can satisfy the process. We confirmed that only 1 message was triggered at a time. In this study, we had a formal guarantee that 2 *once a day* messages cannot be triggered simultaneously and there cannot be a model output by HermiT every time for every possible variable combination. If we put the different variables used in the first 10 rules (Table S3 in Multimedia Appendix 3) into the propositional variables (Table S1 in Multimedia Appendix 1), we will have an exponential number of *possible participants*.

As 2 messages cannot be triggered simultaneously to meet the exact requirements, we added a rule (rule 11), and the variable used in the proposal starts *once a day*. If rule 11 is false, the entire ruleset (deemed as significant conjunction) will be set to false, and then, there will be no model as output, and we will be able to *debug* our rules if needed. If it is set to true, we will have a formal guarantee that regardless of the true value we put in the rule base, 2 *once a day* messages will not be triggered at the same time.

Figure 7. Unified Modeling Language sequence diagram for personalized recommendation generation and delivery. SPARQL: SPARQL Protocol and Resource Description Framework Query Language; TDB: tuple database.



Ethics Approval

We have used simulated data for this study. Therefore, participants' data have not been recorded or disclosed.

Results

An e-coach system can use the messages presented in this study (Table S1 in Multimedia Appendix 1) to improve individual activities with proper goal management. Therefore, the e-coach system must access these messages stored in a KB during tailored recommendation generation. Both the asserted and inferred knowledge obtained through the reasoning method will be helpful to determine the most appropriate message.

The TDB database, as shown in Figure 7, was used as a KB in this study. The test used to verify the performance and reliability of the proposed *OntoRecoModel* ontology included SPARQL queries and a rule base. In ontology verification, we generated personalized and contextual activity recommendations according to the semantic rules to improve the individual's physical activity to meet their activity goals. We executed all the semantic rules described in Table S3 in Multimedia Appendix 3 and used the Jena ARQ engine to run relevant SPARQL queries on the simulated data for the 8 test cases described in Table S4 in Multimedia Appendix 4. This helped to determine the type of recommendation message that would be generated, and we have presented our findings (rule-based recommendation generation for different cases) in Table 2. Several individual SPARQL queries are provided in Textbox S1 in Multimedia Appendix 6 as examples, and their results need to be combined to generate personalized recommendations to meet the e-coaching requirements. We achieved 100% precision in executing SPARQL queries to retrieve the necessary data.

Table 2 shows that participants can receive multiple motivational recommendation messages under ToDo and informal categories. The purpose of the e-coaching is to motivate participants (with motivational recommendation messages) for activities on day *n+1* based on the activity progression on day *n*, so that they can meet their weekly activity goals (*generic* or *personalized*) and maintain a healthy lifestyle. Proposition variable A-15 and A-16 (Table S1 in Multimedia Appendix 1) were the determinant of the weekly goal achievement and the delivery of the corresponding recommendation messages.

Table 2. Recommendation generation for different cases on day n for day n+1 (n>0).

Case	Activity status on day *n*	Recommendations for day *n+1*	
		ToDo	Informal
1	Goal achieved	A[a]-3, A-6, A-8, A-10, and A-12	A-13 and C[b]-1
2	Goal partially achieved	A-2, A-5, A-8, A-10, and A-11	A-14 and C-1
3	Goal partially achieved	A-1, A-5, A-7, A-9, and A-12	A-14 and C-1
4	Goal not achieved	A-1, A-5, A-7, A-9, and A-11	A-14 and C-1
5	Goal achieved	A-4, A-6, A-8, A-10, and A-12	A-13 and C-1
6	Goal partially achieved	A-4, A-5, A-8, A-9, and A-11	A-14 and C-1
7	Goal partially achieved	A-3, A-5, A-7, A-9, and A-12	A-14 and C-1
8	Goal not achieved	A-3, A-5, A-7, A-9, and A-11	A-14 and C-1

[a]A: activity recommendations.

[b]C: contextual recommendations.

Discussion

Principal Findings

The recommendation generation module used SPARQL queries and a rule base to generate personalized and contextual activity recommendations. There is no *false positive* situation based on the proposed ontology. According to the test cases in Table S4 in Multimedia Appendix 4, case 1 and case 5 achieved the daily activity goal; case 2, case 3, case 6, and case 7 achieved partial daily activity goal; and case 4 and case 8 ultimately failed to attain the daily activity goal. After combining the results of SPARQL queries with semantic rules, the related recommendation messages were updated, as shown in Table 2. The average execution time for all the SPARQL queries was between 0.1 and 0.3 seconds. The semantic rules described in Table S3 in Multimedia Appendix 3 represent the logic behind personalized recommendation message generation. The rule-based binary reasoning (if $\rightarrow$ 1, else $\rightarrow$ 0) helps to interpret the reason behind the delivery of a personal recommendation message.

The reasoning time of the proposed ontology was measured against the following reasoners available in Protégé: HermiT, Pellet, FaCT++, RacerPro, and KAON2; the corresponding processing times are shown in Table 3. The HermiT reasoner performed the best without reporting any inconsistencies.

Table 3. Comparative performance analysis of different reasoners available in Protégé.

Reasoner	Approximate reasoning time (seconds)
HermiT	2-3
Pellet	4-5
FaCT++	5-6
RacerPro	4-5
KAON2	5-6

The reading time after loading the ontology into the Jena workspace was approximately 1 to 2.5 seconds, with the *OWL_MEM_MICRO_RULE_INF* ontology specification (OWL full) in the *Terse RDF Triple Language* format, *in-memory* storage, and *optimized rule-based reasoner OWL rules*. Then, we used the Jena framework to query the ontology classes, predicates, subjects, and individuals in <1, <0.3, <0.4, and <2 seconds, respectively. Each ontology model (complete RDF diagram) was associated with a document manager (default global document manager: *OntDocumentManager*) to assist in processing ontology documents. All classes that represent the value of the ontology in the ontology API had *OntResource* as a general superclass with attributes (version information, comment, label, seeAlso, isDefinedBy, sameAs, and differentFrom) and methods (add, set, list, get, update, and delete). We implemented the RDF interface provided by Jena to maintain the modeled ontology and its instances in the TDB and load them back for further processing. Jena Fuseki was tightly integrated with TDB to provide a robust transactional persistent storage layer.

Limitations and Future Scope

As explained in this study, we conducted the overall experiment on simulated data in a modeled e-coaching environment. This concept must be tested after integrating with a real-time activity e-coaching system, in which actual participants will be involved. Here, the personalized recommendation generation is rule-driven and straightforward. In Figure 2, the health state monitor and prediction module can be upgraded using data-driven ML approaches, followed by annotation of prediction results into the ontology. However, it is the future scope of this study.

In our conceptualized activity e-coaching, the recommendation generation module successfully searched the KB of motivational

recommendation messages based on the rules in addition to the SPARQL results. The recommendation messages can be further personalized based on human behavior, liking for sports (eg, soccer), and the concept of reward bank. The components of the activity-related message can be further divided into indoor, outdoor, morning, afternoon, evening, and night activities. If a person has a dog and the e-coach system is aware of it, its recommendation generation module may suggest some activity recommendations involving the dog.

Table 2 shows that a participant can receive >1 recommendation message. It may lead to a message overloading problem. In future research, the recommendation process can be automated with ML algorithms (eg, time series and regression model) to select an optimal set of recommendations from feasible recommendations. The scope of the proposed ontology can be enhanced by conducting a study on a cluster of trials.

Conclusions

This study created the *OntoRecoModel* ontology to generate and model personalized recommendation messages for physical activity coaching. The proposed ontology not only semantically annotates recommendation messages, their intention, and components but also models personal preference data, individual activity data, and contextual weather information (required for personalized recommendation generation). Moreover, we successfully verified the use of the proposed ontology in rule-based recommendation generation using the SPARQL query engine. This study also showed a direction to categorize recommendation messages according to the defined ontology rules. Furthermore, reasoning has helped to organize the recommendation messages into multiple aspects. The recommendation message categorization, their semantic annotation, and the ontological SPARQL queries enable the recommendation generation module to generate them based on preferences, activity data, and contextual weather data.

The *OntoRecoModel* ontology uses the OWL-based web language to represent the collected data in the RDF triple storage format. The performance of the proposed ontology was evaluated using simulated data from 8 test cases. The structure and logical consistency of the proposed ontology were evaluated using the HermiT reasoner. In future studies, we will recruit actual participants following the inclusion and exclusion criteria to replicate the entire test scenario and assess the effectiveness of the recommendation generation plan for goal evaluation.

Acknowledgments

The authors acknowledge the funding and infrastructure obtained from the University of Agder, Center for e-Health, Norway, to conduct this study.

Authors' Contributions

AC contributed to conceptualization, formal analysis, investigation, methodology, obtaining resources, and writing the original draft. AP was involved in funding acquisition, reviewing, and supervision. All authors read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Propositional variables and corresponding recommendation messages.
[DOCX File , 17 KB - medinform_v10i6e33847_app1.docx]

Multimedia Appendix 2
Different data types used in this study and their nature.
[DOCX File , 15 KB - medinform_v10i6e33847_app2.docx]

Multimedia Appendix 3
In-context recommendation conditions and corresponding rules (rule base) for test setup.
[DOCX File , 16 KB - medinform_v10i6e33847_app3.docx]

Multimedia Appendix 4
Description of the test cases.
[DOCX File , 16 KB - medinform_v10i6e33847_app4.docx]

Multimedia Appendix 5
Description of the contextual weather data.
[DOCX File , 15 KB - medinform_v10i6e33847_app5.docx]

Multimedia Appendix 6
Selected list of SPARQL Protocol and Resource Description Framework Query Language queries used in this study.
[DOCX File , 16 KB - medinform_v10i6e33847_app6.docx]

References

1. Chatterjee A, Gerdes MW, Martinez SG. Identification of risk factors associated with obesity and overweight-a machine learning overview. Sensors (Basel) 2020 May 11;20(9):2734 [FREE Full text] [doi: 10.3390/s20092734] [Medline: 32403349]
2. Chatterjee A, Gerdes M, Martinez S. eHealth initiatives for the promotion of healthy lifestyle and allied implementation difficulties. In: Proceedings of the 2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). 2019 Presented at: 2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob); Oct 21-23, 2019; Barcelona, Spain. [doi: 10.1109/wimob.2019.8923324]
3. Wagner K, Brath H. A global view on the development of non communicable diseases. Prev Med 2012 May;54 Suppl:S38-S41. [doi: 10.1016/j.ypmed.2011.11.012] [Medline: 22178469]
4. Noncommunicable diseases. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [accessed 2022-02-24]
5. Obesity and overweight. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2022-02-24]
6. Chatterjee A, Prinz A, Gerdes M, Martinez S. Digital interventions on healthy lifestyle management: systematic review. J Med Internet Res 2021 Nov 17;23(11):e26931 [FREE Full text] [doi: 10.2196/26931] [Medline: 34787575]
7. Gerdes M, Martinez S, Tjondronegoro D. Conceptualization of a personalized ecoach for wellness promotion. In: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2017 Presented at: PervasiveHealth '17: 11th EAI International Conference on Pervasive Computing Technologies for Healthcare; May 23 - 26, 2017; Barcelona Spain. [doi: 10.1145/3154862.3154930]
8. Rutjes H, Willemsen M, IJsselsteijn W. Understanding effective coaching on healthy lifestyle by combining theory-and data-driven approaches. In: Proceedings of the Personalization in Persuasive Technology Workshop, Persuasive Technology 2016. 2016 Presented at: Personalization in Persuasive Technology Workshop, Persuasive Technology 2016; Apr 5, 2016; Salzburg, Austria URL: http://ceur-ws.org/Vol-1582/13Rutjes.pdf
9. Dijkhuis TB, Blaauw FJ, van Ittersum MW, Velthuijsen H, Aiello M. Personalized physical activity coaching: a machine learning approach. Sensors (Basel) 2018 Mar 19;18(2):623 [FREE Full text] [doi: 10.3390/s18020623] [Medline: 29463052]
10. Chatterjee A, Gerdes M, Prinz A, Martinez S. Human coaching methodologies for automatic electronic coaching (eCoaching) as behavioral interventions with information and communication technology: systematic review. J Med Internet Res 2021 Mar 24;23(3):e23533. [doi: 10.2196/23533] [Medline: 33759793]
11. Chatterjee A, Prinz A, Gerdes M, Martinez S. An automatic ontology-based approach to support logical representation of observable and measurable data for healthy lifestyle management: proof-of-concept study. J Med Internet Res 2021 Apr 09;23(4):e24656 [FREE Full text] [doi: 10.2196/24656] [Medline: 33835031]
12. Kim H, Park H, Min YH, Jeon E. Development of an obesity management ontology based on the nursing process for the mobile-device domain. J Med Internet Res 2013 Jul 28;15(6):e130 [FREE Full text] [doi: 10.2196/jmir.2512] [Medline: 23811542]
13. Sojic A, Terkaj W, Contini G, Sacco M. Modularising ontology and designing inference patterns to personalise health condition assessment: the case of obesity. J Biomed Semantics 2016 May 04;7(1):12 [FREE Full text] [doi: 10.1186/s13326-016-0049-1] [Medline: 29764473]
14. Kim H, Mentzer J, Taira R. Developing a physical activity ontology to support the interoperability of physical activity data. J Med Internet Res 2019 Apr 23;21(4):e12776 [FREE Full text] [doi: 10.2196/12776] [Medline: 31012864]
15. Lasierra N, Alesanco A, O'Sullivan D, García J. An autonomic ontology-based approach to manage information in home-based scenarios: from theory to practice. Data Knowl Eng 2013 Sep;87:185-205. [doi: 10.1016/j.datak.2013.06.004]
16. Yao W, Kumar A. CONFlexFlow: integrating flexible clinical pathways into clinical decision support systems using context and rules. Decision Support Syst 2013 May;55(2):499-515. [doi: 10.1016/j.dss.2012.10.008]
17. Chi Y, Chen T, Tsai W. A chronic disease dietary consultation system using OWL-based ontologies and semantic rules. J Biomed Inform 2015 Mar;53:208-219 [FREE Full text] [doi: 10.1016/j.jbi.2014.11.001] [Medline: 25451101]
18. Rhayem A, Ahmed Mhiri MB, Salah MB, Gargouri F. Ontology-based system for patient monitoring with connected objects. Procedia Comput Sci 2017;112:683-692. [doi: 10.1016/j.procs.2017.08.127]
19. Galopin A, Bouaud J, Pereira S, Seroussi B. An ontology-based clinical decision support system for the management of patients with multiple chronic disorders. Stud Health Technol Inform 2015;216:275-279. [Medline: 26262054]
20. Sherimon PC, Krishnan R. OntoDiabetic: an ontology-based clinical decision support system for diabetic patients. Arab J Sci Eng 2015 Dec 16;41(3):1145-1160. [doi: 10.1007/s13369-015-1959-4]
21. Hristoskova A, Sakkalis V, Zacharioudakis G, Tsiknakis M, De Turck F. Ontology-driven monitoring of patient's vital signs enabling personalized medical detection and alert. Sensors (Basel) 2014 Jan 17;14(1):1598-1628 [FREE Full text] [doi: 10.3390/s140101598] [Medline: 24445411]

22. Riaño D, Real F, López-Vallverdú JA, Campana F, Ercolani S, Mecocci P, et al. An ontology-based personalization of health-care knowledge to support clinical decisions for chronically ill patients. J Biomed Inform 2012 Jul;45(3):429-446 [FREE Full text] [doi: 10.1016/j.jbi.2011.12.008] [Medline: 22269224]
23. Jin W, Kim DH. Design and implementation of e-health system based on semantic sensor network using IETF YANG. Sensors (Basel) 2018 Mar 20;18(2):629 [FREE Full text] [doi: 10.3390/s18020629] [Medline: 29461493]
24. Ganguly P, Chattopadhyay S, Paramesh N, Ray P. An ontology-based framework for managing semantic interoperability issues in e-health. In: Proceedings of the HealthCom 2008 - 10th International Conference on e-health Networking, Applications and Services. 2008 Presented at: HealthCom 2008 - 10th International Conference on e-health Networking, Applications and Services; Jul 7-9, 2008; Singapore. [doi: 10.1109/health.2008.4600114]
25. Bouza A, Reif G, Bernstein A, Gall H. Semtree: ontology-based decision tree algorithm for recommender systems. In: Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008). 2008 Presented at: Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008); Oct 28, 2008; Karlsruhe, Germany URL: https://www.researchgate.net/publication/221466395_SemTree_Ontology-Based_Decision_Tree_Algorithm_for_Recommender_Systems
26. Villalonga C, op den Akker H, Hermens H, Herrera L, Pomares H, Rojas I, et al. Ontological modeling of motivational messages for physical activity coaching. In: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2017 Presented at: PervasiveHealth '17: 11th EAI International Conference on Pervasive Computing Technologies for Healthcare; May 23 - 26, 2017; Barcelona Spain. [doi: 10.1145/3154862.3154926]
27. Lasierra N, Alesanco A, Guillén S, García J. A three stage ontology-driven solution to provide personalized care to chronic patients at home. J Biomed Inform 2013 Jul;46(3):516-529 [FREE Full text] [doi: 10.1016/j.jbi.2013.03.006] [Medline: 23567539]
28. MOX2 Bluetooth LE activity monitor. Accelerometry.eu. URL: https://www.accelerometry.eu/products/wearable-sensors/mox2/ [accessed 2022-02-24]
29. Eriksson O, Johannesson P, Bergholtz M. Institutional ontology for Conceptual Modeling. J Inf Technol 2018 Jun 01;33(2):105-123. [doi: 10.1057/s41265-018-0053-2]
30. Bell D. An introduction to the Unified Modeling Language. IBM. URL: https://developer.ibm.com/articles/an-introduction-to-uml/ [accessed 2022-04-19]
31. Johnson I, Abécassis J, Charnomordic B, Destercke S, Thomopoulos R. Making Ontology-Based Knowledge and Decision Trees interact: an approach to enrich knowledge and increase expert confidence in data-driven models. In: Knowledge Science, Engineering and Management. Berlin, Heidelberg: Springer; 2010.
32. Gajderowicz B, Sadeghian A, Soutchanski M. Ontology enhancement through inductive decision trees. In: Uncertainty Reasoning for the Semantic Web II. Berlin, Heidelberg: Springer; 2013.
33. Holsapple CW, Joshi KD. A collaborative approach to ontology design. Commun ACM 2002 Feb;45(2):42-47. [doi: 10.1145/503124.503147]
34. Weather API. Open Weather. URL: https://openweathermap.org/api [accessed 2022-02-24]
35. Sirin E, Parsia B, Cuenca Grau B, Kalyanpur A, Katz Y. Pellet: a practical OWL-DL reasoner. SSRN J 2007. [doi: 10.2139/ssrn.3199351]
36. Parsia B, Matentzoglu N, Gonçalves RS, Glimm B, Steigmiller A. The OWL Reasoner Evaluation (ORE) 2015 competition report. J Autom Reason 2017;59(4):455-482 [FREE Full text] [doi: 10.1007/s10817-017-9406-8] [Medline: 30069067]
37. Knublauch H, Fergerson R, Noy N, Musen M. The Protégé OWL plugin: an open development environment for semantic web applications. In: The Semantic Web – ISWC 2004. Berlin, Heidelberg: Springer; 2004.
38. Editors. Semantic Web. URL: http://semanticweb.org/wiki/Editors [accessed 2022-02-24]
39. Reasoners. Semantic Web. URL: http://semanticweb.org/wiki/Reasoners [accessed 2022-02-24]
40. Shearer R, Motik B, Horrocks I. HermiT: A highly-efficient OWL reasoner. In: Proceedings of the Fifth OWLED Workshop on OWL: Experiences and Directions, collocated with the 7th International Semantic Web Conference (ISWC-2008). 2008 Presented at: Proceedings of the Fifth OWLED Workshop on OWL: Experiences and Directions, collocated with the 7th International Semantic Web Conference (ISWC-2008); Oct 26-27, 2008; Karlsruhe, Germany.
41. Tsarkov D, Horrocks I. FaCT++ description logic reasoner: system description. In: Automated Reasoning. Berlin, Heidelberg: Springer; 2006.
42. Haarslev V, Möller R. RACER system description. In: Automated Reasoning. Berlin, Heidelberg: Springer; 2001.
43. Getting started with Apache Jena. Apache Jena. URL: https://jena.apache.org/getting_started/index.html [accessed 2022-02-24]
44. SPARQL 1.1 Query Language. W3C. URL: http://www.w3.org/TR/2013/REC-sparql11-query-20130321/ [accessed 2022-02-24]
45. Appreciating SPARQL CONSTRUCT more. SPARQL. URL: http://www.snee.com/bobdc.blog/2009/09/appreciating-sparql-construct.html [accessed 2022-02-24]
46. Jena Ontology API. Apache Jena. URL: http://jena.apache.org/documentation/ontology/ [accessed 2022-02-24]
47. Cataletto M. World Health Organization issues new guidelines on physical activity and sedentary behavior. Pediatric Allergy Immunol Pulmonol 2020 Dec 01;33(4):167 [FREE Full text] [doi: 10.1089/ped.2020.29005.mca]

48. Sedentary behaviour for adults. KFL&A Public Health. URL: https://www.kflaph.ca/en/healthy-living/sedentary-behaviour-for-adults.aspx [accessed 2022-02-24]
49. Physical activity. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/physical-activity [accessed 2022-02-24]
50. How many pedometer steps should you aim for each day? Verywellfit. URL: https://www.verywellfit.com/how-many-pedometer-steps-per-day-are-enough-3432827 [accessed 2022-02-24]
51. Weather conditions. Open Weather. URL: https://openweathermap.org/weather-conditions [accessed 2022-04-19]
52. Seligman M, Csikszentmihalyi M. Positive psychology: an introduction. In: Flow and the Foundations of Positive Psychology. Dordrecht: Springer; 2014.
53. The SROIQ(D) description logic. Leslie Sikos. URL: http://www.lesliesikos.com/sroiqd-description-logic/ [accessed 2022-02-24]

Abbreviations

API: application programming interface
CDSS: clinical decision support system
KB: knowledge base
LPA: low physical activity
ML: machine learning
MPA: medium physical activity
OWL: Web Ontology Language
OWL-DL: Web Ontology Language–Description Logic
RDF: Resource Description Framework
SPARQL: SPARQL Protocol and Resource Description Framework Query Language
SSN: semantic sensor network
SWRL: Semantic Web Rule Language
TDB: tuple database
UiAeHo: University of Agder eHealth Ontology
VPA: vigorous physical activity

Edited by C Lovis; submitted 26.09.21; peer-reviewed by JR Yu, Á Sobrinho; comments to author 18.10.21; revised version received 05.03.22; accepted 21.04.22; published 23.06.22.

Please cite as:
Chatterjee A, Prinz A
Personalized Recommendations for Physical Activity e-Coaching (OntoRecoModel): Ontological Modeling
JMIR Med Inform 2022;10(6):e33847
URL: https://medinform.jmir.org/2022/6/e33847
doi:10.2196/33847
PMID:35737439

Original Paper

Experiences and Challenges of Emerging Online Health Services Combating COVID-19 in China: Retrospective, Cross-Sectional Study of Internet Hospitals

Fangmin Ge[1*], MPH; Huan Qian[1*], MD; Jianbo Lei[2], MD, PhD; Yiqi Ni[1], MS; Qian Li[1], BA; Song Wang[3], PhD; Kefeng Ding[1], MD, PhD

[1]The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China

[2]Center for Medical Informatics, Health Science Center, Peking University, Beijing, China

[3]School of Management, Zhejiang University, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Kefeng Ding, MD, PhD
The Second Affiliated Hospital
School of Medicine
Zhejiang University
No 88 Jiefang Road
Hangzhou, 310000
China
Phone: 86 13906504783
Email: dingkefeng@zju.edu.cn

Abstract

Background: Internet-based online virtual health services were originally an important way for the Chinese government to resolve unmet medical service needs due to inadequate medical institutions. Its initial development was not well received. Then, the unexpected COVID-19 pandemic produced a tremendous demand for telehealth in a short time, which stimulated the explosive development of internet hospitals. The Second Affiliated Hospital of Zhejiang University (SAHZU) has taken a leading role in the construction of internet hospitals in China. The pandemic triggered the hospital to develop unique research on health service capacity under strict quarantine policies and to predict long-term trends.

Objective: This study aims to provide policy enlightenment for the construction of internet-based health services to better fight against COVID-19 and to elucidate future directions through an in-depth analysis of 2 years of online health service data gleaned from SAHZU's experiences and lessons learned.

Methods: We collected data from SAHZU Internet Hospital from November 1, 2019, to September 16, 2021. Data from over 900,000 users were analyzed with respect to demographic characteristics, demands placed on departments by user needs, new registrations, and consultation behaviors. Interrupted time series (ITS) analysis was adopted to evaluate the impact of this momentous emergency event and its long-term trends. With theme analysis and a defined 2D model, 3 investigations were conducted synchronously to determine users' authentic demands on online hospitals.

Results: The general profile of internet hospital users is young or middle-aged women who live in Zhejiang and surrounding provinces. The ITS model indicated that, after the intervention (the strict quarantine policies) was implemented during the outbreak, the number of internet hospital users significantly increased (β_2=105.736, P<.001). Further, long-term waves of COVID-19 led to an increasing number of users following the outbreak (β_3=0.167, P<.001). In theme analysis, we summarized 8 major demands by users of the SAHZU internet hospital during the national shutdown period and afterwards. Online consultations and information services were persistent and universal demands, followed by concerns about medical safety and quality, time, and cost. Users' medical behavior patterns changed from onsite to online as internet hospital demands increased.

Conclusions: The pandemic has spawned the explosive growth of telehealth; as a public tertiary internet hospital, the SAHZU internet hospital is partially and irreversibly integrated into the traditional medical system. As we shared the practical examples of 1 public internet hospital in China, we put forward suggestions about the future direction of telehealth. Vital experience in the construction of internet hospitals was provided in the normalization of COVID-19 prevention and control, which can be demonstrated as a model of internet hospital management practice for other medical institutions.

(JMIR Med Inform 2022;10(6):e37042) doi:10.2196/37042

KEYWORDS

COVID-19; telehealth; e-consultation; dynamics of health care topics; China health system

Introduction

Digital health care is on the frontline in the fight against COVID-19, during which strict quarantine policy deterred most public access to onsite health care [1]. In response, online services from internet hospitals were quickly mobilized and dramatically upgraded as emergency relief measures for COVID-19 prevention and control in China. The pandemic accelerated the development of telehealth, including complementary services to existing departments, policy support, and rapid development and implementation of internet hospitals [2]. Internet hospitals, in general, are online medical platforms that combine online and offline access to medical institutions to provide a variety of telehealth services directly to patients. To date, 4 kinds of services have been equipped, including convenience services (booking appointments, checking test results), online medical services (electronic prescriptions), telemedicine (health education), and related support (follow-up consultations) [3]. There are 3 types of internet hospitals in China: government-oriented, hospital-oriented, and enterprise-oriented [4]. Routinely, Chinese patients go to hospitals, repeatedly queuing for registration, inquiries, or medical checkups in different departments [5]. Internet hospitals are able to render multiple services and offer essential medical support, surmounting geographical and time-related barriers [6]. Generally, internet hospitals in huge demand alleviate the imbalance of limited medical resources and increased burden of chronic diseases [7].

In China, the internet hospital is part of an ambitious plan, “Healthy China 2030 initiative,” released by the Chinese government in 2016. One of the important tenants of the plan is to make full use of internet technology to promote the integration of the internet and medical care, which is known as the “internet-plus-healthcare plan.” The construction of internet hospitals is an important part of the plan. The outbreak of COVID-19 greatly stimulated the explosive growth of internet hospitals. In 2016, there were only 25 internet hospitals [8]. By December 31, 2020, 1004 [9] had been established, and by June 2021, this number had increased to 1600 [10]. By 2017, the market size of internet medical services in China will be 32.5 billion yuan, with an estimated 250 million users [11]. The rapid excessive growth of internet hospitals urgently calls for a summary of relevant experience and construction guidance to ensure the healthy development of internet hospitals.

Worldwide, studies verified that, in the initial stage of the pandemic, there was a considerable amount of emerging literature on telehealth in most high-income countries [11]. For example, in the United States, approximately 60.0% of health care consumers reported that they first search online for information about an intended doctor through the internet hospital site or physician-rating websites before making a choice [12]. In addition, 59% of health care consumers confirm their choice based on the evaluation of doctors by internet hospitals or physician-rating websites [13]. However, the application of telehealth in resource-limited settings and low- and middle-income countries must be established to make the most of its potential and transform health care for the world’s population [14]. The definition, functions, boundaries, and hidden problems of internet hospitals are in urgent need of updated consensus to realize their potential and promote health care in the future.

SAHZU has played a pioneering and leading role in China's development of internet hospitals. SAHZU is located in Hangzhou, the cradle of well-known internet enterprises (eg, Alibaba and NetEase). SAHZU is one of the oldest public general hospitals, ranking in the top 10 among national general hospitals in China. As the “forerunner” of online medical care in China, the hospital took the lead in 2017 to launch an internet hospital and has continuously updated its capabilities. To date, almost 900,000 users have registered at SAHZU's Internet Hospital, which provides services to more than 1 million individuals per month. It initiated free online consultation and medication delivery services from January 27, 2020, to March 27, 2020, covering the lockdown period of the province. On March 15, 2020, China's first relevant group standard, Regulations on Online Consultation Services for Infectious Diseases, was released [15]. In the following 18 months, the hospital modified its strategies in promoting smart-assisted services as COVID-19 surged and subsided.

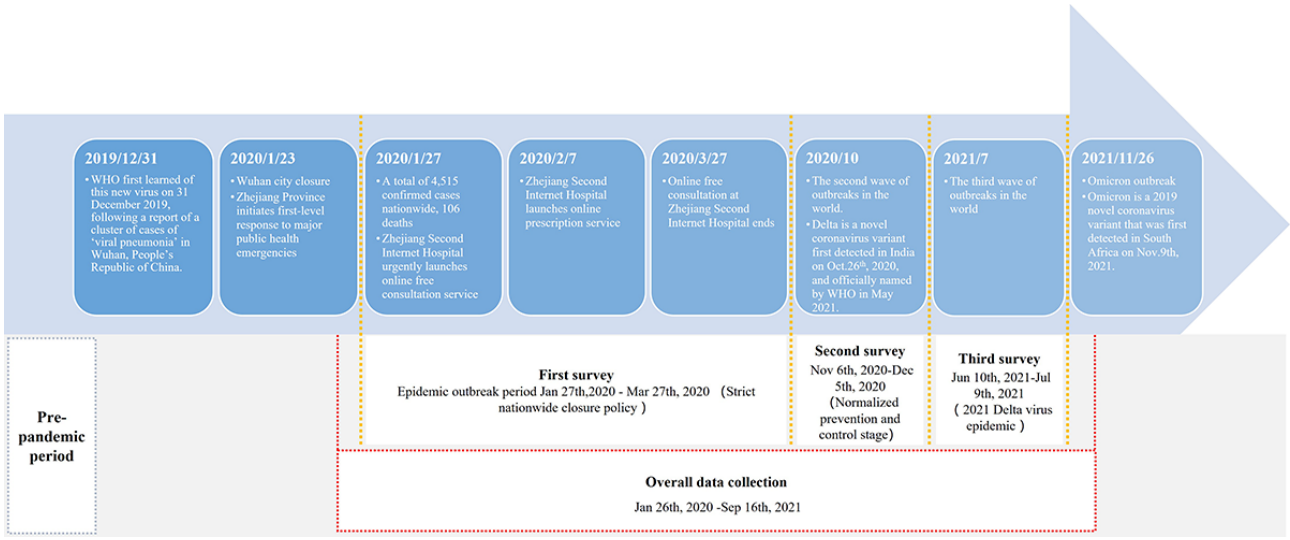
The pandemic triggered this unique research on health service capacity under strict quarantine policies. Only a few studies have reported the maximum usage of online hospital services during major public health emergencies. The strength of this study might be long-term follow-ups that further identify users’ behavior patterns and provide implications for the construction of global internet hospitals in the COVID-19 era. Of note, we captured the authentic demands of online health care seekers during the national blockade at the beginning of the pandemic. This study aimed to demonstrate the changes and trends in internet hospital applications during the COVID-19 pandemic in China.

Methods

Study Timeline

On January 30, 2020, the World Health Organization declared COVID-19 to be a public health emergency of international concern. SAHZU efficiently responded by establishing a free online consultation portal and received more than 10,000 consultation requests in the following 2 months. On February 7, 2020, a special online pharmacy was created to circumvent difficulties in obtaining regular medications. As an observational, cross-sectional study, we defined time periods as shown in Figure 1, whereas the investigation timeline primarily synchronized with national policies.

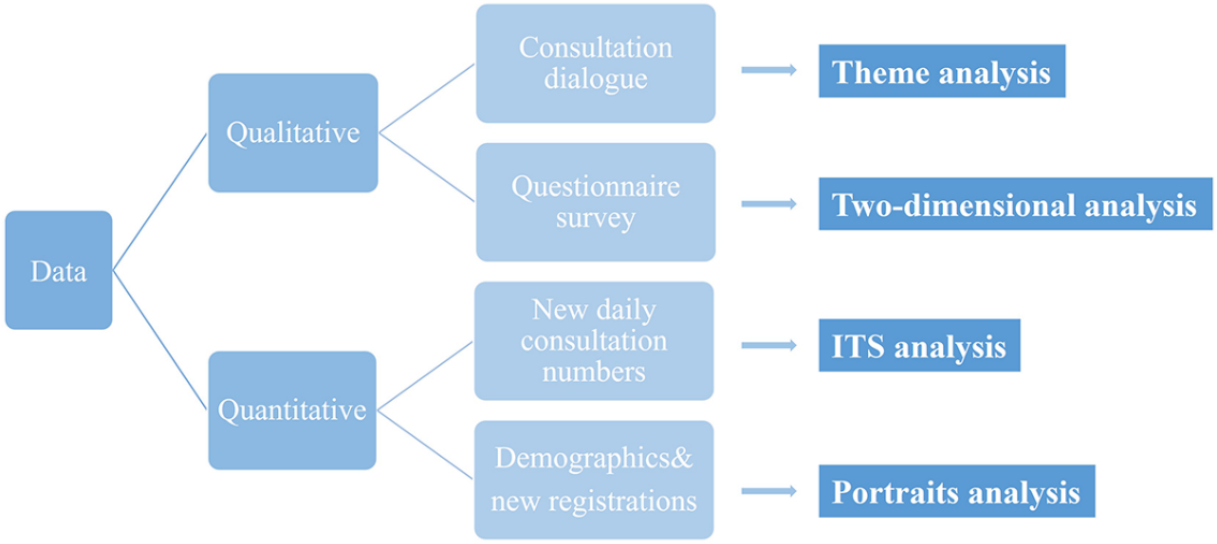
Figure 1. Study timeline. WHO: World Health Organization.



Data Collection and Analysis Methods

The data sources in this paper were divided into 2 categories: quantitative data and qualitative data. Multiple data sources and research methods are illustrated in Figure 2.

Figure 2. Data sources and analysis methods. ITS: intermittent time series.



Data Collection

Quantitative Data

Data from all internet hospital users, including daily registration numbers and user demographics (gender, age, geolocation, choice of specific department, and service needs), were collected on August 15, 2021.

Qualitative Data

As the basis of the initial research (first survey) and the qualitative research of this study, 15,990 conversation flows during the outbreak period under the national shutdown were reviewed in this study.

The first step was data reduction. This study is based on real-world data, and dialogue between doctors and patients contains grammar errors, dialects, and nonstandard expressions. Thus, the dialogue flow needs to be manually analyzed to address one of the major demands we classified.

Second, there was required prework before the analysis. For instance, more than one kind of demand could be included in a consultation case; therefore, the sum of the counts is not equal to the total number of cases. Moreover, a single user could have initiated several consultation cases with different needs or different patients because the platform allows users to initiate consultations for themselves or their family members using valid patient IDs.

Third, every entry in the data set was de-identified, including the gender, age, patient's chief complaint and previous diagnoses, the department of the clinician the patient consulted, and the content of the conversation without private information.

Thus, we preprocessed the dialogue between doctors and patients before formally processing the data.

Following the data reduction task, our theme analyses followed 3 steps: (1) theme formation, (2) theme matching along themes and patterns observed in the conversations, and (3) theme comparison across practice sites. Finally, we identified 8 major categories of demands from users, as listed in Table 1.

Table 1. Major categories of demands under the national shutdown period.

Number	Major categories of demands
1	Follow-up consultation
2	Drug refill
3	Consultation on common symptoms
4	Consultation on suspected symptoms of COVID-19
5	Information service
6	Psychological support
7	Rescheduling of treatment
8	Guidance on protective measures

The data were independently reviewed by 8 members of the investigative team, making methodological memos, theoretical memos, and preliminary interpretations. Individual researcher analyses and interpretations were discussed by the research team throughout the project. The themes and patterns were further refined, and new themes were cogenerated. All themes were developed through a process of articulating a unifying idea that represented interpretations from multiple data points. Conceptual labels were assigned to organize themes according to a common thread among ideas. In each step, themes were refined, whereby similarly labelled ideas were combined into themes and given more general labels. Disagreements were resolved through group discussion until consensus was reached. Finally, unstructured data were assigned to 8 major categories and then cleaned by the same researcher to keep the classification results unified. In our following question surveys, the users of our sample survey came from the users of our theme analysis.

To further investigate the needs and trends in patient behavior during online health service, we conducted 2 follow-up questionnaire surveys in November 2020 and June 2021, corresponding to the start and end of the second wave of the pandemic, respectively. The questionnaire (see Multimedia Appendix 1) includes 3 questions with fixed options: main reasons for using internet hospitals during the pandemic, changes in their way of accessing medical care after the pandemic, and major concerns about telehealth services. Interviewees were randomly selected from those who used internet services during the pandemic outbreak. At the start and finish of this second wave, 2 random online or telephone surveys were completed. Among the randomly selected 1100 actual internet hospital users for each survey, 1060 and 805 valid questionnaires were collected, respectively. All surveys were conducted anonymously, and each person only answered 3 questions from the predesigned questionnaire.

Ethics Approval

Our study protocol and procedures for informed consent before the formal survey were approved by the hospital ethics committee (Approval Number 2021-0761). Participants were required to answer a yes or no question to confirm their willingness to participate voluntarily. After confirmation of the question, the participant was directed to complete the questionnaire.

Analytical Methods

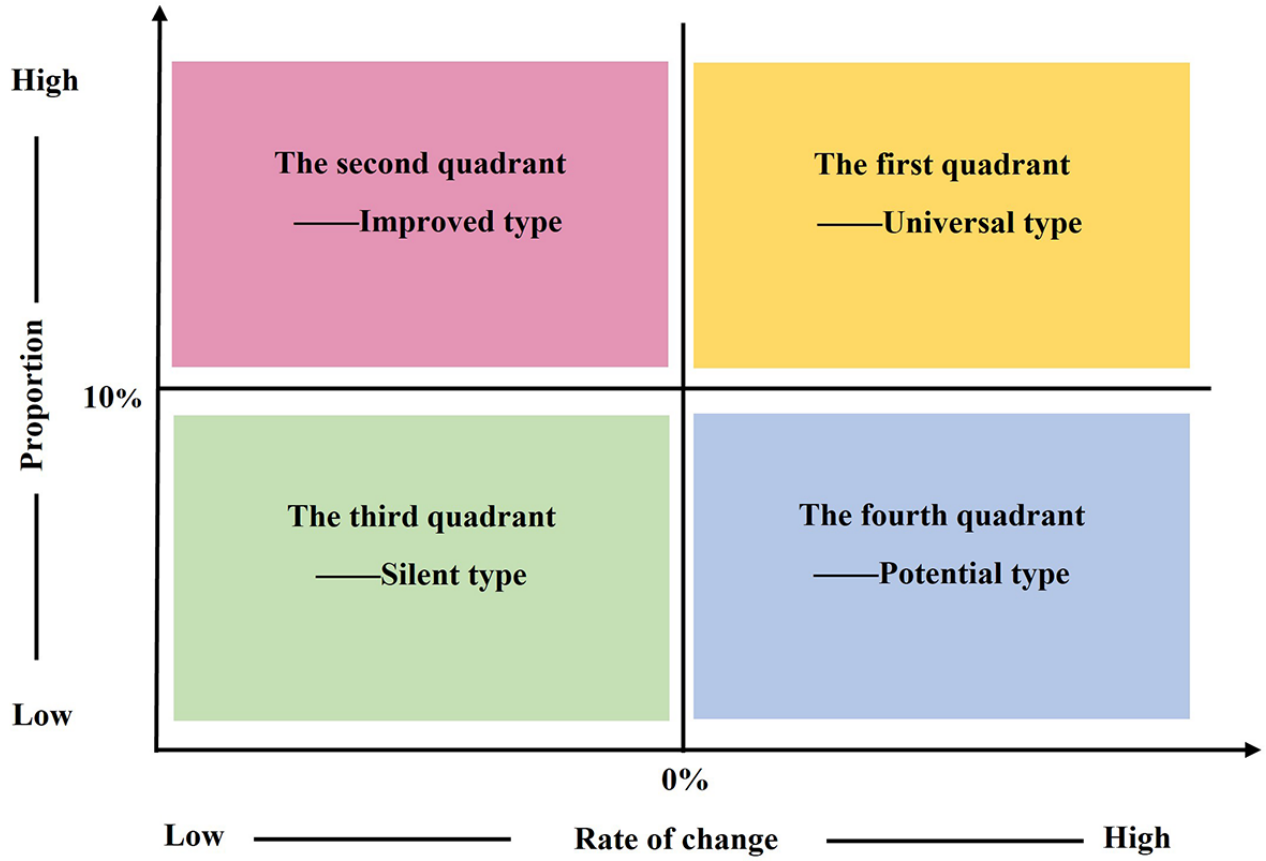
Interrupted Time Series Analysis

Interrupted time series (ITS) analysis was adopted to evaluate the impact of momentous emergency events (Multimedia Appendix 2). To identify use prompted by the pandemic, we assessed differences in new daily consultation numbers from July 15, 2019, to September 15, 2021, with a baseline period (November 1, 2019, to January 26, 2020) and after the national lockdown response (March 18, 2020, to June 30, 2020). We conducted a multistage comparison of the development of online medical treatment (October 1, 2020, to April 10, 2021, and April 11, 2021, to September 16, 2021); thus, the long-term impact of the pandemic was also considered in the ITS analysis.

2D Analysis

To gain deep insight into the changes in patient demands, we modeled the proportion and increase or decrease of each demand through a 2D model. A matrix chart with 4 quadrants was designed; the abscissa is the rate of change, and the ordinate is the distribution. Different functions fall into typical quadrants, as expressed in Figure 3.

Figure 3. 2D model.



The first quadrant is a universal type, with a large proportion of demand and upwards change in different periods. The second quadrant is an improved choice, relatively stable, with a large proportion of demand but downward changes. The third quadrant is a silent type, with a small proportion of demand and downward changes in different periods. The fourth quadrant is the potential needs; the proportion of demand is small, but the changes are upwards. Furthermore, this 2D model was also used to classify the data on changes in patients' ways of accessing medical care and concerns.

Statistical Analysis and Visualization

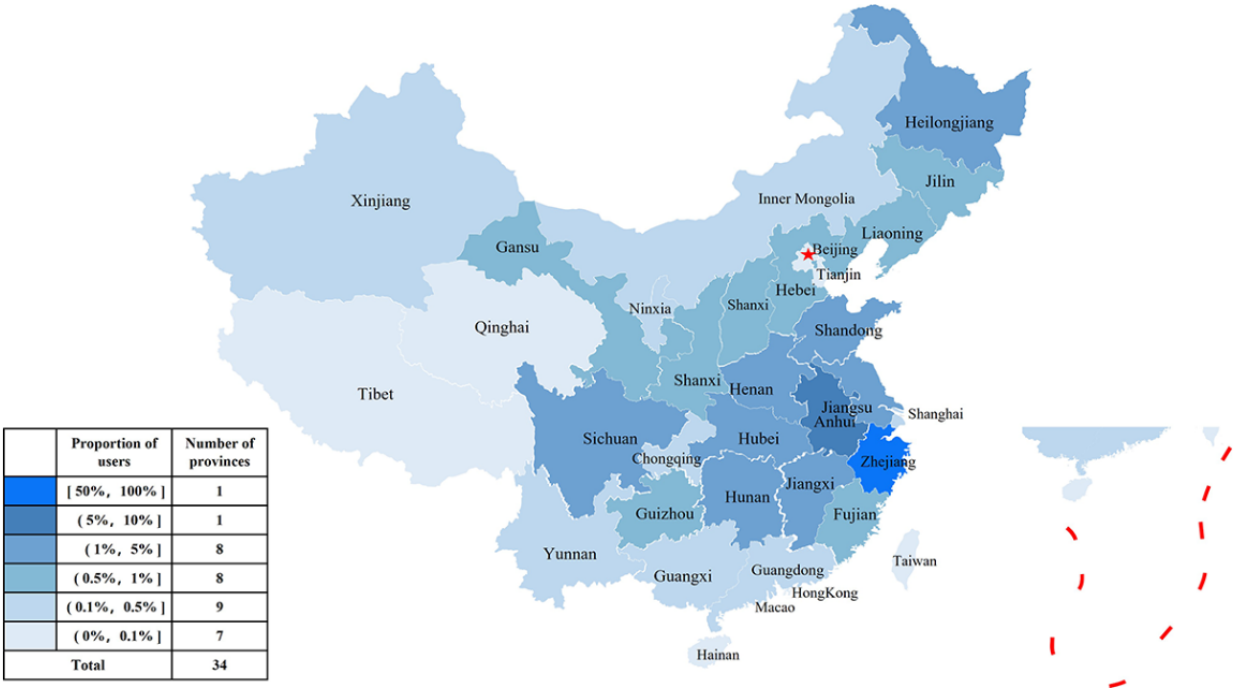
SPSS 25.0 software (28.0.1; IBM Corp, Armonk, NY) was used for data analysis. Frequencies and percentages were used for categorical data. The chi-square test was adopted for comparisons between groups. A 2-tailed *P* value <.05 was considered statistically significant.

Results

Demographic Composition of All Users and Their Online Choices

The demographic composition of internet hospital service users was analyzed based on August 15, 2021 (Figures 4 and 5). In comparison, we also describe the demographic characteristics of online patients from January 27, 2020, to March 27, 2020 (Figure 6) and those active in online consultation during quarantine (Figure 7).

Figure 4. Regional composition of all Second Affiliated Hospital of Zhejiang University School of Medicine (SAHZU) internet hospital users before August 15, 2021.



	Proportion of users	Number of provinces
	[50%, 100%]	1
	(5%, 10%]	1
	(1%, 5%]	8
	(0.5%, 1%]	8
	(0.1%, 0.5%]	9
	(0%, 0.1%]	7
Total		34

Figure 5. Regional composition of Second Affiliated Hospital of Zhejiang University School of Medicine (SAHZU) internet hospital users in Zhejiang Province before August 15, 2021.

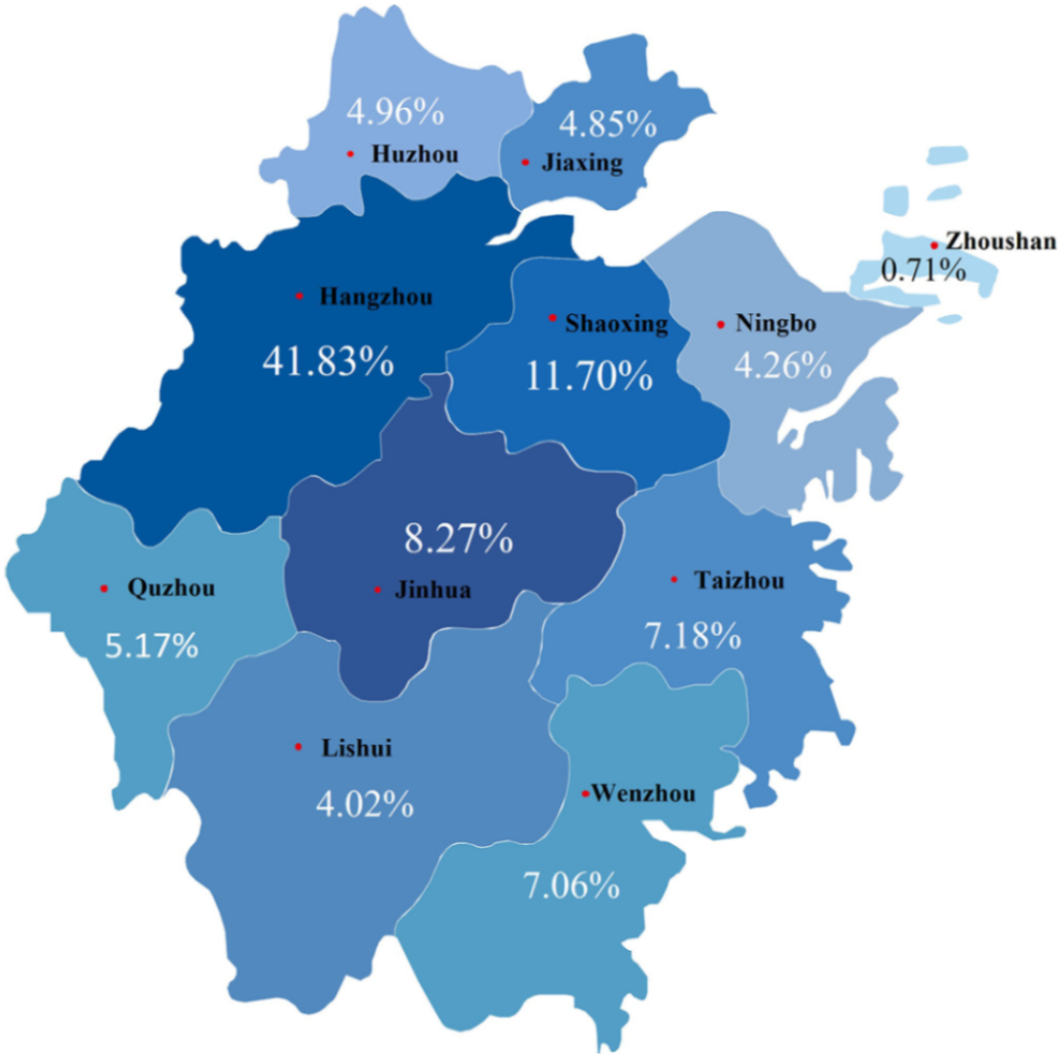


Figure 6. Comparison of the age group composition of Second Affiliated Hospital of Zhejiang University School of Medicine (SAHZU) internet hospital users before August 15, 2021, and those during the pandemic outbreak period from January 27, 2020, to March 27, 2020.

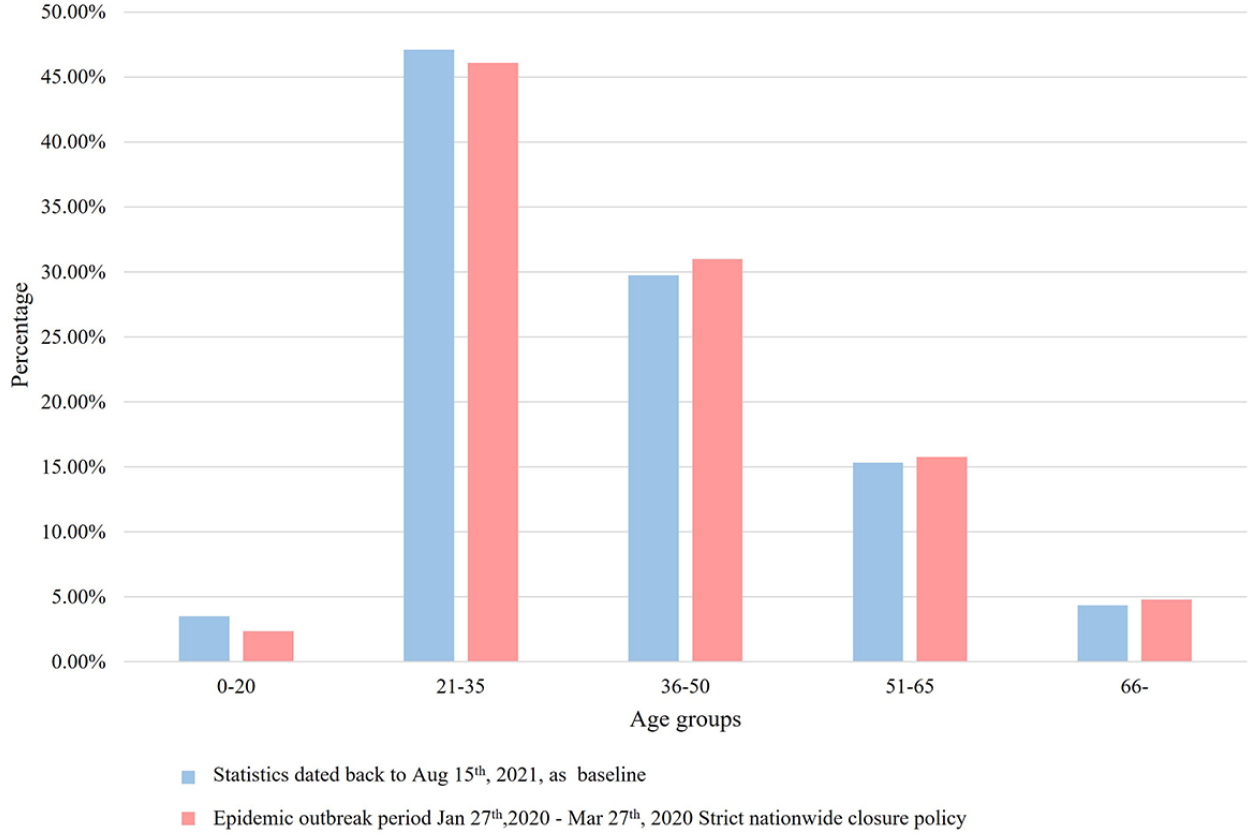
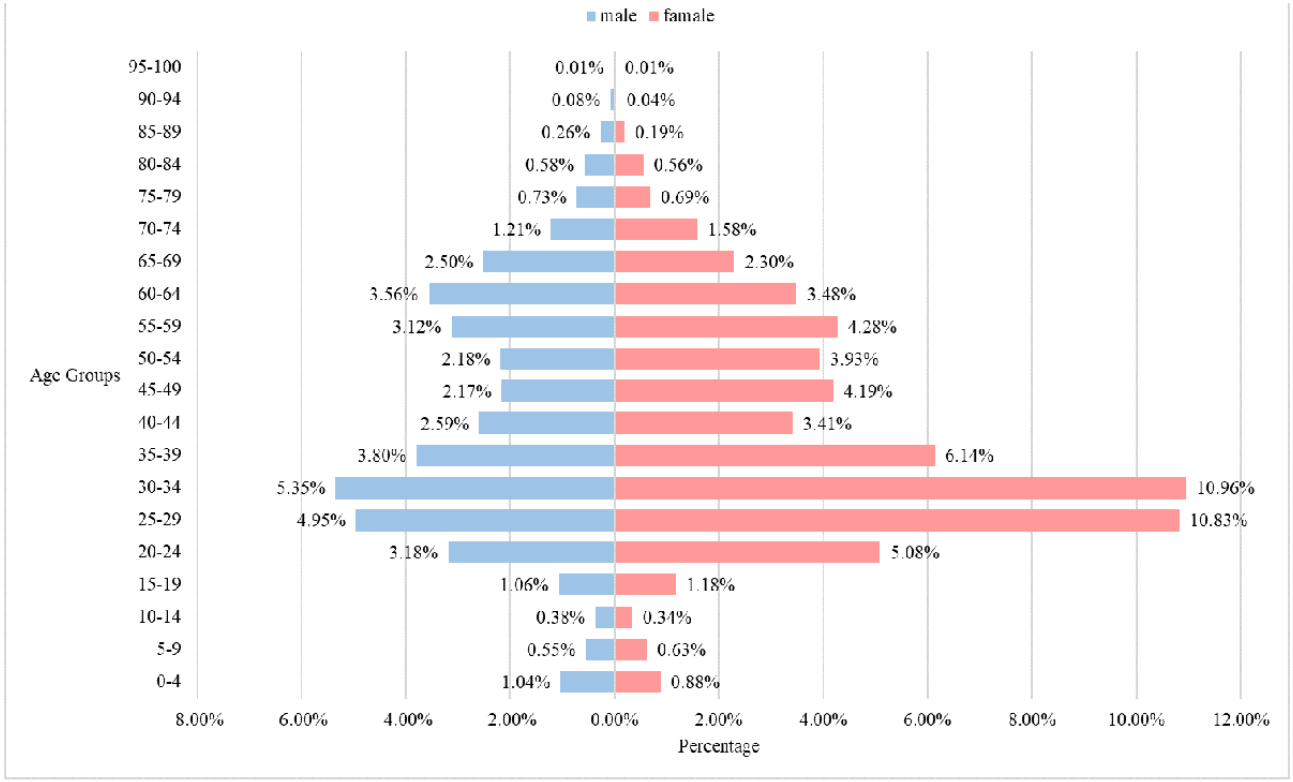


Figure 7. Age and gender distributions for all online consultation participants during the pandemic outbreak period (January 27, 2020, to March 27, 2020). Information for cases such as infants or older adults was apparently completed by their family members.



We illustrate a change in department choice among requests for online consultation (Figure 8), mainly an increase for the gynecology, endocrinology, obstetrics, and orthopedics departments. During the COVID-19 outbreak, the SAHZU internet hospital began to offer free online consultations and assembled a highly responsive team that remained stable afterwards. With respect to the departments focused on the most common and chronic diseases, the number of online consultations increased significantly over the long term. The choice of internet hospital functions (Figure 9) showed that users were mainly interested in appointment registration and test result queries, followed by online consultation. During the initial outbreak, the use of online consultation services such as team consultation, picture and text consultation, and medication consultation increased significantly. In the fairly stable period that followed, users gradually returned to making appointments and requesting test results.

Figure 10 lists the departments that received the most consultation requests in the first outbreak period, January 27, 2020, to March 27, 2020. We identified 8 major categories of demands (Figure 11), which have shown different usage patterns over time. The distribution of the major demands was plotted in chronological order, revealing the trends in detail (Multimedia Appendix 3).

Figure 8. The clinical departments that received the most consultation requests.

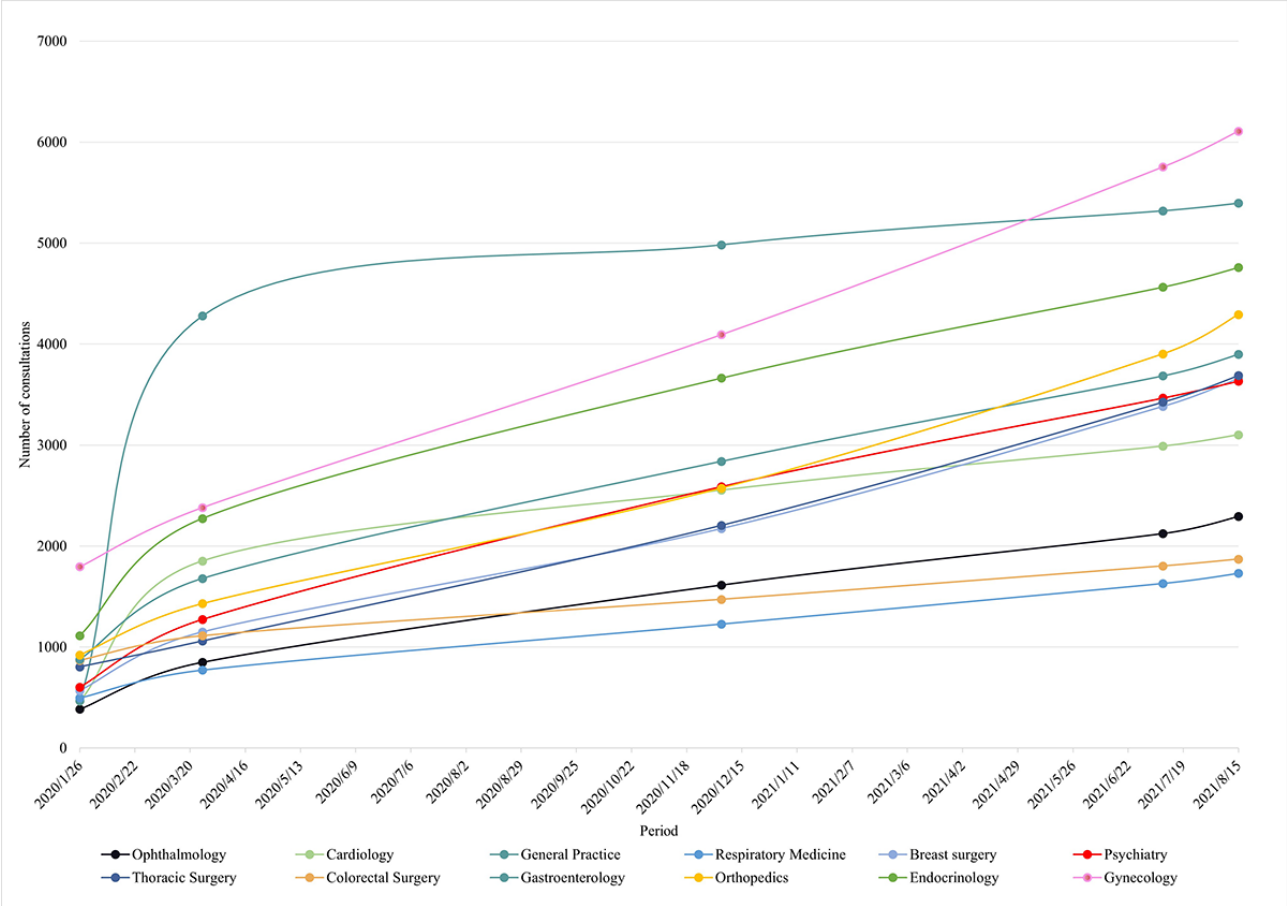


Figure 9. Changes in Second Affiliated Hospital of Zhejiang University School of Medicine (SAHZU) internet hospital users' demands over time.

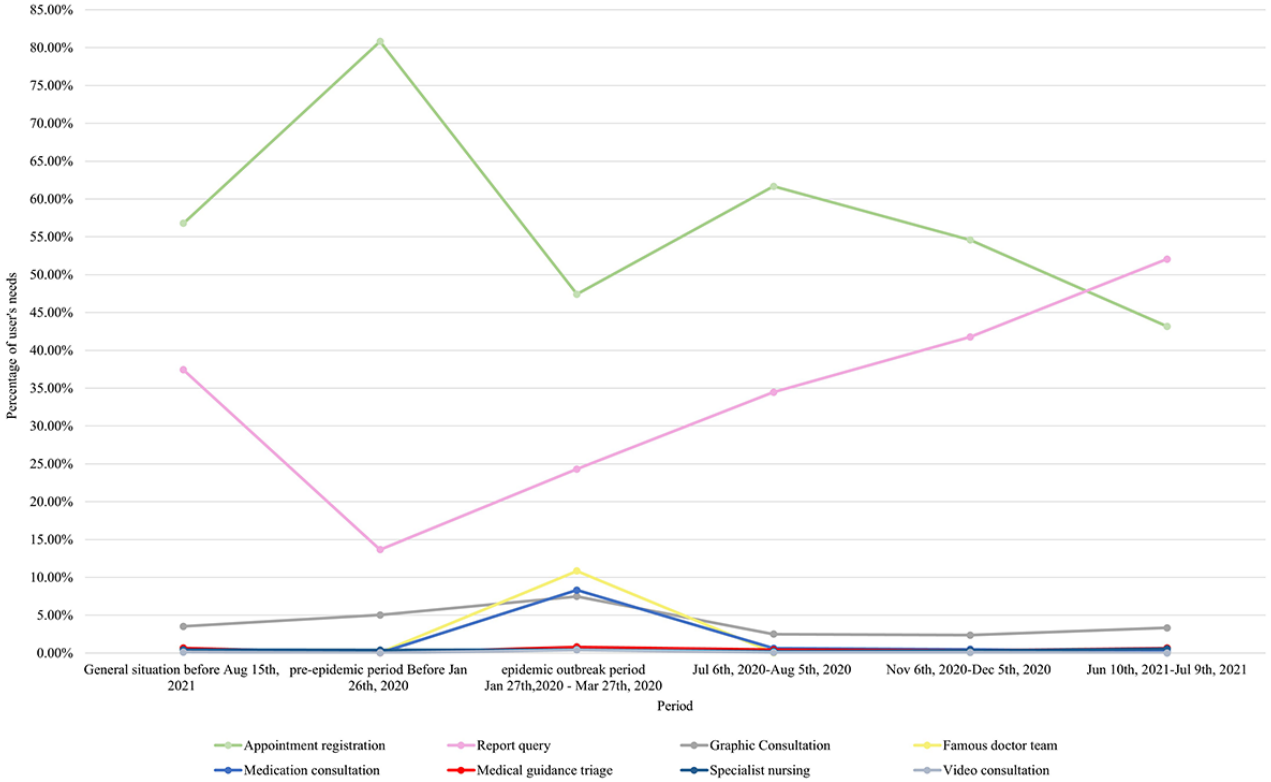


Figure 10. The most frequently consulted departments from January 27, 2020, to March 27, 2020.

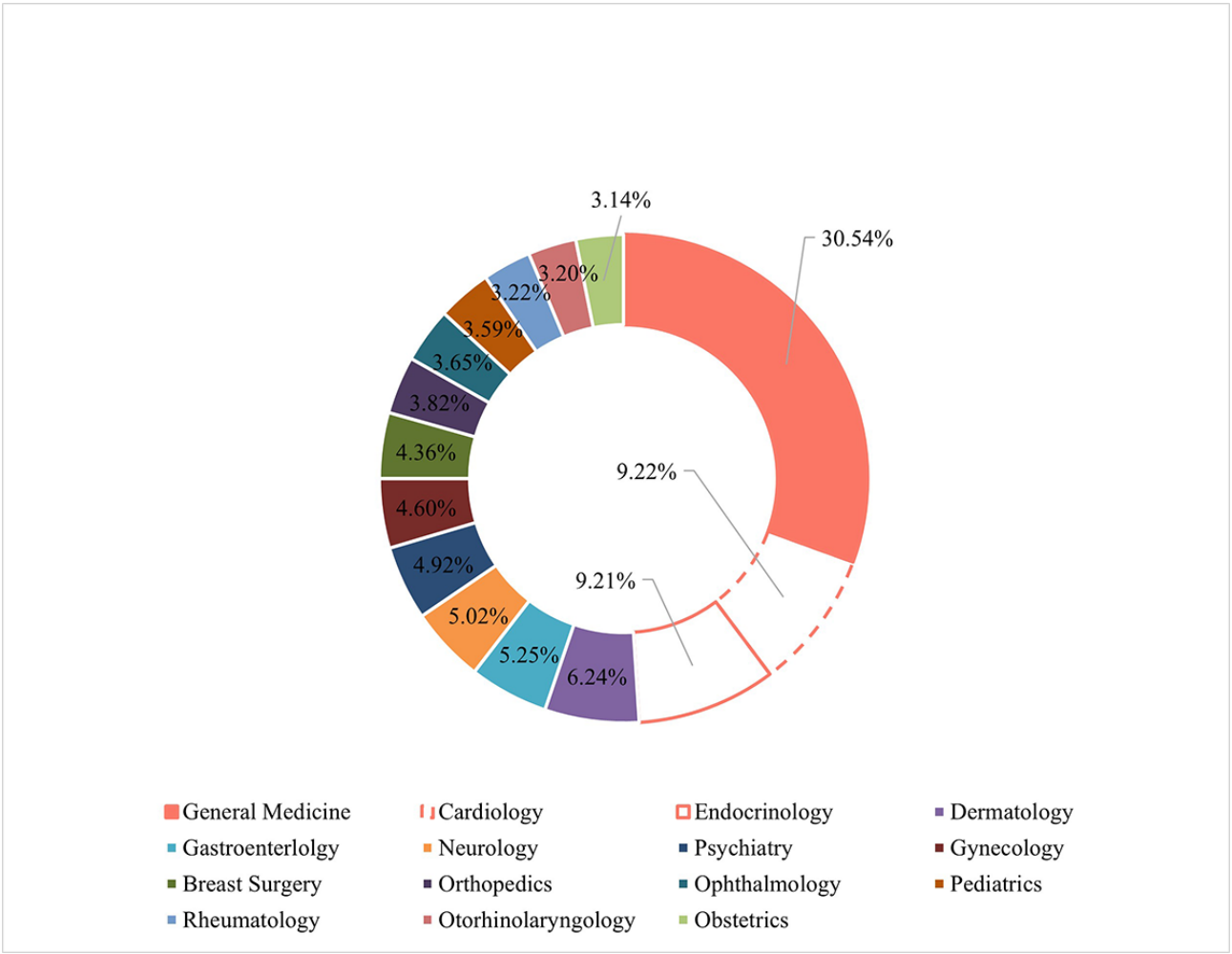
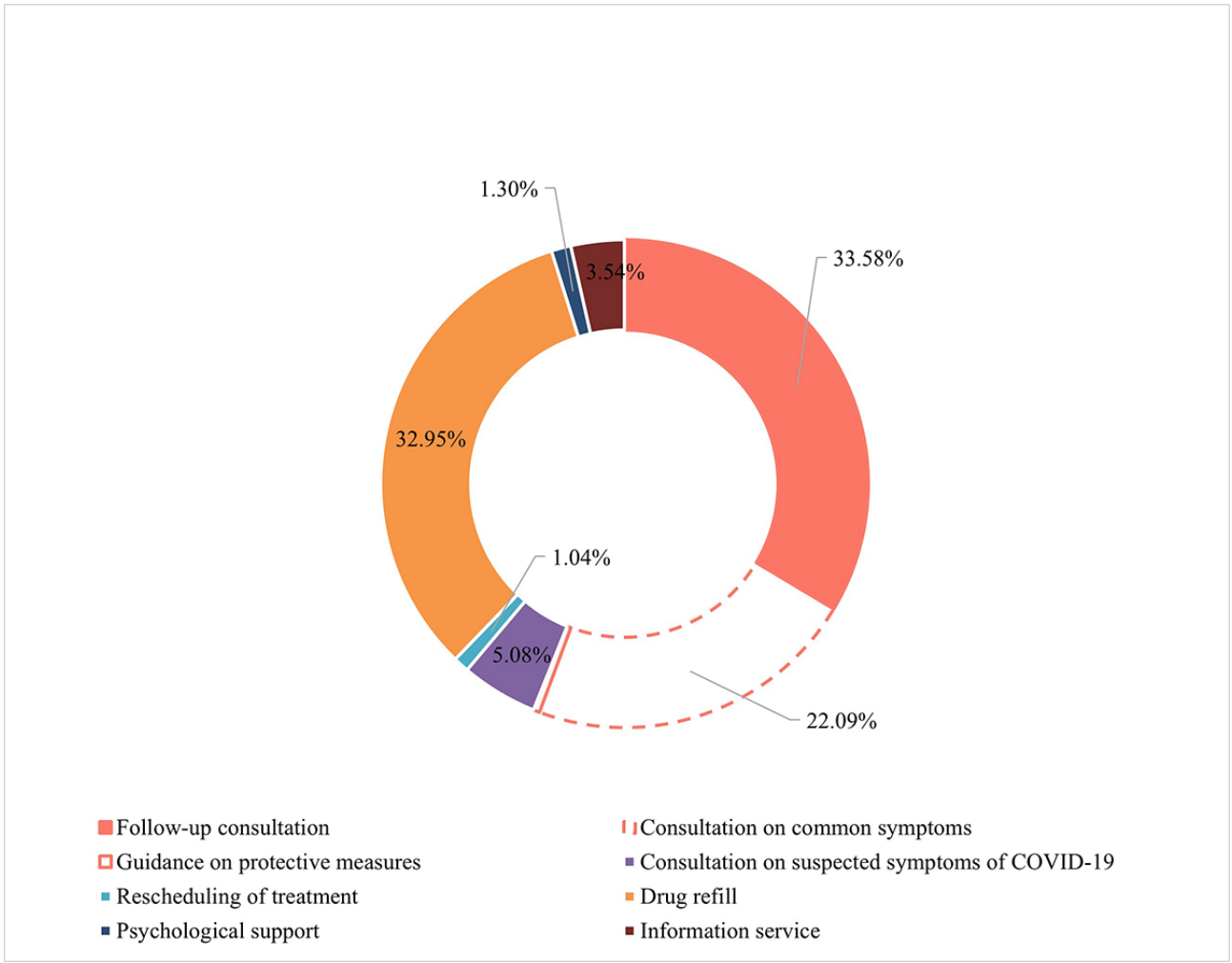


Figure 11. The main demands from Second Affiliated Hospital of Zhejiang University School of Medicine (SAHZU) internet hospitals patients from January 27, 2020, to March 27, 2020.



Changes in New Registrations

The number of new registrations changed over time, while the main users of the SAHZU internet hospital remained 21- to 35-year-old women (Figure 12). Sudden peaks corresponded to the free consultation activities and Spring Festival. The number of consultations increased significantly during the national quarantine (Figure 13).

Figure 12. Daily Second Affiliated Hospital of Zhejiang University School of Medicine (SAHZU) internet hospital user growth from August 16, 2019, to August 15, 2021.

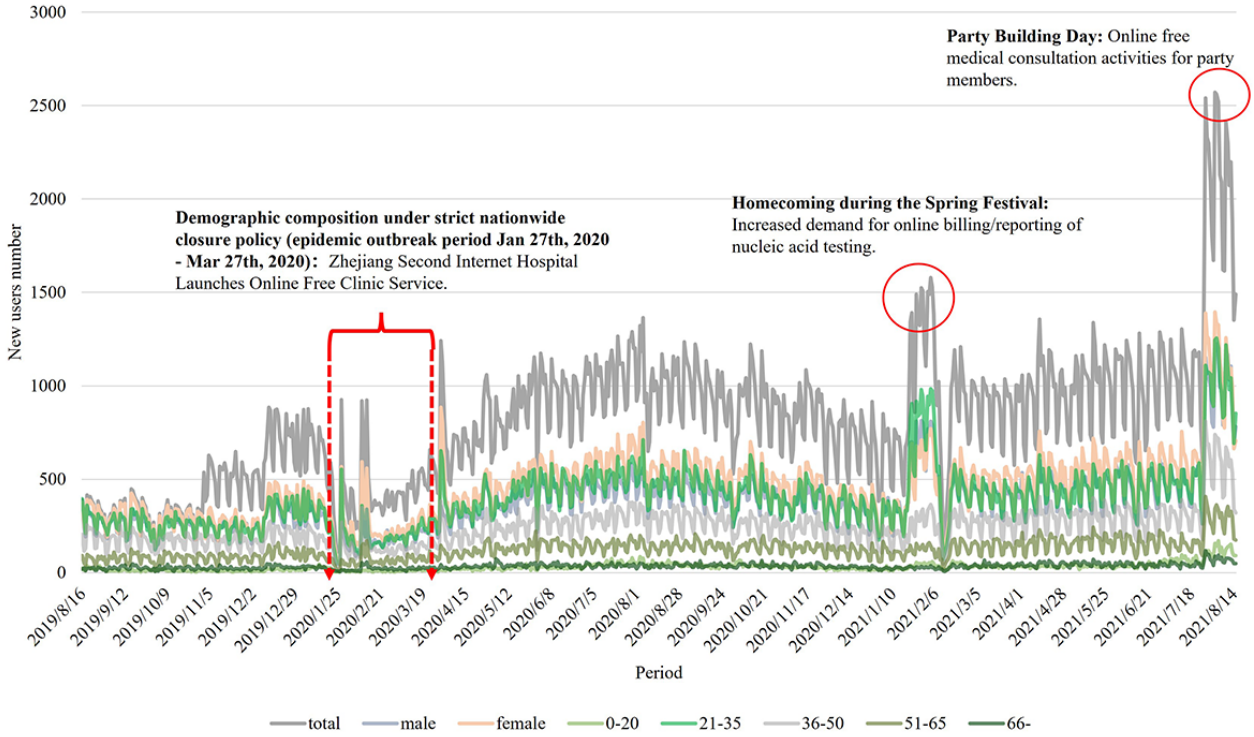
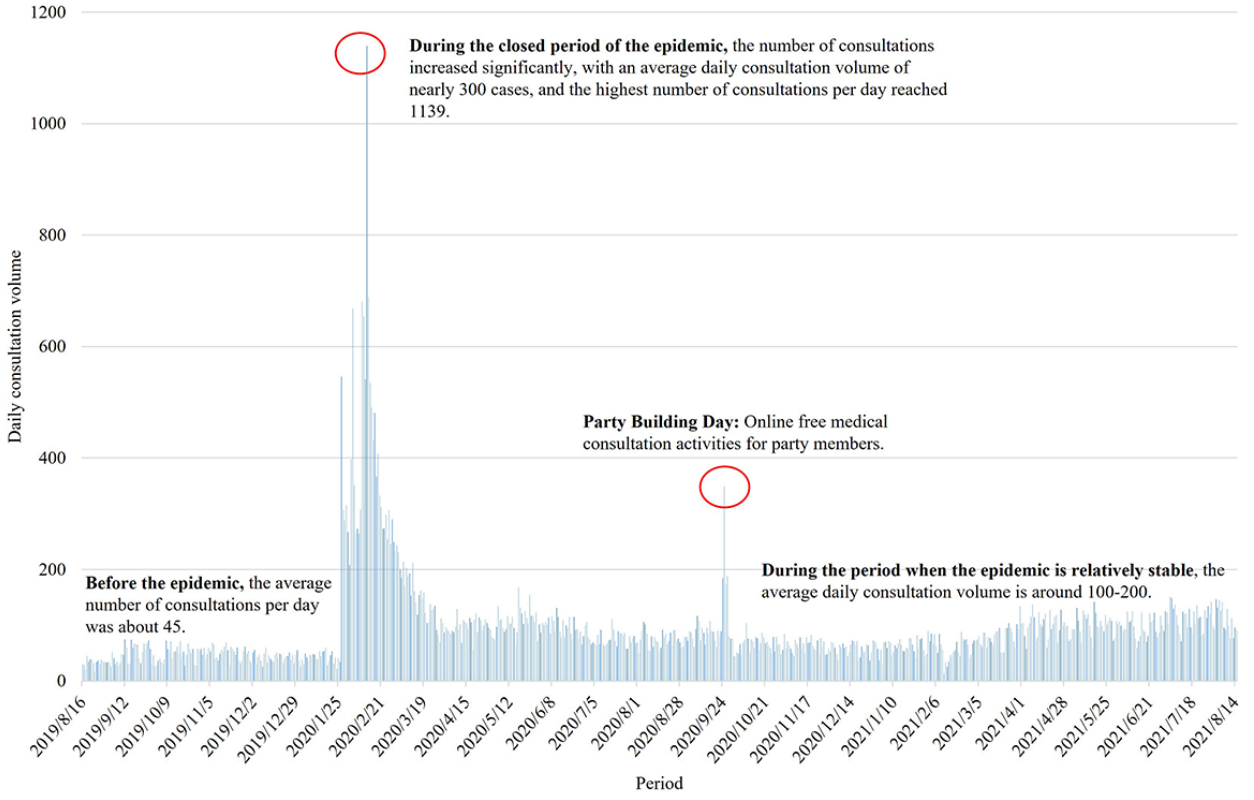


Figure 13. The number of new Second Affiliated Hospital of Zhejiang University School of Medicine (SAHZU) internet hospital consultations per day from August 16, 2019, to August 15, 2021.



ITS Analysis of New Registrations

We conducted a 2-stage ITS analysis targeting new consultation users from July 2019 to September 2021, using data from the SAHZU internet hospital online consultation platform. The first group of data (July 15, 2019, to September 15, 2020) was used to assess the temporary effect on the utilization rate of online medical treatment attributed to domestic COVID-19 (Figure 14). The sectional point of analysis in GROUP1 was determined by the length of the lockdown in Hangzhou during the COVID-19 outbreak. Given the globalization of the COVID-19 pandemic, GROUP2 (October 1, 2020, to September 13, 2021) was used to measure long-term trends in patients' use of online medical platforms (Figure 15).

The overall model ($P_{overall}$<.001) and individual coefficients (P_{x1}=.008, P_{x2}<.001, P_{x3}<.001, $P_{intercept}$<.001) were significant. The starting point of online medical service users was estimated at 38,162 (Table 2), with the number of active users increasing every day until January 23, 2020, according to the pre-intervention slope (n=0.052, 95% CI 0.014 to 0.091; *P*=.008). After the COVID-19 intervention was implemented (the strict quarantine policy), the number of users increased significantly (n=105.736, 95% CI 92.773 to 118.787; *P*<.001). After March 27, 2020, the development of online medical services decreased over time compared with the pre-intervention period, with a coefficient of –0.235 (95% CI –0.293 to –0.180; *P*<.001), thereby indicating a decline in telehealth users after the COVID-19 outbreak. Although the mean number of users increased, the overall development showed a downwards trend. To some extent, the pandemic had a temporary impact on the utilization rate of online medical services, although the effect on patients' habits remained to be seen.

The number of active users was initially 61,738 (Table 3). The trend in the utilization of online medical platforms was insignificant before the peak of the second wave of global COVID-19 outbreak (n=0.009, 95% CI –0.052 to 0.062; *P*=.81). COVID-19 deaths increased significantly (April 11, 2021), accounting for the change in the global situation (n=25.226, 95% CI 18.258 to 33.722; *P*<.001). The number of online users increased over time after April 9, 2020, with a coefficient of 0.167 (95% CI 0.087 to 0.247; *P*<.001). Along with Figure 15, the trend in the number of online users before and after the global pandemic peak was revealed. Before the COVID-19 pandemic, people's enthusiasm for online medical treatment had reached a plateau, and their behaviors undoubtedly began to change after the outbreak.

Figure 14. Segmented regression model for users of the online consultation platform from July 15, 2019, to September 13, 2020, using the generalized least square method; interrupted time series analysis (ITSA) to evaluate the impact of COVID-19 on telemedicine.

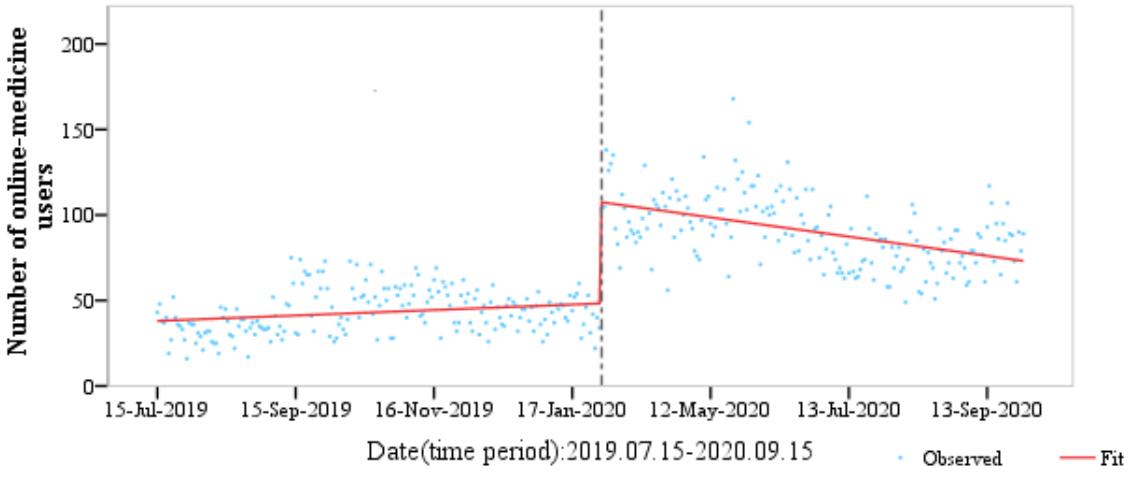


Figure 15. Segmented regression model for users of the online consultation platform from October 1, 2020, to August 7, 2021, using the generalized least squares method; prediction of secular trends of the prevalence of online consultations.

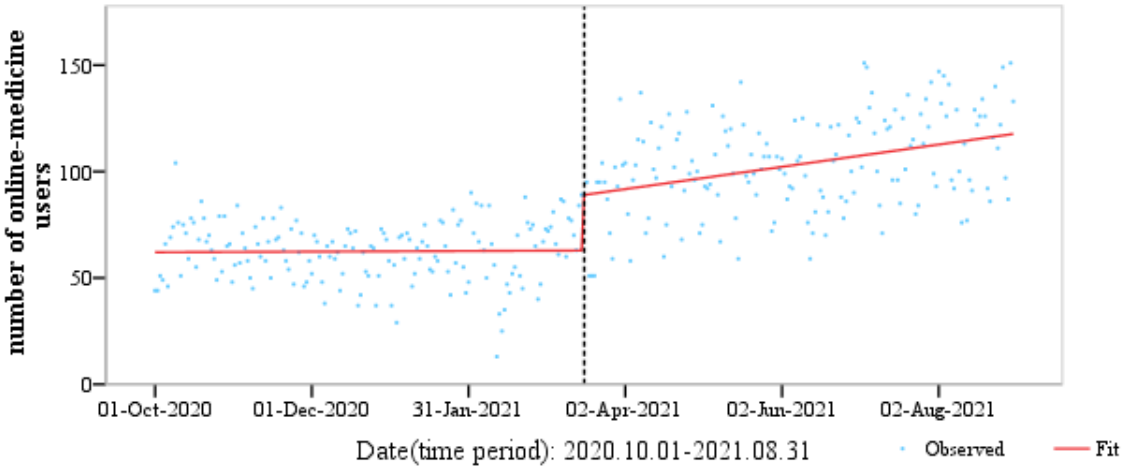


Table 2. Regression using a generalized least squares model on data from the development of the domestic pandemic from July 15, 2019, to September 15, 2020, with the pre-intervention period between July 15, 2019, and January 23, 2020; the strict quarantine period between January 23, 2020, and March 27, 2020; and the postintervention period between March 27, 2020, and September 15, 2020. Maximum lag: 1; number of observations=379; F3,379=150.3; *P*<.001.

Model	Estimation	SE	*t* value (df=379)	*P* value	95% CI
Y (intercept)	38.162	2.234	17.008	<.001	(33.611 to 42.398)
x1 (pre-intervention slope)	0.052	0.02	2.657	.008	(0.014 to 0.091)
x2 (change in intercept)	105.736	6.615	15.99	<.001	(92.773 to 118.787)
x3 (change in slope/interaction)	–0.235	0.029	–8.203	<.001	(–0.293 to 0.180)

Table 3. Regression to determine the long-term impact of COVID-19 on internet hospitals, using a generalized least squares model on data from the development of the domestic pandemic from October 1, 2020, to September 16, 2021, with the pre-intervention period between October 1, 2020, and April 10, 2021, and the postintervention period between April 11, 2021, and September 16, 2021. Maximum lag: 1; number of observations=331; F3,331=89.13; *P*<.001.

Model	Estimation	SE	*t* value (df=331)	*P* value	95% CI
Y (intercept)	61.738	2.796	22.194	<.001	(56.555 to 67.555)
x1 (pre-intervention slope)	0.009	0.029	0.167	.81	(–0.052 to 0.062)
x2 (change in intercept)	25.226	3.931	6.612	<.001	(18.258 to 33.722)
x3 (change in slope/Interaction)	0.167	0.041	4.104	.002	(0.087 to 0.247)

Changes in Medical Behaviors

Integrated with the follow-up surveys, we compared the changes in SAHZU internet hospital users' major demands, medical behaviors, and concerns, and we modelled the proportion and fluctuation of each demand through the 2D model (Table 4).

Table 4. Changes in medical behaviors in the 2D model.

Quadrant	Connotation	Patients' demands: March 2020 versus June 2021	Way of accessing medical care: November 2020 versus June 2021	Users' concerns: November 2020 versus June 2021
The first quadrant (universal type)	Specific weight, positive growth	1. Follow-up consultation 2. Consultation on common symptoms 3. Information service 4. Others	1. If you feel unwell, go directly to an offline hospital, and no longer use internet hospitals. 2. When the condition is relatively stable, the follow-ups are only conducted through the internet hospital. 3. Due to traffic or other factors, it is hoped that most diagnosis and treatment can be carried out through internet hospitals.	1. Doubts about the medical safety and quality of online diagnosis and treatment
The second quadrant (improved type)	Specific weight, negative growth	1. Drug refill	1. First, consult the internet hospital for advice; then, go offline for medical service according to the doctor's advice.	1. Poor timeliness of online text consultation and interaction; long wait times 2. Higher fee 3. Personal privacy and data security protection
The third quadrant (silent type)	Small proportion, negative growth	1. Rescheduling of treatment 2. Consultation on suspected symptoms of COVID-19	1. Others	1. Others
The fourth quadrant (potential type)	Small proportion, positive growth	1. Psychological support 2. Guidance on protective measures 3. Rescheduling of treatment	N/A[a]	N/A

[a]N/A: not applicable.

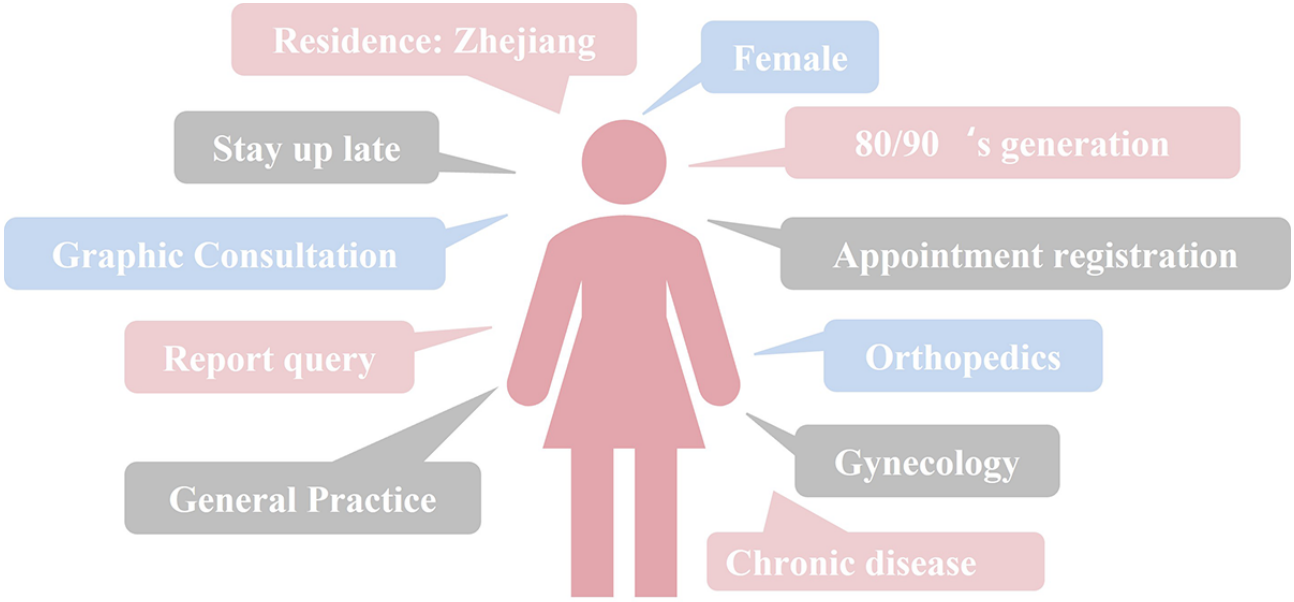
Discussion

This study conducted the first in-depth pandemic-related quantitative and qualitative analyses on changes in public behavior and the perceived factors influencing the use of internet hospitals. As a representative example of a public hospital in China, SAHZU has productively employed the practical experience of telehealth in the normalization of COVID-19 prevention and control.

Portraits of Internet Hospital Users in China

As a first-class public hospital and regional medical center in Zhejiang Province, the SAHZU internet hospital attracted multifarious users. The main profiles of users are shown in Figure 16.

Figure 16. Main profiles of Second Affiliated Hospital of Zhejiang University School of Medicine (SAHZU) internet hospital users.



There were slight changes in user profiles at different time points. The utilization rates of young and middle-aged women increased significantly during the pandemic isolation period. Traditional Chinese women care for the elderly and children and thus may have taken advantage of online access for their needs and those of family members.

Moreover, the popularity of the internet and improvement in digital literacy among women in China have also been reported [16,17]. Internet hospitals do not yet cover the entire population, which is reflected in the unbalanced distribution of patient profiles. Adults aged ≥66 years utilized telehealth services the least, demonstrating the greater barrier between the older population and online services. Performance risks, legal concerns, and privacy risks perceived by older adults may substantially decrease their intention to use telehealth applications [18-20].

Acceleration of Online Medical Demands During the COVID-19 Outbreak in China

During the COVID-19 outbreak, internet access served patients and counsellors without time and space restrictions. This essential medical support reduced panic, enhanced self-protective abilities, corrected improper medical-seeking behaviors, and facilitated epidemiological screening, thereby significantly improving COVID-19 prevention and control [21,22]. The number of consultations changed with the national quarantine policy and reached a peak on February 12, 2020. The ITS analysis verified that the extreme quarantine period caused an overall increase in the number of online consultations, showing a short-term leap upwards. It turned out to be a trend to employ the internet throughout entire treatment processes. Internet hospitals enhanced patients' sense of security, and after initial diagnosis and treatment, they could obtain online follow-ups, which further altered their medical behaviors [5].

We summarized the 8 major internet hospital user demands during the national shutdown period. In China, 60.0% of all internet hospitals provided telehealth services to address COVID-19. Internet hospitals mainly conducted consultations and psychological counselling, provided pandemic control information, filled prescriptions (86.6%), delivered drugs (74.5%), handled medical insurance claims (67.5%), and accepted or distributed donations [4]. Our results showed that follow-up consultations and drug refills were among the top requests, demonstrating efficiency and reducing unnecessary onsite visits. During the outbreak, online consultations virtually assisted in diagnosis and treatment, while surgical patients used the service for preoperative appointments and postoperative follow-up. Users with moderate health problems sought consultations more frequently than did individuals with severe conditions, indicating the value of online platforms for them, while in-person visits were essential and irreplaceable for patients with severe conditions. Additional online services, such as drug delivery, helped relieve the additional pressure on hospitals by reducing the influx of patients. In-depth efforts should be made to improve the management of prescription refills, quality and safety, dispensing acceptance, and standardization [23]. Online services offering real-time information regarding the hospital and national policies and

study results indicated that users were most concerned about symptoms, up-to-date knowledge, transmission routes, and preventive measures for COVID-19 [24].

Change in Behavior Patterns in Seeking Medical Resources

The results of this study agree with long-term data observations and surveys in different periods. During the pandemic, people's medical behavior patterns partially changed from onsite to online, which is irreversible for our public tertiary hospital. Before the pandemic, there was a temporary shock and reluctance to seek medical treatment online, but this did not produce a long-term change in the following waves of the global pandemic; people's health care habits began to change. According to the ITS analysis, this change can evolve into a long-term trend of choosing online health care and self-management. The trend in online visits followed that of overall telehealth visits, with the rates increasing dramatically after the start of the pandemic and then progressively decreasing, but users in need retained their online habits [25]. With this upwards trend, people's medical habits might have completely changed. We may carefully conclude that the response to COVID-19 will result in more than a temporary increase in online hospitals. For the predictable future, internet hospitals will reinforce the medical system by offering health care while minimizing potential exposure to the disease [26]. Once people have taken advantage of medical services via digital technologies, there is little reason to give them up.

This ongoing phenomenon should encourage health practitioners to move promptly towards digital transformation. However, how to make the ecological development of internet hospitals benign remains vague [27]. Based on the comparison of medical resource-seeking behaviors at the early and late stages of the pandemic, the changes in demands were sorted with the 2D model.

Follow-up consultation and information services were revealed as the basic needs of patients who revisited the online hospital and will be primary considerations in the future direction of internet hospital construction. Information services are capable of strengthening triage and convenience of services of internet hospitals before diagnosis.

The demand for prescriptions has dropped, indicating that some requirements cannot be resolved online, although this internet hospital function may undergo a functional iteration in the future. The limitations of internet hospitals are obvious. For instance, the demand for drug refills is significantly affected by policies (eg, the lack of health insurance or cash payments) [5]. This situation was a top-ranking medical concern regarding costs in our surveys.

Transient pandemic-related demands included consultation for common symptoms, suspected symptoms of COVID-19, psychological support, and guidance on protective measures. Online consultation serves as an indirect means of communication. Doctors consulted online are limited by the lack of information (eg, physical and auxiliary examinations) and may give only cursory medical advice, which cannot replace a hospital visit. Nondisease-specific issues and moderate health problems were much more frequent consultation requests rather than severe clinical conditions [28]. The proportion of pandemic screening was small, and although it was related to the crisis period, it may continue to rise. It is necessary to cooperate with offline approaches to more effectively draw upon the internet hospital's online advantages.

As revealed from the analysis of the path to select medical care, as an improved demand in 2D analysis, an increasing number of patients are willing to conduct research on the internet before going offline or following a doctor's online advice afterwards. It is noteworthy that the choice of "directly choosing offline medical treatment" fluctuates to a certain extent, thereby reflecting a temporary return to traditional habits with the decline of the pandemic.

Implications for Internet Hospital Development

Public online hospitals offer reliable resources and complete functions but have limited profit models. Long-term effective operation is an issue. Meanwhile, with novel vaccines and drugs targeting SARS-CoV-2 being developed, the challenge faced by the internet hospital community is to continuously update solutions for the majority of users. It will be crucial to consider the benefit-risk ratio for optimal therapies and minimize onsite visits. To sustain the online health care system, governments and societies are recognizing technology as a promising solution for innovative health service delivery and expansion with minimal investments.

Our surveys showed that the optimal demands of online users appear in prediagnosis and posttreatment. Human medical behaviors cannot be comprehensively shifted to digital access. Key online service implications include the entire closed loop of one's medical behaviors, as well as refined services before and after onsite visits. In addition, the construction of a portal for specialized diseases through online forms is an important direction for future internet hospital construction.

The pandemic has given health services an impetus for managing chronic conditions in innovative ways. Notably, the departments that received the most consultations (General Medicine, Cardiology, and Endocrinology) complied with the incidence of internal medicine [29]. Patients expect online consultations to provide professional advice and personalized care. To date, personalized telehealth solutions and clear implementation recommendations are being fully explored by internet hospitals [30]. Cancer patients, as another chronic condition population, are more susceptible to infection owing to the immunosuppressed state caused by anticancer therapy. In the spectrum of high-incidence tumor diseases in China, the most common types are lung cancer (approximately 17.9% of the total new incidence), colorectal cancer (12.2%), gastric cancer (10.5%), breast cancer (9.1%), and liver cancer (9.0%) [31]. These data are echoed in the top consultation rankings of thoracic, breast, and gastroenterology departments. The online hospital acts as a vital solution for various cancer patients, along with important support for many oncologists to help with decision-making [32]. Telehealth steps are recommended for postoperative patients and those on interventions for multiple adjuvant treatments [33].

While current user profiles indicate that these are the most popular departments for online consultation, diseases with strong stigmas, such as gynecology, infertility, dermatological problems, and mental diseases, are also benefitting from telehealth solutions. Patients experiencing stigma, which refers to the inner shame of patients suffering from certain diseases who are experiencing psychological stress [34], are taking advantage of online medical resources that allow them to seek the help they need with maximum privacy, by keeping these patients at a safe distance away from virtue circumstance. For instance, telehealth has been utilized as a useful communication method in the treatment of depression, anxiety, and posttraumatic stress disorder (PTSD) during the pandemic [24,35]. By the end of January 2020, consultation rates for psychiatric issues surged. Nevertheless, online psychological support peaked during the initial lockdown weeks. In view of isolation, misinformation and rumors spread via social media. Likewise, individuals worried about contracting this unknown virus and consultation for suspected COVID-19 accounted for 5.2% of all consultations, which is consistent with parallel studies [36]. During the early outbreak of COVID-19, internet hospitals assisted in relieving psychological burdens and increased disease awareness by providing official and responsible information. Furthermore, up-to-date health information was provided to relieve social anxiety.

Derivative Problems of Internet Hospitals

Perceived risk, defined as one's perception of uncertainty in the use of telehealth services and the severity of its consequences, is measured with 4 constructs: privacy risk [37], performance risk, legal concern, and trust [38]. Rectifying the concerns of users for online medical behaviors is also an important issue. All the derivative problems of online medical care in the survey were sorted out, as well as the corresponding reasons and possible solutions (see Table 5). Our 2D analysis highlighted the poor timeliness of online text consultation interactions, followed by high online fees and concerns about personal privacy and data security. The potential construction direction of internet hospitals refers to how to ensure the medical safety and quality of online diagnosis and treatment. During the COVID-19 pandemic, health care professionals, designers of telehealth applications, and policy makers devised more practical functions, user-friendly interfaces, and reasonable policy guidance for internet hospitals to upgrade the existing model and to deal with future crises.

Table 5. Derivation problems of internet hospitals.

Problems	Corresponding reasons	Possible solutions
Form	1. Poor timeliness 2. Insufficient doctor-patient interaction 3. Unfriendly experience	1. Enhance information interaction reminder. 2. Cultivate user service. 3. Optimize the rationality of the app interface. 4. Improve app functions, and enhance user experience.
Cost	1. The big price gap between internet hospitals 2. Unable to pay with medical insurance	1. Issue government policies to guide prices. 2. It is recommended that the medical insurance department include online diagnosis and treatment fees in the scope of medical insurance payment.
Ethics	1. Insufficient patient privacy protection technology 2. Concerns with patient privacy leakage 3. Lack of laws, regulations, and policy guidance	1. Issue policy documents to provide legal support. 2. Clarify the identification of medical malpractice and the division of responsibility. 3. Deidentify private data.
Platform	1. Incomplete and inadequate consideration of front-end, back-end, and bottom construction	1. Strengthen the technical team, and improve the data sharing ability and operability. 2. Consider data security needs.

Limitations and Improvements

The data were collected from a single institution in China. SAHZU is a public tertiary hospital and may not be representative of other levels of hospitals and different regions. However, Zhejiang University ranks the third highest among China's universities, and the hospital works with over 200 primary and secondary hospitals. Our online hospitals cover the nation, and almost 900,000 users had registered by January 2022. Therefore, we assume, to a certain extent, that our conclusions are representative. At the same time, the application prospects of internet hospitals in primary or secondary hospitals and multicenter studies are required to validate our conclusions. The surveys we utilized enabled users to fully express their experiences without the pressure of delivering socially acceptable opinions to an interviewer. However, the actual data acquired through face-to-face interviews were more authentic. Furthermore, the consultation database might be managed by artificial intelligence (AI) and build functions such as internet AI-assisted triage.

Conclusions

Since the outbreak of COVID-19 at the end of 2019, the pandemic has imposed great economic and social burdens worldwide. We conducted a retrospective cross-sectional study by analyzing online medical behaviors over 2 years. Our findings imply that, as a public tertiary internet hospital, the SAHZU internet hospital is partially and irreversibly integrated into the traditional medical system.

Acknowledgments

The authors would like to formally acknowledge the contributions of Dr. Yinjun Li for the intermittent time series (ITS) analysis. This work was partly supported by the Municipal Natural Science Foundation of Beijing of China (grant number 7222306) and the National Traditional Chinese Medicine innovation team and talent support projects (grant numbers ZYYCXTD-C-202210).

Data Availability

The data set supporting the conclusions of this article is included within the article. Additional data are available for review upon request.

Authors' Contributions

KD and FG planned and designed the study. HQ conducted the database search and screened studies for inclusion. YN, QL, and JL extracted data. YN, QL, and HQ were assessed in the investigation study. FG planned and performed the statistical analysis. HQ wrote the first draft of the manuscript. KD revised the manuscript. All authors contributed to manuscript revision and read and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire about internet hospital usage.
[DOCX File , 40 KB - medinform_v10i6e37042_app1.docx]

Multimedia Appendix 2
Interrupted time series (ITS) analysis.
[DOCX File , 38 KB - medinform_v10i6e37042_app2.docx]

Multimedia Appendix 3
Daily flow of the change in major demands for online consultation during the nationwide strict quarantine policy.
[DOCX File , 518 KB - medinform_v10i6e37042_app3.docx]

References

1. Lian W, Wen L, Zhou Q, Zhu W, Duan W, Xiao X, et al. Digital health technologies respond to the COVID-19 pandemic in a tertiary hospital in China: development and usability study. J Med Internet Res 2020 Nov 24;22(11):e24505 [FREE Full text] [doi: 10.2196/24505] [Medline: 33141679]
2. Lai Y, Chen S, Li M, Ung COL, Hu H. Policy interventions, development trends, and service innovations of internet hospitals in China: documentary analysis and qualitative interview study. J Med Internet Res 2021 Jul 20;23(7):e22330 [FREE Full text] [doi: 10.2196/22330] [Medline: 34283025]
3. Han Y, Lie RK, Guo R. The internet hospital as a telehealth model in China: systematic search and content analysis. J Med Internet Res 2020 Jul 29;22(7):e17995 [FREE Full text] [doi: 10.2196/17995] [Medline: 32723721]
4. Xu X, Cai Y, Wu S, Guo J, Yang L, Lan J, et al. Assessment of internet hospitals in China during the COVID-19 pandemic: national cross-sectional data analysis study. J Med Internet Res 2021 Jan 20;23(1):e21825 [FREE Full text] [doi: 10.2196/21825] [Medline: 33417586]
5. Liu L, Shi L. Chinese patients' intention to use different types of internet hospitals: cross-sectional study on virtual visits. J Med Internet Res 2021 Aug 13;23(8):e25978 [FREE Full text] [doi: 10.2196/25978] [Medline: 34397388]
6. Hong Z, Li N, Li D, Li J, Li B, Xiong W, et al. Telemedicine during the COVID-19 pandemic: experiences from Western China. J Med Internet Res 2020 May 08;22(5):e19577 [FREE Full text] [doi: 10.2196/19577] [Medline: 32349962]
7. Katayama Y, Kiyohara K, Hirose T, Matsuyama T, Ishida K, Nakao S, et al. A mobile app for self-triage for pediatric emergency patients in Japan: 4 year descriptive epidemiological study. JMIR Pediatr Parent 2021 Jun 30;4(2):e27581 [FREE Full text] [doi: 10.2196/27581] [Medline: 34255709]
8. Tencent Research Institute, Vcbeat Research. 2016 China Internet Hospital White Paper. 2016 Nov 15. URL: http://www.360doc.com/content/16/1127/01/34899478_609811863.shtml [accessed 2022-05-16]
9. 2021 China e-Hospital Development Report. National Telemedicine and Connected HealthCare Center. 2021 May 21. URL: https://zk.cn-healthcare.com/doc-show-53644.html [accessed 2022-05-16]
10. Develop rapidly! There are more than 1,600 Internet hospitals in my country. Xinhua News Agency. 2021 Aug 23. URL: http://www.gov.cn/xinwen/2021-08/23/content_5632844.htm [accessed 2022-05-16]

11. Wu H, Deng Z. Do Physicians' Online Activities Impact Outpatient Visits? An Examination of Online Health Communities Completed Research Paper. 2019 Presented at: 23rd Pacific Asia Conference on Information Systems; July 8-12, 2019; X'ian, China URL: https://dblp.org/rec/conf/pacis/WuD19
12. Lagu T, Norton CM, Russo LM, Priya A, Goff SL, Lindenauer PK. Reporting of patient experience data on health systems' websites and commercial physician-rating websites: mixed-methods analysis. J Med Internet Res 2019 Mar 27;21(3):e12007 [FREE Full text] [doi: 10.2196/12007] [Medline: 30916654]
13. Pike CW, Zillioux J, Rapp D. Online ratings of urologists: comprehensive analysis. J Med Internet Res 2019 Jul 02;21(7):e12436 [FREE Full text] [doi: 10.2196/12436] [Medline: 31267982]
14. Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of telehealth during the COVID-19 pandemic: scoping review. J Med Internet Res 2020 Dec 01;22(12):e24087 [FREE Full text] [doi: 10.2196/24087] [Medline: 33147166]
15. Specification for online consultation service for infectious disease epidemic situation. Zhejiang Digital Economy Association. 2020 Mar 15. URL: http://www.ttbz.org.cn/StandardManage/Detail/33876/ [accessed 2022-05-16]
16. Huang Q, Chen X, Huang S, Shao T, Liao Z, Lin S, et al. Substance and internet use during the COVID-19 pandemic in China. Transl Psychiatry 2021 Sep 23;11(1):491 [FREE Full text] [doi: 10.1038/s41398-021-01614-1] [Medline: 34556627]
17. Xiao Y, Liu X, Ren T. Internet use and pro-environmental behavior: Evidence from China. PLoS One 2022 Jan 27;17(1):e0262644 [FREE Full text] [doi: 10.1371/journal.pone.0262644] [Medline: 35085292]
18. Roberts ET, Mehrotra A. Assessment of disparities in digital access among Medicare beneficiaries and implications for telemedicine. JAMA Intern Med 2020 Oct 01;180(10):1386-1389 [FREE Full text] [doi: 10.1001/jamainternmed.2020.2666] [Medline: 32744601]
19. Lam K, Lu AD, Shi Y, Covinsky KE. Assessing telemedicine unreadiness among older adults in the United States during the COVID-19 pandemic. JAMA Intern Med 2020 Oct 01;180(10):1389-1391 [FREE Full text] [doi: 10.1001/jamainternmed.2020.2671] [Medline: 32744593]
20. Klaver NS, van de Klundert J, van den Broek RJGM, Askari M. Relationship between perceived risks of using mHealth applications and the intention to use them among older adults in the Netherlands: cross-sectional study. JMIR Mhealth Uhealth 2021 Aug 30;9(8):e26845 [FREE Full text] [doi: 10.2196/26845] [Medline: 34459745]
21. Gong K, Xu Z, Cai Z, Chen Y, Wang Z. Internet hospitals help prevent and control the epidemic of COVID-19 in China: multicenter user profiling study. J Med Internet Res 2020 Apr 14;22(4):e18908 [FREE Full text] [doi: 10.2196/18908] [Medline: 32250962]
22. Coffey JD, Christopherson LA, Glasgow AE, Pearson KK, Brown JK, Gathje SR, et al. Implementation of a multisite, interdisciplinary remote patient monitoring program for ambulatory management of patients with COVID-19. NPJ Digit Med 2021 Aug 13;4(1):123 [FREE Full text] [doi: 10.1038/s41746-021-00490-9] [Medline: 34389787]
23. Ding L, She Q, Chen F, Chen Z, Jiang M, Huang H, et al. The internet hospital plus drug delivery platform for health management during the COVID-19 pandemic: observational study. J Med Internet Res 2020 Aug 06;22(8):e19678 [FREE Full text] [doi: 10.2196/19678] [Medline: 32716892]
24. Li L, Liu G, Xu W, Zhang Y, He M. Effects of internet hospital consultations on psychological burdens and disease knowledge during the early outbreak of COVID-19 in China: cross-sectional survey study. J Med Internet Res 2020 Aug 04;22(8):e19551 [FREE Full text] [doi: 10.2196/19551] [Medline: 32687061]
25. 2020 Internet Hospital research report. Vcbeat Research. 2020 Aug 05. URL: https://vbdata.cn/reportDetail/f1c06232d56799502ba821edcd02830f [accessed 2022-05-16]
26. Wilson N, Mansoor OD, Boyd MJ, Kvalsvig A, Baker MG. We should not dismiss the possibility of eradicating COVID-19: comparisons with smallpox and polio. BMJ Glob Health 2021 Aug 09;6(8):e006810 [FREE Full text] [doi: 10.1136/bmjgh-2021-006810] [Medline: 34373261]
27. Li F. The status quo and development of Internet smart medical care in the context of the epidemic-Changchun City as an example (in Chinese). Zhongguo Xinxihua 2021 Jun 20;6:107-108. [doi: 10.3969/j.issn.1672-5158.2021.06.045]
28. Jiang X, Xie H, Tang R, Du Y, Li T, Gao J, et al. Characteristics of online health care services from China's largest online medical platform: cross-sectional survey study. J Med Internet Res 2021 Apr 15;23(4):e25817 [FREE Full text] [doi: 10.2196/25817] [Medline: 33729985]
29. The Healthy China Action Plan (2019-2030). Healthy China Action Promotion Committee. 2019 Jul 15. URL: http://www.gov.cn/xinwen/2019-07/15/content_5409694.htm [accessed 2022-05-16]
30. Knitza J, Simon D, Lambrecht A, Raab C, Tascilar K, Hagen M, et al. Mobile health usage, preferences, barriers, and eHealth literacy in rheumatology: patient survey study. JMIR Mhealth Uhealth 2020 Aug 12;8(8):e19661 [FREE Full text] [doi: 10.2196/19661] [Medline: 32678796]
31. Liu Z, Li Z, Zhang Y. Interpretation on the report of Global Cancer Statistics 2020. Journal of Multidisciplinary Cancer Management 2021;7(02):1-14 [FREE Full text] [doi: 10.12151/JMCM.2021.02-01]
32. Cortiula F, Pettke A, Bartoletti M, Puglisi F, Helleday T. Managing COVID-19 in the oncology clinic and avoiding the distraction effect. Ann Oncol 2020 May;31(5):553-555 [FREE Full text] [doi: 10.1016/j.annonc.2020.03.286] [Medline: 32201224]
33. Belkacemi Y, Grellier N, Ghith S, Debbi K, Coraggio G, Bounedjar A, et al. A review of the international early recommendations for departments organization and cancer management priorities during the global COVID-19 pandemic:

applicability in low- and middle-income countries. Eur J Cancer 2020 Aug;135:130-146 [FREE Full text] [doi: 10.1016/j.ejca.2020.05.015] [Medline: 32580130]
34. Geng F, Dong Y, Michael K. Reliability and validity of the Chinese-version of Stigma Scale for Mental Illness. Chinese Mental Health Journal 2020;24(05):343-346. [doi: 10.3969/j.issn.1000-6729.2010.05.007]
35. Tan Y, Teng Z, Qiu Y, Tang H, Xiang H, Chen J. Potential of mobile technology to relieve the urgent mental health needs in China: web-based survey. JMIR Mhealth Uhealth 2020 Jul 07;8(7):e16215 [FREE Full text] [doi: 10.2196/16215] [Medline: 32673239]
36. Juanjuan L, Santa-Maria CA, Hongfang F, Lingcheng W, Pengcheng Z, Yuanbing X, et al. Patient-reported outcomes of patients with breast cancer during the COVID-19 outbreak in the epicenter of China: a cross-sectional survey study. Clin Breast Cancer 2020 Oct;20(5):e651-e662 [FREE Full text] [doi: 10.1016/j.clbc.2020.06.003] [Medline: 32709505]
37. Sadilek A, Liu L, Nguyen D, Kamruzzaman M, Serghiou S, Rader B, et al. Privacy-first health research with federated learning. NPJ Digit Med 2021 Sep 07;4(1):132 [FREE Full text] [doi: 10.1038/s41746-021-00489-2] [Medline: 34493770]
38. Deng Z, Hong Z, Ren C, Zhang W, Xiang F. What predicts patients' adoption intention toward mHealth services in China: empirical study. JMIR Mhealth Uhealth 2018 Aug 29;6(8):e172 [FREE Full text] [doi: 10.2196/mhealth.9316] [Medline: 30158101]

Abbreviations

AI: artificial intelligence
GLSM: generalized least square method
ITS: interrupted time series
PTSD: posttraumatic stress disorder
SAHZU: The Second Affiliated Hospital of Zhejiang University School of Medicine

Edited by C Lovis; submitted 05.02.22; peer-reviewed by H Wei, H Chen, S Hajesmaeel Gohari; comments to author 28.02.22; revised version received 21.03.22; accepted 28.04.22; published 01.06.22.

Please cite as:
Ge F, Qian H, Lei J, Ni Y, Li Q, Wang S, Ding K
Experiences and Challenges of Emerging Online Health Services Combating COVID-19 in China: Retrospective, Cross-Sectional Study of Internet Hospitals
JMIR Med Inform 2022;10(6):e37042
URL: https://medinform.jmir.org/2022/6/e37042
doi:10.2196/37042
PMID:35500013

Review

Combating COVID-19 Using Generative Adversarial Networks and Artificial Intelligence for Medical Images: Scoping Review

Hazrat Ali[1], PhD; Zubair Shah[1], PhD

College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

Corresponding Author:
Zubair Shah, PhD
College of Science and Engineering
Hamad Bin Khalifa University
Al Luqta St
Ar-Rayyan
Doha, 34110
Qatar
Phone: 974 50744851
Email: zshah@hbku.edu.qa

Abstract

Background: Research on the diagnosis of COVID-19 using lung images is limited by the scarcity of imaging data. Generative adversarial networks (GANs) are popular for synthesis and data augmentation. GANs have been explored for data augmentation to enhance the performance of artificial intelligence (AI) methods for the diagnosis of COVID-19 within lung computed tomography (CT) and X-ray images. However, the role of GANs in overcoming data scarcity for COVID-19 is not well understood.

Objective: This review presents a comprehensive study on the role of GANs in addressing the challenges related to COVID-19 data scarcity and diagnosis. It is the first review that summarizes different GAN methods and lung imaging data sets for COVID-19. It attempts to answer the questions related to applications of GANs, popular GAN architectures, frequently used image modalities, and the availability of source code.

Methods: A search was conducted on 5 databases, namely PubMed, IEEEXplore, Association for Computing Machinery (ACM) Digital Library, Scopus, and Google Scholar. The search was conducted from October 11-13, 2021. The search was conducted using intervention keywords, such as “generative adversarial networks” and “GANs,” and application keywords, such as “COVID-19” and “coronavirus.” The review was performed following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) guidelines for systematic and scoping reviews. Only those studies were included that reported GAN-based methods for analyzing chest X-ray images, chest CT images, and chest ultrasound images. Any studies that used deep learning methods but did not use GANs were excluded. No restrictions were imposed on the country of publication, study design, or outcomes. Only those studies that were in English and were published from 2020 to 2022 were included. No studies before 2020 were included.

Results: This review included 57 full-text studies that reported the use of GANs for different applications in COVID-19 lung imaging data. Most of the studies (n=42, 74%) used GANs for data augmentation to enhance the performance of AI techniques for COVID-19 diagnosis. Other popular applications of GANs were segmentation of lungs and superresolution of lung images. The cycleGAN and the conditional GAN were the most commonly used architectures, used in 9 studies each. In addition, 29 (51%) studies used chest X-ray images, while 21 (37%) studies used CT images for the training of GANs. For the majority of the studies (n=47, 82%), the experiments were conducted and results were reported using publicly available data. A secondary evaluation of the results by radiologists/clinicians was reported by only 2 (4%) studies.

Conclusions: Studies have shown that GANs have great potential to address the data scarcity challenge for lung images in COVID-19. Data synthesized with GANs have been helpful to improve the training of the convolutional neural network (CNN) models trained for the diagnosis of COVID-19. In addition, GANs have also contributed to enhancing the CNNs’ performance through the superresolution of the images and segmentation. This review also identified key limitations of the potential transformation of GAN-based methods in clinical applications.

(JMIR Med Inform 2022;10(6):e37365) doi:10.2196/37365

KEYWORDS

augmentation; artificial intelligence; COVID-19; diagnosis; generative adversarial networks; diagnostic; lung image; imaging; data augmentation; X-ray; CT scan; data scarcity; image data; neural network; clinical informatics

Introduction

Background

In December 2019, COVID-19 broke out and spread at an unprecedented rate, given the highly contagious nature of the virus. As a result, the World Health Organization (WHO) declared it a global pandemic in March 2020 [1]. Therefore, a response to combat the spread through speedy diagnosis became the most critical need of the time. A common method for diagnosing COVID-19 is the use of a real-time reverse transcription–polymerase chain reaction (RT-PCR) test. However, with the increasing number of cases worldwide, the health care sector was overloaded as it became challenging to cope with the requirements of the tests with the available testing facilities. In addition, research has shown that RT-PCR may result in false negatives or fluctuating results [2]. Hence, diagnosis through computed tomography (CT) and X-ray images of lungs may supplement performance. Motivated by this need, alternative methods, such as automatic diagnosis of COVID-19 from lung images, were explored and encouraged. In this regard, it is well understood that artificial intelligence (AI) techniques could help inspect chest CTs and X-rays within seconds and augment the public health care sector. The use of properly trained AI models for diagnosis of COVID-19 is promising for scaling up the capacity and accelerating the process as computers are, in general, faster than humans in computations.

Many AI and medical imaging methods were explored to provide support in the early diagnosis of COVID-19, for example, AI for COVID-19 [3-5], machine learning for COVID-19 [6], and data science for COVID-19 [7]. However, AI techniques rely on large data. For example, training a convolutional neural network (CNN) to perform classification of COVID-19 versus normal chest X-ray images requires training of the CNN with a large number of chest X-ray images both for COVID-19 and for normal cases. Since the diagnosis of COVID-19 requires studying of lung CT or X-ray images, the availability of lung imaging data is vital to develop medical imaging methods. However, the lack of data on COVID-19 hampered the initial progress in developing these methods to combat COVID-19.

Many early attempts were made to collect imaging data for lungs infected with COVID-19—specifically CT and X-ray images either through a private collection in hospitals or through crowdsourcing using public platforms. In parallel, many studies have explored the use of generative adversarial networks (GANs) to generate synthetic imaging data that can improve the training of AI models to diagnose COVID-19.

GANs are a family of deep learning models that consist of 2 neural networks trained in an adversarial fashion [8-15]. The 2 neural networks, namely the generator and the discriminator, attempt to minimize their losses, while maximizing the loss of the other. This training mechanism improves the overall learning task of the GAN model, particularly for generating data. GANs have recently been studied for computer vision and medical imaging tasks, such as image generation, superresolution, and segmentation [9,10]. Given the significant potential of GANs in medical imaging, it was intuitive that many researchers were tempted to explore the use of GANs for data augmentation of imaging data on COVID-19. In addition, some researchers also used GANs for segmentation and superresolution of lung images.

This scoping review focuses on providing a comprehensive review of the GAN-based methods used to combat COVID-19. Specifically, it covers the studies where GANs have been used for lung CT and X-ray images to diagnose COVID-19 or to enhance the performance of CNNs for the diagnosis of COVID-19 (eg, by data augmentation or superresolution).

Research Problem

GANs have gained the attention of the medical imaging research community. As the COVID-19 pandemic continued to grow in 2020 and 2021, the research community faced a significant challenge due to the scarcity of medical imaging data on COVID-19 that can be used to train AI models (eg, CNN) to perform COVID-19 diagnosis automatically. Given the popularity of GANs for image synthesis, researchers turned to exploring the use of GANs for data augmentation of lung radiology images. Many studies were conducted to use different variants of GANs for data augmentation of lung CT images and lung X-ray images. Similarly, a few studies also used GANs for the diagnosis of COVID-19 from lung radiology images. However, to the best of our knowledge, there is no review on the role of GANs in addressing the challenges related to COVID-19 data scarcity and diagnosis. The following research questions related to COVID-19 imaging data were considered for this review:

What were the common applications of GANs proposed for challenges related to COVID-19?

- Which architectures of GANs are most commonly applied for data augmentation tasks related to COVID-19?
- Which imaging modality is the popular choice for the diagnosis of COVID-19?
- What were the most commonly used data sets of CT and X-ray images for COVID-19?
- What studies were conducted with open-source code to reproduce the results?
- What studies were conducted and presented to radiology experts for evaluation of the suitability toward future use in clinical applications?

The results of this review will be helpful for researchers and professionals in the medical imaging and health care domain who are considering using GAN-based methods to address challenges related to COVID-19 imaging data and to address the challenge in improving automatic diagnosis using radiology images.

Methods

Study Design

In this work, a scoping review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) guidelines [16]. The methods for performing the study are described next.

Search Strategy

Search Sources

A search was conducted from October 11-13, 2021. The search was performed on the following 5 databases: PubMed, IEEEXplore, Association for Computing Machinery (ACM) Digital Library, Scopus, and Google Scholar. In the case of Google Scholar, only the first 99 results were retained as the results beyond 99 items were highly irrelevant to the scope of the study. Similarly, in the case of ACM Digital Library, the first 100 results were retained as a lack of relevancy to the study was obvious in results beyond 100.

Search Terms

The search terms used in this study were chosen from the literature with guidance from experts in the field. The terms were chosen based on the intervention (eg, "generative adversarial networks," "GANs," "cycleGANs") and the target application (eg, "COVID-"19", "coronavirus," "corona pandemic"). The exact search strings used in the search for this study are available in Multimedia Appendix 1.

Search Eligibility Criteria

This study focused on the applications of GANs in analyzing radiology images of lungs for COVID-19, used for any purpose such as data augmentation or synthesis, diagnosis, superresolution, and prognosis. Only those studies were included that reported GAN-based methods for analyzing chest X-ray images, chest CT images, and chest ultrasound images. Studies that reported GAN-based methods for analyzing nonlung images were removed. Any studies that used deep learning methods but did not use GANs were also excluded. Studies reporting GANs for nonimaging data were also excluded. To provide a list of reliable studies, only peer-reviewed articles, conference papers, and book chapters were included. Preprints, conference abstracts, short letters, and commentaries were excluded. Similarly, review articles were also excluded. No restrictions were imposed on the country of publication, study design, or outcomes. Studies that were written in English and were published from 2020 to 2022 were included. No studies before 2020 were included.

Study Selection

Two reviewers (authors HA and ZS) screened the titles and abstracts of the search results. Initial screening by the 2 reviewers was performed independently. Disagreement occurred for only 9 articles. The disagreement was resolved through mutual discussion and consensus. For measuring the disagreement, Cohen κ [17] was calculated to be 0.89, which shows good agreement between the 2 independent reviewers. Multimedia Appendix 2 shows the matrix for the agreement between the 2 independent reviewers.

Data Extraction

Multimedia Appendix 3 shows the form for extraction of the key characteristics. The form was pilot-tested and refined in 2 rounds, first by data extraction for 5 studies and then by data extraction for another 5 studies. This refinement of the form ensured that only relevant data were extracted from the studies. The 2 reviewers (HA and ZS) extracted the data from the included studies, related to the GAN-based method, applications, and data sets. Any disagreement between the reviewers was resolved through mutual consensus and discussions. As the disagreements at the study selection stage were resolved through careful and lengthy discussions, the disagreement at the data extraction was only minor.

Data Synthesis

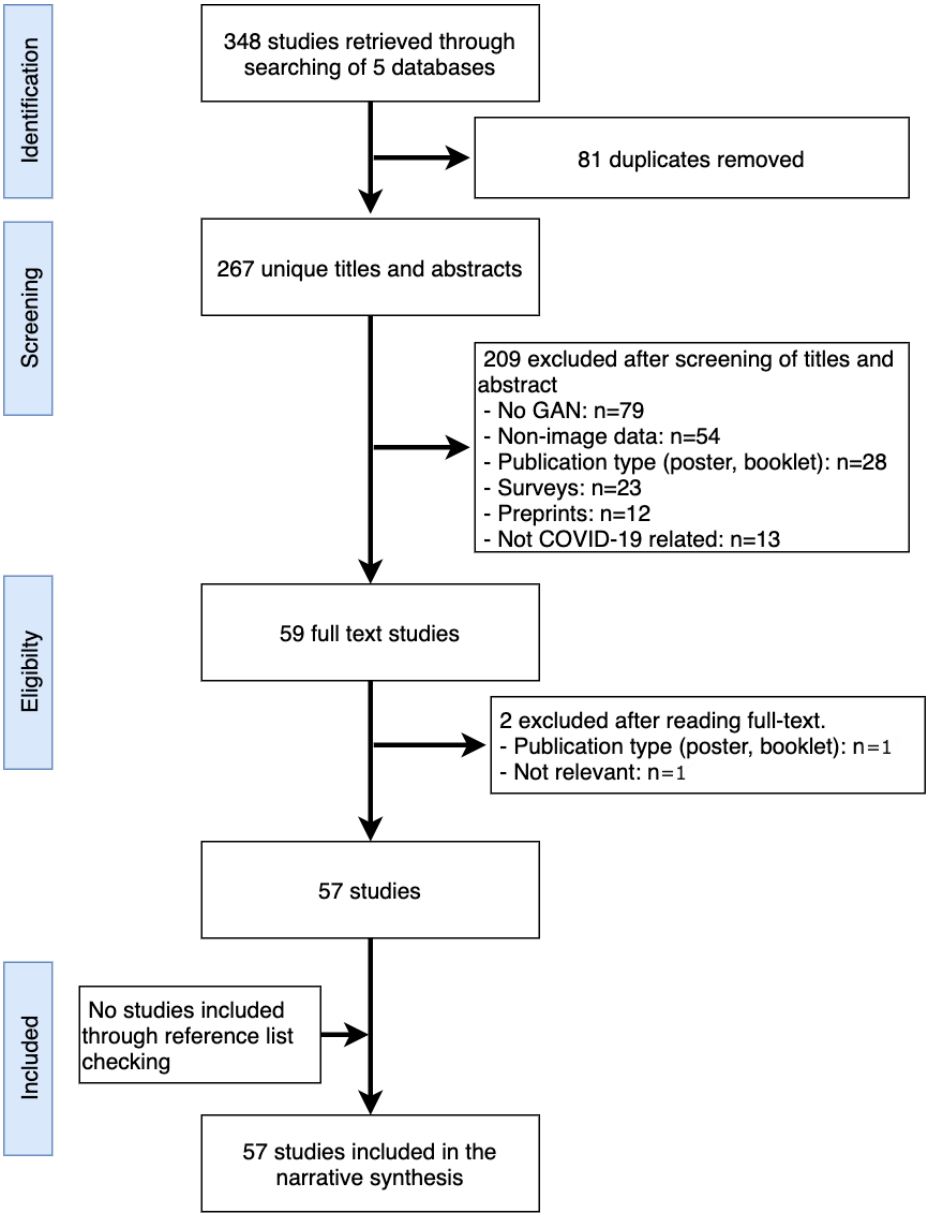
After extraction of the data from the full text of the identified studies, a narrative approach was used to synthesize the data. The use of GAN-based methods was classified in terms of the application of GANs (eg, augmentation, segmentation of lungs); the type of GAN architecture, if reported (eg, conditional GAN or cycleGAN); and the modality of the imaging data for which the GAN was used (eg, CT or X-ray imaging). Similarly, the studies were classified based on the availability of the data set (eg, public or private), the size of the data set (eg, the number of images in the original images and the number of images after augmentation with the GAN, if applicable), and the proportion of the training and test sets as well as the type of cross-validation. The data synthesis was managed and performed using Microsoft Excel.

Results

Search Results

From 5 online databases, a total of 348 studies were retrieved (see Figure 1). Of the 348 studies, 81 (23.3%) duplicates were removed. The titles and abstracts of the remaining 267 (76.7%) studies were carefully screened as per the criteria of inclusion and exclusion. The screening of the titles and abstracts resulted in the exclusion of 208 (77.9%) studies (see Figure 1 for reasons of exclusion). After the full-text reading of the remaining 59 (22.1%) studies, 2 (3%) studies were excluded following the inclusion/exclusion criteria. Finally, a total of 57 (97%) studies were included in this review. No additional studies were found through reference list checking. As per the yearwise publication, 15 (26%) of 57 studies were published in 2020 and 41 (72%) of 57 were published in 2021.

Figure 1. PRISMA-ScR flowchart for the search outcomes and selection of studies. GAN: generative adversarial network; PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews.



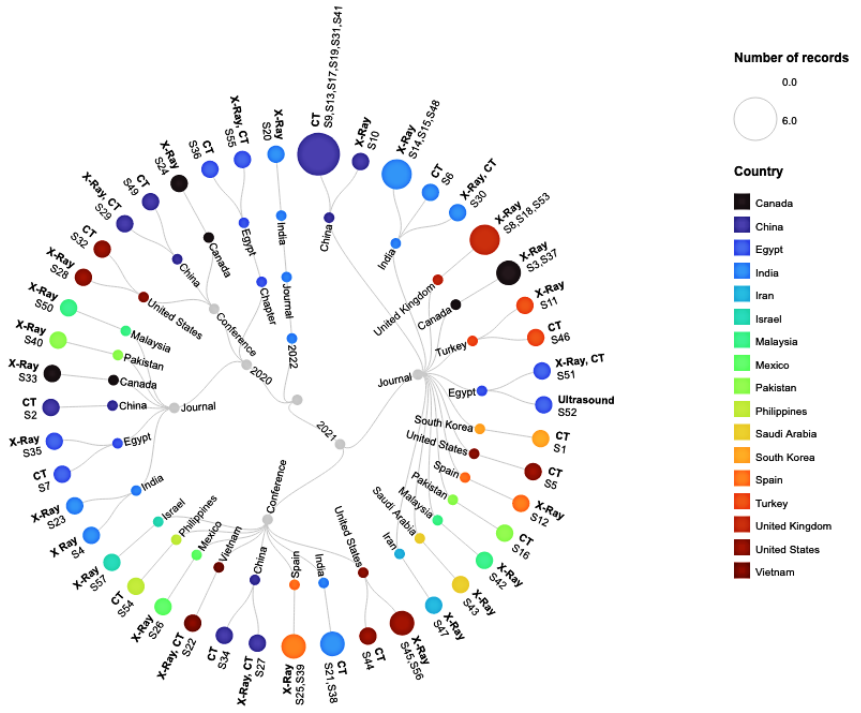
Demographics of the Included Studies

Among the included studies (N=57), 37 (65%) studies were published articles in peer-reviewed journals, 18 (32%) studies were published in conference proceedings, and 2 (4%) studies were published as book chapters. No thesis publication was found relevant to the scope of this review. Around one-fourth of the studies (n=15, 26%) were published in 2020. Most of the studies were published in 2021 (n=41, 72%). The included studies were published in 14 countries. The largest number of publications were from China (n=12, 21%), followed by India (n=10, 18%). Both the United States and Egypt published the same number of studies (n=6, 11%, each). The characteristics are summarized in Table 1 and Multimedia Appendix 4. Figure 2 (see [18-74]) shows the demographics of the included studies, along with the modality of the chest images used.

Table 1. Characteristics of the included studies (N=57). Demographics are shown for type of publication, country of publication, and year of publication.

Characteristics	Studies, n (%)
Publication type	
Journal	37 (65)
Conference	18 (32)
Book chapter	2 (4)
Country	
China	12 (21)
India	10 (18)
United States	6 (11)
Egypt	6 (11)
Canada	4 (7)
Spain	3 (5)
Malaysia	2 (4)
Turkey	2 (4)
Pakistan	2 (4)
Vietnam	1 (2)
Mexico	1 (2)
South Korea	1 (2)
Philippines	1 (2)
Israel	1 (2)
Year of publication	
2020	15 (26)
2021	41 (72)
2022	1 (2)

Figure 2. Characteristics of the included studies showing the publication type, country of publication, and modality of data. The number of studies is reflected by the size of the terminal node. The numbers S1-S57 refer to the included studies. CT: computed tomography.



Application of the Studies

As shown in Table 2, the included studies have reported 5 different tasks being addressed: augmentation (data augmentation), diagnosis of COVID-19, prognosis, segmentation (to identify the lung region), and diagnosis of lung diseases. As the diagnosis of COVID-19 using medical imaging has been a priority since the pandemic started, 39 (68%) of 57 studies reported the diagnosis of COVID-19 as the main focus of their work [19-21, 23-33, 35-37, 39, 41, 42, 44, 46, 50, 52, 53, 55, 56, 58-60, 63-69, 71, 72]. In addition, 9 (16%) studies reported data augmentation as the main task addressed in the work [18,43,45,49,54,61,62], 1 (2%) study reported prognosis of COVID-19 [22], 3 (5%) studies reported segmentation of lungs [34,51,57], and 1 (2%) study reported diagnosis of multiple lung diseases [47].

Table 2. Applications of using GAN[a]-based methods and types of GANs.

Applications	Studies (N=57), n (%)
Applications addressed in the studies	
Diagnosis	39 (68)
Data augmentation	9 (16)
Segmentation+diagnosis	3 (5)
Segmentation	3 (5)
Diagnosis of lung disease	1 (2)
Prognosis	1 (2)
Prognosis+diagnosis	1 (2)
Applications of using GANs	
Augmentation	42 (74)
Diagnosis	5 (9)
Superresolution	3 (5)
Segmentation	3 (5)
Feature extraction	2 (4)
Prognosis	1 (2)
3D synthesis	1 (2)
Type of GAN used	
GAN	17 (30)
CycleGAN	9 (16)
Conditional GAN	9 (16)
Deep convolutional GAN	4 (7)
Auxiliary classifier GAN	4 (7)
Superresolution GAN	2 (4)
3D conditional GAN	2 (4)
BiGAN	1 (2)
Random GAN	1 (2)
Pix2pix GAN	1 (2)

[a]GAN: generative adversarial network.

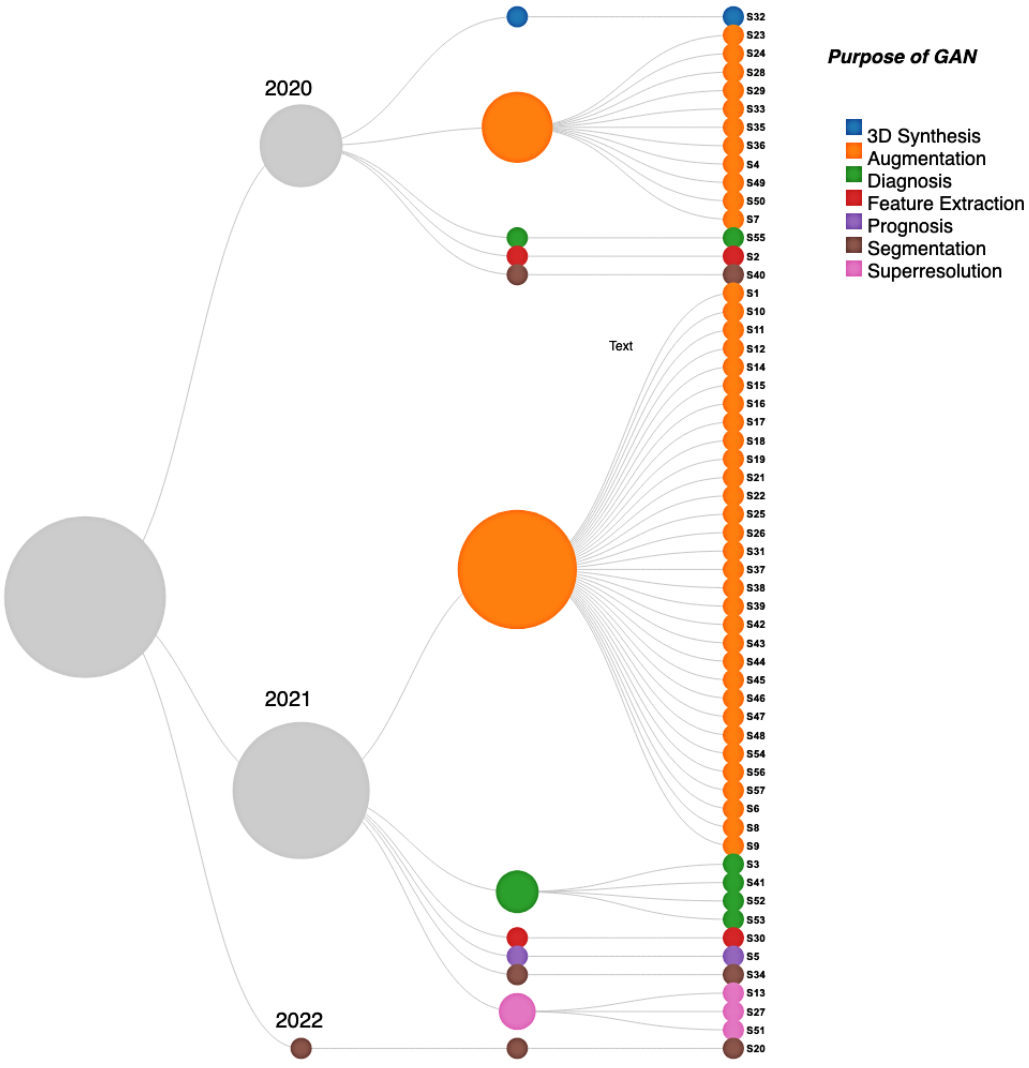
The majority of the studies used GANs to augment the data, where they reported the use of GANs to increase the data set size. Specifically, 42 (74%) studies used GAN-based methods for data augmentation [18, 21, 23-29, 31-36, 38-43, 45, 46, 48, 50, 52-56, 59-67, 71, 73, 74]. The augmented data were then used to improve the training of different CNNs to diagnose COVID-19. In addition, 3 (5%) studies used GANs for segmentation of the lung region within the chest radiology images [37,51,57], 3 (5%) studies used GANs for superresolution to improve the quality of the images before using them for diagnosis purposes [30,44,68], 5 (9%) studies used GANs for the diagnosis of COVID-19 [20,58,69,70,72], 2 (4%) studies used GANs for feature extraction from images [19,47], and 1 (2%) study used a GAN-based method for prognosis of COVID-19 [22]. The prevalent mode of imaging is the use of 2D imaging data, and 1 (2%) study reported a GAN-based method for synthesizing 3D data [49]. Figure 3

(see [18-74]) shows the mapping of the applications of GAN-based methods for all the included studies.

Different variants have been proposed for GAN architectures since their inception. The most common type of GAN used in these studies was the cycleGAN, used in 9 (16%) studies [29,35,36,42,46,54,56,70,74]. The cycleGAN is an image translation GAN that does not require paired data to transform images from one domain to another. Other popular types of GANs were conditional GAN used by 9 (16%) studies [18,22,24,25,33,37,41,57,60], deep convolutional GAN used by 4 (7%) studies [21,38,43,67], and auxiliary classifier GAN used by 4 (7%) studies [32,40,55,69]. The superresolution GAN was used by 2 (4%) studies [44,68], and 1 (2%) study reported the use of multiple GANs, namely Wassertein GAN, auxiliary classifier GAN, and deep convolutional GAN, and compared their performances for improving the quality of images [31].

Of the 57 studies, only 10 (18%) [18,19,26,27,30,34,43,61-73] reported changes to the architecture of the GAN they were using. In the rest of the studies, no major changes were reported to the architecture of the GAN.

Figure 3. Major applications of GANs in the included studies. The number of publications for each application is reflected by the size of the circle in the second-last layer. The numbers S1-S57 refer to the included studies. GAN: generative adversarial network.



Characteristics of the Data Sets

The included studies applied GANs on lung radiology images obtained using various modalities. Specifically, the use of X-ray images dominated the studies. In total, 29 (51%) studies used X-ray images of lungs [20,21, 25, 27-29, 31, 32, 35, 37, 40-43, 45, 50, 52, 54, 56, 57, 59, 60, 62, 64, 65, 67, 70, 73, 74], while 21 (37%) studies used CT images [18,19,22-24,26,30,33,34,36,38,48,49,51,53,55,58,61,63,66,71], and 6 (11%) studies reported the use of both X-ray and CT images [39,44,46,47,68,72]. Only 1 (2%) study used ultrasound images for COVID-19 diagnosis [69], which shows that ultrasound is not a popular imaging modality for training GANs and other deep learning models for COVID-19 detection (also see Figure 4). Of the 57 studies, most (n=47, 82%) used image data sets that are publicly available. In 10 (18%) studies, the data sets used are private. Table 3 provides a list of the various data sets used in the included studies and whether they are publicly available data sets or private. The most commonly used data set was the COVIDx data set available on Github, used by 26 (46%) studies.

Figure 4. Venn diagram showing the number of studies using CT vs X-ray images. Only 1 (2%) study reported the use of ultrasound images (not reflected here). CT: computed tomography.

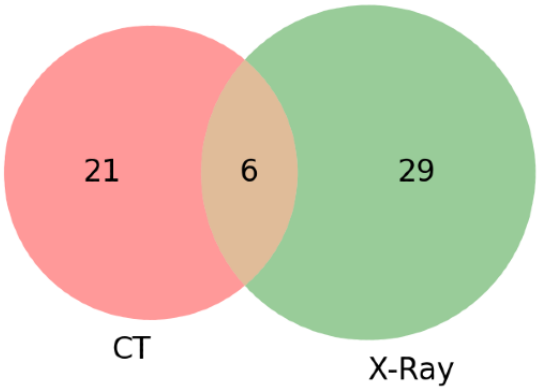


Table 3. Resources of the data sets used in the included studies. The name is provided only if available.

Platform (name)	Public or private	Modality of imaging
Kaggle	Public [75]	CT[a]
Github	Public [76]	CT
Github	Public [77]	CT
Github (Covidx)	Public [78]	X-ray, CT
Github	Public [79]	X-ray
Kaggle (Tawsif)	Public [80]	X-ray
Github	Public [81]	X-ray
Kaggle	Public [82]	X-ray
Mendeley	Public [83]	CT
Website	Public [84]	CT
Kaggle (Allen Institute)	Public [85]	CT
Kaggle (RSNA)	Public [86]	X-ray
Website	Public [87]	CT
Github	Public [88]	Ultrasound
Kaggle	Public [89]	X-ray
Website (Italian Society of Medical and Interventional Radiology)	Public [90]	X-ray
First Affiliated Hospital of the University of Science and Technology China	Private	CT
Massachusetts General Hospital, Brigham and Women's Hospital	Private	CT
Comlejo Hospitalario Universitario de A Coruna Spain	Private	X-ray

[a]CT: computed tomography.

The majority of the studies reported the size of the data set in terms of the number of images. The number of images used was greater than 10,000 in only 7 (12%) studies [20,22,30,39,63,66,74], while 3 (5%) studies used images between 5000 and 10,000 [33,47,64]. The most common range for the number of images used was 1000-5000 images used in 15 (26%) studies. Around one-fifth of the studies (n=11, 19%) used between 500 and 1000 images. In 11 (19%) other studies, the number of images used was less than 500. No study reported a number of images less than 100. The maximum number of images was 84,971, used by Uemura et al [22]. Only a few of the studies reported the number of patients for whom the data were used: 1 (2%) study used data for more than 1000 patients [26], 2 (4%) studies used data for 500-1000 patients [29,42], 6 (11%) studies used data for 100-500 patients [19,22,24,30,38,71], and 4 (7%) studies used data for less than 100 patients [18,49,66,69]. The number of patients was not reported in the rest of the studies.

After augmentation using GANs, the studies increased the number of images to several thousand, with a maximum number of 21,295 [54]. In 6 (11%) studies using GANs for data augmentation, the number of images increased to more than 10,000. In 3 (5%) studies, the number of images increased to 5000-10,000. In 9 (16%) studies, the number of images

increased to 1000-5000, and in 2 (4%) studies, the number of images increased between 500 and 1000. No study reported data augmentation output below 500 images.

Evaluation Mechanisms

Generally, the popular metrics for evaluating the diagnosis and classification performances of neural networks are accuracy, precision, recall, dice score, and area under the receiver operating characteristic curve (AUROC). To evaluate the performance of neural networks for diagnosis of COVID-19, 38 (67%) of the 57 studies used accuracy, along with metrics such as precision, recall, and dice score [21,23-28,31-34,36,38,40,43-48,52,53,55,56,58-60,63-72,74]. Around one-fourth of the studies (n=18, 32%) used sensitivity and specificity. In addition, 12 (21%) studies used the AUROC [19,20,26,30,32,46-48,50,51,68,74]. The numbers do not add up, as many studies used more than 1 metric for evaluation. In addition to the metrics mentioned here, 1 (2%) study used additional metrics, namely concordance index and relative absolute error, to evaluate prognosis and survival prediction for patients with COVID-19 [22].

Likewise, the popular metrics used to assess the quality of the synthesized images are the structural similarity measure (SSIM), the peak signal-to-noise ratio (PSNR), and the Fréchet inception distance (FID). Of the 57 studies included, 6 (11%) used the SSIM [18,30,49,60-62], 5 (9%) used the PSNR [18,30,49,61,62], and 3 (5%) used the FID metric [18,43,62] for evaluation.

The majority of the studies (n=42, 74%) reported having the data split between independent training and test sets. A few of the studies (n=6, 11%) reported 5-fold or 10-fold cross-validation for training and evaluation of the model. For almost one-sixth of the studies (n=9, 16%), the information on cross-validation was not available.

Reproducibility and Secondary Evaluation

This review also summarizes the studies in which the authors provided the implementation code. Only 7 (12%) of the 57 studies provided links for their code [19,20,34,47,48,66,70]. Only 2 (4%) studies reported a secondary evaluation by radiologists/doctors/experts by presenting the outcome of the results obtained by their models [19,45]. In addition, 1 (2%) study presented the results of end-to-end diagnosis of COVID-19 from CT images to 3 radiologists for a second opinion [19], and 1 (2%) study presented synthetic X-ray images to 2 radiologists for a second opinion on the quality of the generated X-ray images [45].

Discussion

Principal Findings

In this review, a significant rise in the number of studies on the topic was found in 2021 compared to 2020. This makes sense as the first half of 2020 saw only initial cases of COVID-19 infection, and research on the use of GANs for COVID-19 had yet to gain pace. Lung radiology image data for COVID-19-positive examples gradually became available during this period and increased only in the latter part of 2020. The highest number of studies were published from China and India (n=22). There can be 2 possible reasons for this. First, the 2 countries hold the top 2 spots on the ranking of the world's most populous countries. Second, the COVID-19 pandemic started in China, hence prompting earlier research efforts there.

Interestingly, the same number of studies (n=6) were published from the United States and Egypt each. The correlation mapping in Figure 5 shows that most of the studies published in 2020 originated from China, India, Egypt, and Canada. However, in 2021, many other countries also contributed to the published research. The number of journal papers was twice that of conference papers. This is surprising as journal publications would typically require more time in paper processing compared to conferences. It can be possible that many authors turned to journal submissions as, during the start of the pandemic, many conferences were suspended initially before moving to the online (virtual) mode.

In the majority of the included studies (n=39), the main task was to perform diagnosis of COVID-19 using lung CT or X-ray images. In these studies, a GAN was used as a submodule of the overall framework, and diagnosis was performed with the help of variants of CNNs, such as ResNet, VGG16, and Inception-net. In the included studies, GANs were used for 7 different purposes: data augmentation, segmentation of lungs within chest radiology images, superresolution of lung images to improve the quality of the images, diagnosis of COVID-19 within the images, feature extraction, prognosis studies related to COVID-19, and synthesis of 3D volumes of CT. Around 73% of the included studies used GAN-based methods for data augmentation to address the data scarcity challenge of COVID-19. It is not unexpected, as data augmentation is the most popular application of GANs. Only 1 study used the 3D variant of GAN for 3D synthesis of CT volumes. This is not surprising as 3D synthesis of CT volumes using 3D GANs is computationally expensive. The computations for the 3D synthesis of CT volumes may exceed the available resources of the graphics processing unit (GPU).

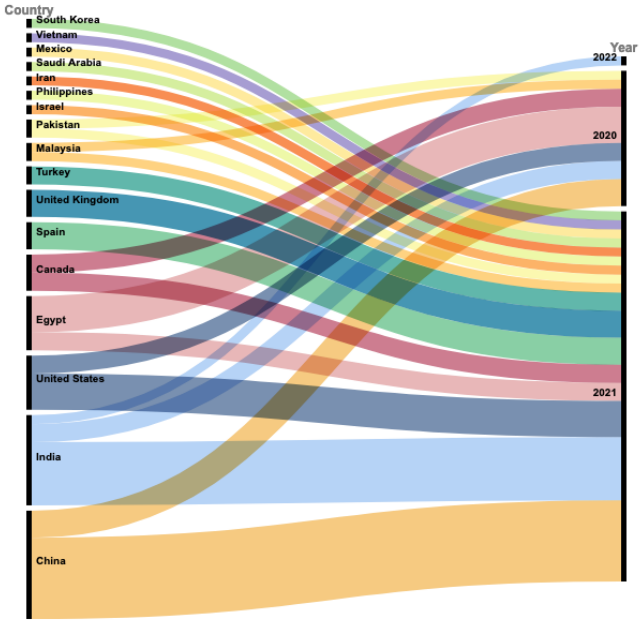
Since there are many variants of GANs, this review also looked at the most commonly used GAN architecture in the included studies. The most common choice of GAN in the included studies was the cycleGAN used in 9 studies. The cycleGAN is a GAN architecture that comprises 2 generators and 2 discriminators and does not require pair-to-pair training data [11]. Hence, it was a popular choice to generate COVID-19–positive images from normal images.

This review analyzed the common imaging modality for the different applications related to COVID-19. As chest X-ray imaging and CT scans are the most popular imaging methods for studying the infection in individuals, the studies included in this review were those that used these 2 imaging modalities. Specifically, 35 studies used X-ray images, and 21 studies used CT images. Some of the studies (n=6) also used both CT and X-ray images for diagnosis by training different models or for the transformation of images from X-ray to CT. Though ultrasound imaging is not prevalent in the clinical diagnosis of COVID-19, 1 study reported using ultrasound images to diagnose COVID-19 with GANs. No other modality of imaging was used by the included studies.

The majority of the included studies (n=47) used data that are available publicly on Github, Kaggle, or other publicly accessible websites. These data are acquired from multiple sources (eg, collected from more than 1 hospital or through crowdsourcing), which makes them more diverse and hence more useful for training of GAN models. Similarly, it is hoped that the use of publicly accessible data will also encourage other researchers to conduct experiments on the data sets. The rise of publications in 2021 can also be linked to the availability of publicly available data sets that continued to rise as the number of COVID-19 cases continued to grow. A few of the included studies (n=10) used private or proprietary data sets, and hence, the details about those data sets are only limited to what has been described in the corresponding studies.

Only 13 studies provided information on the number of individuals whose data were used in the included studies. Among these, only 1 study used data for more than 1000 individuals [26] and 2 studies used data for more than 500 individuals [29,42]. The remaining 10 studies used data for less than 500 individuals. Given the size of the population infected with COVID-19 (418+ million as of writing this, reported from John Hopkins University Coronavirus Resource Center [91]), the need for experiments with much more extensive data is obvious. As a result of having more data, learning inherent features within the radiology images by using GANs will become more generalized with training on larger data. There is still more need to contribute to publicly accessible data.

Figure 5. Mapping of correlation between publications from each country vs year of publication. Studies in 2020 originated mostly from China, India, Egypt, and Canada. In 2021, many other countries also contributed to the published research.



Practical and Research Implications

This review presented the different studies that used GANs for various COVID-19 applications. Data augmentation of COVID-19 imaging data was the most common application in the included studies. The augmented data can significantly improve the training of AI methods, particularly deep learning methods used for COVID-19 diagnosis. This review found that for most of the studies, the current CT and X-ray imaging data (even if smaller in size) are already available through publicly accessible links on Github, Kaggle, or institutional websites. This should encourage more researchers to build upon the available data sets and train more variants of deep learning and GAN-based methods to speed up the research progress on COVID-19. Similarly, researchers can also add to the existing data set on Github by uploading their data to the current data repositories. An example of crowdsourcing of data is the COVIDx image repository for lung X-ray images (see Table 3).

This review identified that the code to reproduce the results was not available for the majority of the studies. Only 7 of the included studies provided a public link to the code. Availability of a public repository to reproduce the results for diagnosis or augmented data can help in advancing the research as well as increase the trust and reliance on the reported results in terms of the quality of the generated images or the accuracy reports for the diagnosis. In addition, the reproducibility by this code was not assessed by this review, as it was beyond the scope of this review. Careful and responsible studies are needed to make an assessment of the published methods for transformation into clinical applications.

The majority of the included studies (n=43) did not provide information on the number of patients, although they did mention the number of images used in the experiments. So, it is unclear how many images were used per individual. Hence, the lack of information limits the ability of the readers to evaluate the performance in the context of the number of patients. Moreover, for public data sets with crowd-sourced contributions, it is challenging to trace back the number of images to the number of individuals.

Validation of the performance of GANs in terms of the quality/usability of the generated images has a significant role in promoting the acceptability of the methods. Of the included

studies, only 2 studies reported that the results were presented to radiologists/clinicians for a secondary validation. In 1 study on the synthesis of X-ray images, the radiologists agreed that the quality of the X-rays has improved but falls short of diagnostic quality for use in clinics [45]. Although using GAN-based methods in COVID-19 is tempting for many researchers, the lack of evaluation by radiologists or using GAN-based methods without radiologists and clinicians in the loop will hinder the acceptability of these methods for clinical applications. In addition, it is beyond the scope of this review to evaluate a study based on reporting of secondary evaluation by the radiologists, though a secondary assessment by the radiologists would have added value to the studies and increased their acceptability. The lack of details related to the individuals whose COVID-19 data were used in these studies may also hinder their acceptance for transformation into clinical applications. The training of GANs is usually computationally demanding, requiring GPUs. More edge computing–based implementations are needed for clinical applications to make these models compatible for implementation on low-power devices. This will increase the acceptability of these methods in clinical devices.

Strengths and Limitations

Strengths

Though several reviews can be found on the applications of AI techniques in COVID-19, no review was found that focused on the potential of GAN-based methods to combat COVID-19. Compared to other reviews [3,4,6,7] where the scope is too broad as they attempted to cover many different AI models, this review provided a comprehensive analysis of the GAN-based approaches used primarily on lung CT and X-ray images. Similarly, many reviews covered the applications of GANs in medical imaging [10,12-15]; their applications in lung images for COVID-19 have not been reviewed before. So, this review may be considered the first comprehensive review that covers all the GAN-based methods used for COVID-19 imaging data for different applications in general and data augmentation in particular. Thus, it is helpful for the readers to understand how GAN-based approaches were used to address the problem of data scarcity and how the synthetic data (generated by GANs) were used to improve the performance of CNNs for COVID-19. This review provided a thorough list of the various publicly available data sets of lung CT, lung X-ray, and lung ultrasound images. Hence, this can serve as a single point of contact for the readers to explore these data set resources and use them in their research work. This review is consistent with the PRISMA-ScR guidelines for scientific reviews [16].

Limitations

This review included studies from 5 databases: PubMed, IEEEXplore, ACM Digital Library, Scopus, and Google Scholar. Hence, it is possible that some literature that is not indexed in these libraries might have been left out. However, given the coverage by these popular databases, the included studies form a comprehensive representation of the applications of GANs in COVID-19. The review, for practical reasons, included studies published only in English and did not include studies in other languages. Since the scope of this review was limited to lung images only, the potential of GANs for other types of medical data, such as electronic health records, textual data, and audio data (recordings of coughing), was not covered in this review. The results and interpretations presented in this review are derived from the available information in the included studies. Since different studies may have variations and even missing details in their reporting of the data set, the training and test sets, and the validation mechanism, a direct comparison of the results might not be possible. Inconsistent information on the number of images, the training mechanism for GANs, and the selection of test set examples may have affected the findings of this review. In addition, by modern standards of training deep learning models, the size of data reported in most included studies is too small. So, the results reported in the studies in terms of diagnosis accuracy may not generalize well. The findings and the discussions of this review are mainly based on the authors' understanding of GANs (and other AI methods) and do not necessarily reflect the comments and feedback of the doctors and clinicians.

Conclusion

This scoping review provided a comprehensive review of 57 studies on the use of GANs for COVID-19 lung imaging data. Similar to other deep learning and AI methods, GANs have demonstrated outstanding potential in research on addressing COVID-19 diagnosis performance. However, the most significant application of GANs has been data augmentation by generating synthetic chest CT or X-ray imaging data from the existing limited-size data, as the synthetic data showed a direct bearing on the enhancement of the diagnosis. Although GAN-based methods have demonstrated great potential, their adoption in COVID-19 research is still in a stage of infancy. Notably, the transformation of GAN-based methods into clinical applications is still limited due to the limitations in the validation of the results, the generalization of the results, the lack of feedback from radiologists, and the limited explainability offered by these methods. Nevertheless, GAN-based methods can assist in the performance enhancement of COVID-19 diagnosis, even though they should not be used as independent tools. In addition, more research and advancements are needed toward the explainability and clinical transformations of these methods. This will pave the way for a broader acceptance of GAN-based methods in COVID-19 applications.

Acknowledgments

HA contributed to the conception, design, literature search, data selection, data synthesis, data extraction, and drafting. ZS contributed to the design, data selection, data synthesis, and critical revision of the manuscript. All authors gave their final approval and accepted accountability for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 19 KB - medinform_v10i6e37365_app1.docx]

Multimedia Appendix 2
Interrater agreement matrices for study selection steps.
[DOCX File , 22 KB - medinform_v10i6e37365_app2.docx]

Multimedia Appendix 3
Data extraction form.
[DOCX File , 24 KB - medinform_v10i6e37365_app3.docx]

Multimedia Appendix 4
Characteristics of the included studies.
[XLSX File (Microsoft Excel File), 23 KB - medinform_v10i6e37365_app4.xlsx]

References

1. Chen Y, Li L. SARS-CoV-2: virus dynamics and host response. Lancet Infect Dis 2020 May;20(5):515-516 [FREE Full text] [doi: 10.1016/S1473-3099(20)30235-8] [Medline: 32213336]
2. Li Y, Yao L, Li J, Chen L, Song Y, Cai Z, et al. Stability issues of RT-PCR testing of SARS-CoV-2 for hospitalized patients clinically diagnosed with COVID-19. J Med Virol 2020 Jul 26;92(7):903-908 [FREE Full text] [doi: 10.1002/jmv.25786] [Medline: 32219885]
3. Abd-Alrazaq A, Alajlani M, Alhuwail D, Schneider J, Al-Kuwari S, Shah Z, et al. Artificial intelligence in the fight against COVID-19: scoping review. J Med Internet Res 2020 Dec 15;22(12):e20756 [FREE Full text] [doi: 10.2196/20756] [Medline: 33284779]
4. Wang L, Zhang Y, Wang D, Tong X, Liu T, Zhang S, et al. Artificial intelligence for COVID-19: a systematic review. Front Med (Lausanne) 2021 Sep 30;8:704256 [FREE Full text] [doi: 10.3389/fmed.2021.704256] [Medline: 34660623]
5. Chowdhury MEH, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, et al. Can AI help in screening viral and COVID-19 pneumonia? IEEE Access 2020;8:132665-132676. [doi: 10.1109/access.2020.3010287]
6. Alimadadi A, Aryal S, Manandhar I, Munroe PB, Joe B, Cheng X. Artificial intelligence and machine learning to fight COVID-19. Physiol Genomics 2020 Apr 01;52(4):200-202 [FREE Full text] [doi: 10.1152/physiolgenomics.00029.2020] [Medline: 32216577]
7. Latif S, Usman M, Manzoor S, Iqbal W, Qadir J, Tyson G, et al. Leveraging data science to combat COVID-19: a comprehensive review. IEEE Trans Artif Intell 2020 Aug;1(1):85-103. [doi: 10.1109/tai.2020.3020521]
8. Ali H, Umander J, Rohlen R, Gronlund C. A deep learning pipeline for identification of motor units in musculoskeletal ultrasound. IEEE Access 2020;8:170595-170608. [doi: 10.1109/access.2020.3023495]
9. Iqbal T, Ali H. Generative adversarial network for medical images (MI-GAN). J Med Syst 2018 Oct 12;42(11):231. [doi: 10.1007/s10916-018-1072-9] [Medline: 30315368]
10. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. Med Image Anal 2019 Dec;58:101552. [doi: 10.1016/j.media.2019.101552] [Medline: 31521965]
11. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proc IEEE Int Conf Comput Vis 2017:2223-2232. [doi: 10.1109/iccv.2017.244]
12. Lan L, You L, Zhang Z, Fan Z, Zhao W, Zeng N, et al. Generative adversarial networks and its applications in biomedical informatics. Front Public Health 2020 May 12;8:164 [FREE Full text] [doi: 10.3389/fpubh.2020.00164] [Medline: 32478029]
13. Singh NK, Raza K. Medical image generation using generative adversarial networks: a review. In: Patgiri R, Biswas A, Roy P, editors. Health Informatics: A Computational Perspective in Healthcare. Studies in Computational Intelligence, Volume 932. Singapore: Springer; 2021.
14. Wang T, Lei Y, Fu Y, Wynne JF, Curran WJ, Liu T, et al. A review on medical imaging synthesis using deep learning and its clinical applications. J Appl Clin Med Phys 2021 Jan 11;22(1):11-36 [FREE Full text] [doi: 10.1002/acm2.13121] [Medline: 33305538]
15. Saeed AQ, Sheikh Abdullah SNH, Che-Hamzah J, Abdul Ghani AT. Accuracy of using generative adversarial networks for glaucoma detection: systematic review and bibliometric analysis. J Med Internet Res 2021 Sep 21;23(9):e27414 [FREE Full text] [doi: 10.2196/27414] [Medline: 34236992]

16. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Tunçalp, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473. [doi: 10.7326/m18-0850]
17. Higgins J, Deeks J. Chapter 7: selecting studies and collecting data. In: Cochrane Handbook for Systematic Reviews of Interventions. Hoboken, NJ: John Wiley & Sons; 2008.
18. Jiang Y, Chen H, Loew M, Ko H. COVID-19 CT image synthesis with a conditional generative adversarial network. IEEE J Biomed Health Inform 2021 Feb;25(2):441-452. [doi: 10.1109/jbhi.2020.3042523]
19. Song J, Wang H, Liu Y, Wu W, Dai G, Wu Z, et al. End-to-end automatic differentiation of the coronavirus disease 2019 (COVID-19) from viral pneumonia based on chest CT. Eur J Nucl Med Mol Imaging 2020 Oct 22;47(11):2516-2524 [FREE Full text] [doi: 10.1007/s00259-020-04929-1] [Medline: 32567006]
20. Motamed S, Rogalla P, Khalvati F. RANDGAN: randomized generative adversarial network for detection of COVID-19 in chest X-ray. Sci Rep 2021 Apr 21;11(1):8602-2524 [FREE Full text] [doi: 10.1038/s41598-021-87994-2] [Medline: 33883609]
21. Autee P, Bagwe S, Shah V, Srivastava K. StackNet-DenVIS: a multi-layer perceptron stacked ensembling approach for COVID-19 detection using X-ray images. Phys Eng Sci Med 2020 Dec 04;43(4):1399-1414 [FREE Full text] [doi: 10.1007/s13246-020-00952-6] [Medline: 33275187]
22. Uemura T, Näppi JJ, Watari C, Hironaka T, Kamiya T, Yoshida H. Weakly unsupervised conditional generative adversarial network for image-based prognostic prediction for COVID-19 patients based on chest CT. Med Image Anal 2021 Oct;73:102159 [FREE Full text] [doi: 10.1016/j.media.2021.102159] [Medline: 34303892]
23. Goel T, Murugan R, Mirjalili S, Chakrabartty DK. Automatic screening of COVID-19 using an optimized generative adversarial network. Cognit Comput 2021 Jan 25:1-16 [FREE Full text] [doi: 10.1007/s12559-020-09785-7] [Medline: 33520007]
24. Loey M, Manogaran G, Khalifa NEM. A deep transfer learning model with classical data augmentation and CGAN to detect COVID-19 from chest CT radiography digital images. Neural Comput Appl 2020 Oct 26:1-13 [FREE Full text] [doi: 10.1007/s00521-020-05437-x] [Medline: 33132536]
25. Karakanis S, Leontidis G. Lightweight deep learning models for detecting COVID-19 from chest X-ray images. Comput Biol Med 2021 Mar;130:104181 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.104181] [Medline: 33360271]
26. Li Z, Zhang J, Li B, Gu X, Luo X. COVID-19 diagnosis on CT scan images using a generative adversarial network and concatenated feature pyramid network with an attention mechanism. Med Phys 2021 Aug 09;48(8):4334-4349 [FREE Full text] [doi: 10.1002/mp.15044] [Medline: 34117783]
27. Zhang L, Shen B, Barnawi A, Xi S, Kumar N, Wu Y. FedDPGAN: federated differentially private generative adversarial networks framework for the detection of COVID-19 pneumonia. Inf Syst Front 2021 Jun 15;23(6):1403-1415 [FREE Full text] [doi: 10.1007/s10796-021-10144-6] [Medline: 34149305]
28. Rasheed J, Hameed AA, Djeddi C, Jamil A, Al-Turjman F. A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images. Interdiscip Sci 2021 Mar 02;13(1):103-117 [FREE Full text] [doi: 10.1007/s12539-020-00403-6] [Medline: 33387306]
29. Morís DI, de Moura Ramos JJ, Buján JN, Hortas MO. Data augmentation approaches using cycle-consistent adversarial networks for improving COVID-19 screening in portable chest X-ray images. Expert Syst Appl 2021 Dec 15;185:115681 [FREE Full text] [doi: 10.1016/j.eswa.2021.115681] [Medline: 34366577]
30. Zhang Q, Chen Z, Liu G, Zhang W, Du Q, Tan J, et al. Artificial intelligence clinicians can use chest computed tomography technology to automatically diagnose coronavirus disease 2019 (COVID-19) pneumonia and enhance low-quality images. IDR 2021 Feb;Volume 14:671-687. [doi: 10.2147/idr.s296346]
31. Singh RK, Pandey R, Babu RN. COVIDScreen: explainable deep learning framework for differential diagnosis of COVID-19 using chest X-rays. Neural Comput Appl 2021 Jan 08;33(14):8871-8892 [FREE Full text] [doi: 10.1007/s00521-020-05636-6] [Medline: 33437132]
32. Karbhari Y, Basu A, Geem ZW, Han G, Sarkar R. Generation of synthetic chest X-ray images and detection of COVID-19: a deep learning based approach. Diagnostics (Basel) 2021 May 18;11(5):895 [FREE Full text] [doi: 10.3390/diagnostics11050895] [Medline: 34069841]
33. Amin J, Sharif M, Gul N, Kadry S, Chakraborty C. Quantum machine learning architecture for COVID-19 classification based on synthetic data generation using conditional adversarial neural network. Cognit Comput 2021 Aug 10:1-12 [FREE Full text] [doi: 10.1007/s12559-021-09926-6] [Medline: 34394762]
34. Zhang J, Yu L, Chen D, Pan W, Shi C, Niu Y, et al. Dense GAN and multi-layer attention based lesion segmentation method for COVID-19 CT images. Biomed Signal Process Control 2021 Aug;69:102901 [FREE Full text] [doi: 10.1016/j.bspc.2021.102901] [Medline: 34178095]
35. Hernandez-Cruz N, Cato D, Favela J. Neural style transfer as data augmentation for improving COVID-19 diagnosis classification. SN Comput Sci 2021 Aug 13;2(5):410 [FREE Full text] [doi: 10.1007/s42979-021-00795-2] [Medline: 34405153]
36. Jiang H, Tang S, Liu W, Zhang Y. Deep learning for COVID-19 chest CT (computed tomography) image analysis: a lesson from lung cancer. Comput Struct Biotechnol J 2021;19:1391-1399 [FREE Full text] [doi: 10.1016/j.csbj.2021.02.016] [Medline: 33680351]

37. Bhattacharyya A, Bhaik D, Kumar S, Thakur P, Sharma R, Pachori RB. A deep learning based approach for automatic detection of COVID-19 cases using chest X-ray images. Biomed Signal Process Control 2022 Jan;71:103182 [FREE Full text] [doi: 10.1016/j.bspc.2021.103182] [Medline: 34580596]
38. Mann P, Jain S, Mittal A, Bhat A. Generation of COVID-19 chest CT scan images using generative adversarial networks. 2021 Presented at: 2021 International Conference on Intelligent Technologies (CONIT); June 25-27, 2021; Hubbali, Karnataka, India p. 1-5. [doi: 10.1109/conit51480.2021.9498272]
39. Quan T, Thanh H, Huy T, Chanh N, Anh N, Vu P, et al. XPGAN: X-ray projected generative adversarial network for improving COVID-19 image classification. 2021 Presented at: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI); 2021; Nice, France p. 1509-1513. [doi: 10.1109/isbi48211.2021.9434159]
40. Waheed A, Goyal M, Gupta D, Khanna A, Al-Turjman F, Pinheiro PR. CovidGAN: data augmentation using auxiliary classifier GAN for improved COVID-19 detection. IEEE Access 2020;8:91916-91923. [doi: 10.1109/access.2020.2994762]
41. Liang Z, Huang J, Li J, Chan S. Enhancing automated COVID-19 chest X-ray diagnosis by image-to-image GAN translation. 2020 Presented at: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 16-19, 2020; Seoul, South Korea p. 1068-1071. [doi: 10.1109/bibm49941.2020.9313466]
42. Morís D, de MJ, Novo J, Ortega M. Cycle generative adversarial network approaches to produce novel portable chest X-rays images for COVID-19 diagnosis. 2021 Presented at: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); June 6-11, 2021; Toronto, Ontario, Canada p. 1060-1064. [doi: 10.1109/icassp39728.2021.9414031]
43. Rodríguez-de-la-Cruz J, Acosta-Mesa H, Mezura-Montes E. Evolution of generative adversarial networks using pso for synthesis of covid-19 chest x-ray images. 2021 Presented at: 2021 IEEE Congress on Evolutionary Computation (CEC) 2021 Jun 28 (pp.). IEEE; 2021; Kraków, Poland p. 2226-2233. [doi: 10.1109/cec45853.2021.9504743]
44. Nneji G, Cai J, Jianhua D, Monday H, Chikwendu I, Oluwasanmi A, et al. Enhancing low quality in radiograph datasets using wavelet transform convolutional neural network and generative adversarial network for COVID-19 identification. 2021 Presented at: 2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI); August 20-22, 2021; Yibin, China p. 146-151. [doi: 10.1109/prai53619.2021.9551043]
45. Menon S, Galita J, Chapman D, Gangopadhyay A, Mangalagiri J, Nguyen P, et al. Generating realistic COVID-19 x-rays with a mean teacher+ transfer learning GAN. 2020 Presented at: 2020 IEEE International Conference on Big Data (Big Data); December 10-13, 2020; Virtual p. 1216-1225. [doi: 10.1109/bigdata50022.2020.9377878]
46. Dong S, Zhang Z. Joint optimization of cycleGAN and CNN classifier for COVID-19 detection and biomarker localization. 2020 Presented at: 2020 IEEE International Conference on Progress in Informatics and Computing (PIC); December 17-19, 2021; Shanghai, China p. 112-118. [doi: 10.1109/pic50277.2020.9350813]
47. Yadav P, Menon N, Ravi V, Vishvanathan S. Lung-GANs: unsupervised representation learning for lung disease classification using chest CT and X-ray images. IEEE Trans Eng Manag 2021:1-13. [doi: 10.1109/tem.2021.3103334]
48. Yang Z, Zhao L, Wu S, Chen CY. Lung lesion localization of COVID-19 from chest CT image: a novel weakly supervised learning method. IEEE J Biomed Health Inform 2021 Jun;25(6):1864-1872. [doi: 10.1109/jbhi.2021.3067465]
49. Mangalagiri J, Chapman D, Gangopadhyay A, Yesha Y, Galita J, Menon S, et al. Toward generating synthetic CT volumes using a 3D-conditional generative adversarial network. 2020 Presented at: 2020 International Conference on Computational Science and Computational Intelligence (CSCI); December 16-18, 2020; Las Vegas, NV p. 858-862. [doi: 10.1109/csci51800.2020.00160]
50. Sakib S, Tazrin T, Fouda MM, Fadlullah ZM, Guizani M. DL-CRC: deep learning-based chest radiograph classification for COVID-19 detection: a novel approach. IEEE Access 2020;8:171575-171589. [doi: 10.1109/access.2020.3025010]
51. Yang Y, Chen J, Wang R, Ma T, Wang L, Chen J, et al. Towards unbiased COVID-19 lesion localisation and segmentation via weakly supervised learning. 2021 Presented at: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI); April 13-16, 2021; Virtual p. 1966-1970. [doi: 10.1109/isbi48211.2021.9433806]
52. Loey M, Smarandache F, M. Khalifa NE. Within the lack of chest COVID-19 X-ray dataset: a novel detection model based on GAN and deep transfer learning. Symmetry 2020 Apr 20;12(4):651. [doi: 10.3390/sym12040651]
53. Khalifa N, Taha M, Hassanien A, Taha S. The detection of COVID-19 in CT medical images: a deep learning approach. In: Big Data Analytics and Artificial Intelligence against COVID-19: Innovation Vision and Approach. Cham: Springer; 2020:73-90.
54. Zunair H, Hamza AB. Synthesis of COVID-19 chest X-rays using unpaired image-to-image translation. Soc Netw Anal Min 2021 Feb 24;11(1):23 [FREE Full text] [doi: 10.1007/s13278-021-00731-5] [Medline: 33643491]
55. Sachdev J, Bhatnagar N, Bhatnagar R. Deep learning models using auxiliary classifier GAN for COVID-19 detection–a comparative study. In: The International Conference on Artificial Intelligence and Computer Vision. Cham: Springer; 2021:12-23.
56. Morís D, de MJ, Novo J, Ortega M. Portable chest X-ray synthetic image generation for the COVID-19 screening. Eng Proc 2021;7(1):6. [doi: 10.3390/engproc2021007006]
57. Munawar F, Azmat S, Iqbal T, Gronlund C, Ali H. Segmentation of lungs in chest X-ray image using generative adversarial networks. IEEE Access 2020;8:153535-153545. [doi: 10.1109/access.2020.3017915]

58. Oluwasanmi A, Aftab MU, Qin Z, Ngo ST, Doan TV, Nguyen SB, et al. Transfer learning and semisupervised adversarial detection and classification of COVID-19 in CT images. Complexity 2021 Feb 13;2021:1-11. [doi: 10.1155/2021/6680455]
59. Sanajalwe Y, Anbar M, Al-E'mari S. COVID-19 automatic detection using deep learning. Comput Syst Sci Eng 2021 Jan 1;39(1):15-35. [doi: 10.32604/csse.2021.017191]
60. Al-Shargabi AA, Alshobaili JF, Alabdulatif A, Alrobah N. COVID-CGAN: efficient deep learning approach for COVID-19 detection based on CXR images using conditional GANs. Appl Sci 2021 Aug 04;11(16):7174. [doi: 10.3390/app11167174]
61. Shivadekar S, Mangalagiri J, Nguyen P, Chapman D, Halem M, Gite R. An intelligent parallel distributed streaming framework for near real-time science sensors and high-resolution medical images. 2021 Presented at: 50th International Conference on Parallel Processing Workshop; August 9-12, 2021; Chicago, IL p. 1-9. [doi: 10.1145/3458744.3474039]
62. Jamal TO, O'Reilly UM. Signal propagation in a gradient-basedevolutionary learning system. 2021 Presented at: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '21); July 10-14, 2021; Lille, France. [doi: 10.1145/3449639.3459319]
63. Acar E, Şahin E, Yılmaz İ. Improving effectiveness of different deep learning-based models for detecting COVID-19 from computed tomography (CT) images. Neural Comput Appl 2021 Jul 29;33(24):1-21 [FREE Full text] [doi: 10.1007/s00521-021-06344-5] [Medline: 34345118]
64. Sheykhivand S, Mousavi Z, Mojtahedi S, Yousefi Rezaii T, Farzamnia A, Meshgini S, et al. Developing an efficient deep neural network for automatic detection of COVID-19 using chest X-ray images. Alex Eng J 2021 Jun;60(3):2885-2903. [doi: 10.1016/j.aej.2021.01.011]
65. Rangarajan AK, Ramachandran HK. A preliminary analysis of AI based smartphone application for diagnosis of COVID-19 using chest X-ray images. Expert Syst Appl 2021 Nov;183:115401. [doi: 10.1016/j.eswa.2021.115401]
66. Li H, Hu Y, Li S, Lin W, Liu P, Higashita R, et al. CT scan synthesis for promoting computer-aided diagnosis capacity of COVID-19. 2020 Presented at: International Conference on Intelligent Computing 2020; October 2-5, 2020; Bari, Italy p. 413-422. [doi: 10.1007/978-3-030-60802-6_36]
67. Zulkifley MA, Abdani SR, Zulkifley NH. COVID-19 screening using a lightweight convolutional neural network with generative adversarial network data augmentation. Symmetry 2020 Sep 16;12(9):1530. [doi: 10.3390/sym12091530]
68. El-Shafai W, Ali A, El-Rabaie E, Soliman N, Algarni A, El-Samie A. Automated COVID-19 detection based on single-image super-resolution and CNN models. Comput Mater Continua 2021:1141-1157. [doi: 10.32604/cmc.2022.018547]
69. Karar M, Shouman M, Chalopin C. Adversarial neural network classifiers for COVID-19 diagnosis in ultrasound images. Comput Mater Continua 2021:1683-1697. [doi: 10.32604/cmc.2022.018564]
70. Zebin T, Rezvy S. COVID-19 detection and disease progression visualization: deep learning on chest X-rays for classification and coarse localization. Appl Intell (Dordr) 2021 Sep 12;51(2):1010-1021 [FREE Full text] [doi: 10.1007/s10489-020-01867-1] [Medline: 34764549]
71. Ambita A, Boquio E, Naval P. COViT-GAN: vision transformer for COVID-19 detection in CT scan images with self-attention GAN for data augmentation. 2021 Presented at: International Conference on Artificial Neural Networks 2021; September 2021; Virtual p. 587-598. [doi: 10.1007/978-3-030-86340-1_47]
72. Elghamrawy S. An H2O's deep learning-inspired model based on big data analytics for coronavirus disease (COVID-19) diagnosis. In: Big Data Analytics and Artificial Intelligence against COVID-19: Innovation Vision and Approach. Cham: Springer; 2021:263-279.
73. Toutouh J, Esteban M, Nesmachnow S. Parallel/distributed generative adversarial neural networks for data augmentation of COVID-19 training images. 2020 Presented at: Latin American High Performance Computing Conference 2020; September 2-4, 2020; Virtual p. 162-177. [doi: 10.1007/978-3-030-68035-0_12]
74. Bar-El A, Cohen D, Cahan N, Greenspan H. Improved cycleGAN with application to COVID-19 classification. Proc SPIE 2021;11596:1159614. [doi: 10.1117/12.2582162]
75. SARS-COV-2 Ct-Scan Dataset: A Large Dataset of CT Scans for SARS-CoV-2 (COVID-19) Identification. URL: https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset [accessed 2022-06-22]
76. COVID-CT. URL: https://github.com/UCSD-AI4H/COVID-CT [accessed 2022-06-22]
77. HKBU_HPML_COVID-19. URL: https://github.com/wang-shihao/HKBU_HPML_COVID-19 [accessed 2022-06-22]
78. covid-chestxray-dataset. URL: https://github.com/ieee8023/covid-chestxray-dataset [accessed 2022-06-22]
79. Actualmed COVID-19 Chest X-ray Dataset Initiative. URL: https://github.com/agchung/Actualmed-COVID-chestxray-dataset [accessed 2022-06-22]
80. COVID-19 Radiography Database. URL: https://www.kaggle.com/tawsifurrahman/covid19-radiography-database [accessed 2022-06-22]
81. Figure 1 COVID-19 Chest X-ray Dataset Initiative. URL: https://github.com/agchung/Figure1-COVID-chestxray-dataset [accessed 2022-06-22]
82. Chest X-Ray Images (Pneumonia). URL: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia [accessed 2022-06-22]
83. Extensive COVID-19 X-Ray and CT Chest Images Dataset. URL: https://data.mendeley.com/datasets/8h65ywd2jr/3 [accessed 2022-06-22]
84. COVID-19 CT Segmentation Dataset. URL: https://medicalsegmentation.com/covid19/ [accessed 2022-06-22]

85. COVID-19 Open Research Dataset Challenge (CORD-19). URL: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge [accessed 2022-06-22]
86. RSNA Pneumonia Detection Challenge: Can You Build an Algorithm That Automatically Detects Potential Pneumonia Cases?. URL: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data [accessed 2022-06-22]
87. CI Images and Clinical Features for COVID-19. URL: http://ictcf.biocuckoo.cn/HUST-19.php [accessed 2022-06-22]
88. Automatic Detection of COVID-19 from Ultrasound Data. URL: https://github.com/jannisborn/covid19_ultrasound [accessed 2022-06-22]
89. COVID-19 Chest Xray. URL: https://www.kaggle.com/bachrr/covid-chest-xray [accessed 2022-06-22]
90. Società Italiana di Radiologia Medica e Interventistica. URL: https://sirm.org/category/senza-categoria/covid-19/ [accessed 2022-06-22]
91. John Hopkins University Coronavirus Resource Center. URL: https://coronavirus.jhu.edu/ [accessed 2022-06-21]

Abbreviations

ACM: Association for Computing Machinery
AI: artificial intelligence
AUROC: area under the receiver operating characteristic curve
CNN: convolutional neural network
CT: computed tomography
FID: Fréchet inception distance
GAN: generative adversarial network
GPU: graphics processing unit
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
PSNR: peak signal-to-noise ratio
RT-PCR: reverse transcription–polymerase chain reaction
SSIM: structural similarity measure

Edited by C Lovis; submitted 17.02.22; peer-reviewed by S Khan, I Gabashvili; comments to author 02.03.22; revised version received 06.03.22; accepted 11.03.22; published 29.06.22.

Please cite as:
Ali H, Shah Z
Combating COVID-19 Using Generative Adversarial Networks and Artificial Intelligence for Medical Images: Scoping Review
JMIR Med Inform 2022;10(6):e37365
URL: https://medinform.jmir.org/2022/6/e37365
doi:10.2196/37365
PMID:35709336

Original Paper

Understanding the Relationship Between Mood Symptoms and Mobile App Engagement Among Patients With Breast Cancer Using Machine Learning: Case Study

Anna N Baglione[1], BS, MS; Lihua Cai[1], BS, MA, MS, PhD; Aram Bahrini[1], BSc, ME, MS; Isabella Posey[2], BS; Mehdi Boukhechba[1], BS, MS, PhD; Philip I Chow[3], BA, PhD

[1]Department of Engineering Systems and Environment, University of Virginia, Charlottesville, VA, United States

[2]Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY, United States

[3]Center for Behavioral Health & Technology, School of Medicine, University of Virginia, Charlottesville, VA, United States

Corresponding Author:
Anna N Baglione, BS, MS
Department of Engineering Systems and Environment
University of Virginia
Olsson Hall
151 Engineer's Way
Charlottesville, VA, 22904
United States
Phone: 1 434 264 7484
Email: ab5bt@virginia.edu

Abstract

Background: Health interventions delivered via smart devices are increasingly being used to address mental health challenges associated with cancer treatment. Engagement with mobile interventions has been associated with treatment success; however, the relationship between mood and engagement among patients with cancer remains poorly understood. A reason for this is the lack of a data-driven process for analyzing mood and app engagement data for patients with cancer.

Objective: This study aimed to provide a step-by-step process for using app engagement metrics to predict continuously assessed mood outcomes in patients with breast cancer.

Methods: We described the steps involved in data preprocessing, feature extraction, and data modeling and prediction. We applied this process as a case study to data collected from patients with breast cancer who engaged with a mobile mental health app intervention (IntelliCare) over 7 weeks. We compared engagement patterns over time (eg, frequency and days of use) between participants with high and low anxiety and between participants with high and low depression. We then used a linear mixed model to identify significant effects and evaluate the performance of the random forest and XGBoost classifiers in predicting weekly mood from baseline affect and engagement features.

Results: We observed differences in engagement patterns between the participants with high and low levels of anxiety and depression. The linear mixed model results varied by the feature set; these results revealed weak effects for several features of engagement, including duration-based metrics and frequency. The accuracy of predicting depressed mood varied according to the feature set and classifier. The feature set containing survey features and overall app engagement features achieved the best performance (accuracy: 84.6%; precision: 82.5%; recall: 64.4%; F1 score: 67.8%) when used with a random forest classifier.

Conclusions: The results from the case study support the feasibility and potential of our analytic process for understanding the relationship between app engagement and mood outcomes in patients with breast cancer. The ability to leverage both self-report and engagement features to analyze and predict mood during an intervention could be used to enhance decision-making for researchers and clinicians and assist in developing more personalized interventions for patients with breast cancer.

(JMIR Med Inform 2022;10(6):e30712) doi:10.2196/30712

KEYWORDS

breast cancer; digital intervention; mobile intervention; mobile health; mHealth; app engagement; user engagement; mental health; depression; anxiety

Introduction

Background

In the United States, 1 in 8 women will receive a breast cancer diagnosis at some point in her lifetime [1]. Breast cancer is currently the leading cause of cancer death in women [2]. Patients with breast cancer encounter a range of psychosocial stressors that extend beyond the physical effects of anticancer treatment, including emotional distress, diminished well-being, and increased symptoms of depression and anxiety [3,4]. Untreated symptoms of depression and anxiety in women with breast cancer can lead to poor quality of life [5], increased mortality [6], and high economic costs [7].

Interventions that emphasize skill acquisition, such as cognitive behavioral therapy, have been shown to effectively reduce symptoms of depression and anxiety in patients with breast cancer [8,9]. However, numerous barriers prevent patients with cancer from receiving adequate treatment, including high financial [10] and time [11] costs, social stigma [12], and a severe shortage of trained psychotherapists, particularly in rural and underserved areas [13]. Combined, these barriers lead to almost half of breast cancer survivors reporting unmet psychosocial needs [14].

Increasingly, researchers are leveraging mobile phone apps to address mental health issues in patients with cancer. Apps are frequently cited as a way of extending cost-effective care [15,16]. In many cases, digital interventions (ie, web-based and app-delivered interventions) that mirror the content of in-person therapy perform just as well in reducing mood symptoms [17,18]. App-delivered interventions can decrease barriers associated with traditional in-person interventions as treatment is affordable, is readily available, offers efficient use of time (ie, no delays to begin treatment and self-pacing), and is no longer limited by factors such as geographic proximity to available psychotherapists. This is particularly relevant for women undergoing anticancer treatment regimens who may only have small pockets of unstructured time in a day. Numerous studies have validated the use of apps to reduce depression and anxiety symptoms [19,20], including in patients with breast cancer.

Although access to high-quality treatment is a major issue that app-delivered interventions are well poised to address, sustained engagement is a common problem [21]. Engagement is critical as it is necessary for treatment success, as studies have documented a dose-response relationship in app interventions [22,23]. A barrier to advancing knowledge of engagement in digital interventions is data density. It is common for app-delivered interventions to be deployed by a user when and where they are most convenient, potentially leading to a large data set. Fortunately, advances in machine learning have made it possible to analyze vast volumes of engagement data. However, translating these raw engagement data into clinically meaningful observations is an ongoing challenge in oncology research using mobile health (mHealth) tools [24]. Moreover, to date, no studies have presented a clear process for analyzing the relationship between engagement with mental health apps and outcomes in cancer populations using machine learning.

Objectives

This study aimed to develop a process for investigating the dynamic relationship between engagement with a mental health app intervention and mood. The process involves several steps, including cleaning and preprocessing the raw app use data, extracting features of mood and engagement, and predicting moods from these features using machine learning algorithms. To demonstrate the application and potential usefulness of this process, we applied it to a limited number of newly diagnosed patients with breast cancer who participated in a 7-week trial that evaluated the efficacy of a suite of mental health apps [25].

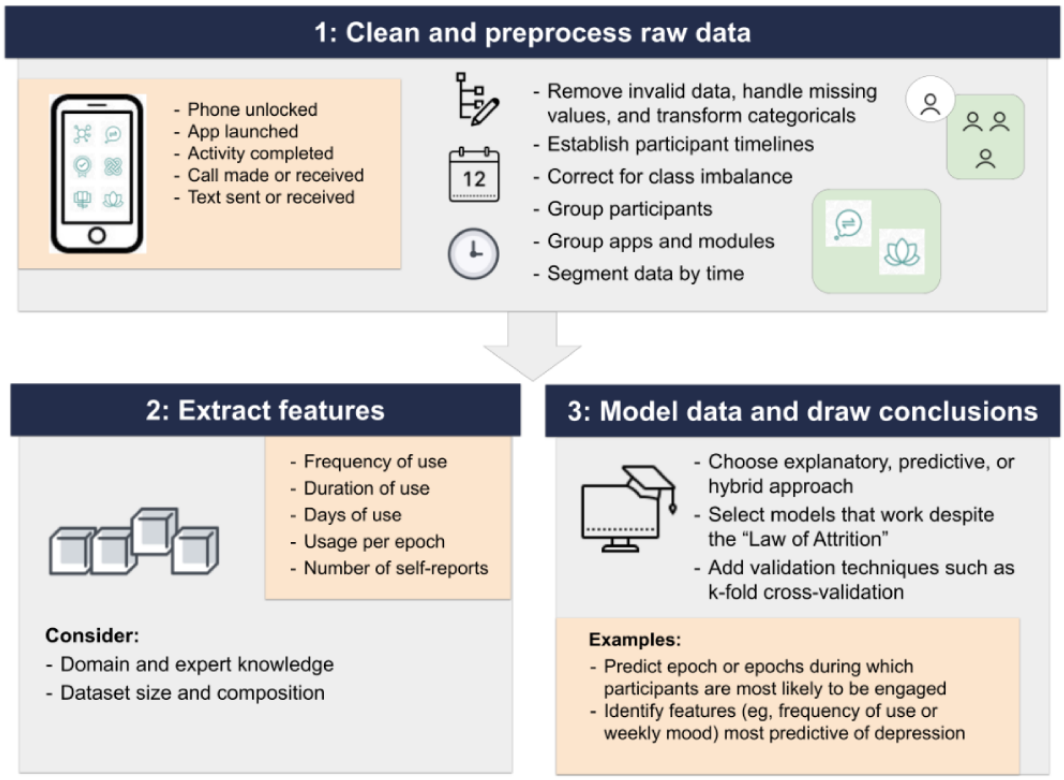
Methods

A Process to Examine the Relationship Between App Engagement and Mood in Patients With Breast Cancer

Overview

The overarching steps for understanding the dynamic relationship between engagement with mental health apps and mood among patients with breast cancer are outlined in Figure 1. Our process is informed by accepted data science techniques for extracting and analyzing features from raw data and gives special consideration to data sets that contain metrics of user engagement. This process assumes that researchers already have a data set that includes a mixture of time-stamped engagement data in addition to self-report data on mood. Mood data should include validated self-report measures administered at baseline, post intervention, and regular intervals (eg, weekly) throughout the study. Engagement data should comprise time-stamped event logs of app launches. It may also include information such as logs of phone lock or unlock events, mobile app launches, completed in-app activities, and outgoing or incoming calls and texts.

Figure 1. Proposed process for extracting and analyzing features of mood and engagement for patients with breast cancer using statistical and machine learning.



Step 1: Preprocess the Raw App Engagement Data

Overview

The first step is to preprocess the raw engagement data. Preprocessing is critical for preparing the data for analysis and includes removing invalid data, handling missing data, transforming categorical variables, normalizing all values, and correcting for class imbalance. In mHealth studies, such as those involving patients with breast cancer, preprocessing entails several additional tasks: establishing participant timelines, identifying time windows of interest, grouping participants, and grouping apps and modules.

Remove Invalid Data, Handle Missing Values, and Transform Categoricals

Invalid and missing data are common to all data sets and can occur because of user error, sensor malfunction, or lack of user action. This may be particularly relevant in the context of patients with breast cancer, given the demands and cognitive effects of treatment (eg, chemotherapy); for example, a GPS sensor may provide an inaccurate reading, or a user may complete a self-report measure on their phone but fail to click the *submit* button. Large swaths of invalid or missing data can degrade the quality of the data set and lead to less accurate analysis, making it imperative that researchers handle both with care. In mHealth studies, *invalid* data are best described as data that fall outside the acceptable range for a given variable. An example is app launches that are too short (eg, <5 seconds) or too long (eg, >5 hours) in duration. In the former case, the user opens the app and immediately closes it. In the latter case, the mobile phone sensor that monitors app use may fail to record the end of the user's use activity period for the given app. Invalid data should be removed at the very beginning of the preprocessing stage to reduce the complexity of the data set and the computing power needed to analyze it.

Missing values are data that should have been recorded but were not. Newly diagnosed patients with breast cancer often struggle with both constraints on their time and the emotional burden of managing their disease [26,27]. As a result, missing data may occur at various points in a trial, such as failure to complete all administered self-report measures. Various techniques are available to account for missing data. For variables that follow a linear pattern, interpolation can be used to impute missing values between 2 time points; that is, $y_i = (y_{i-1} + y_{i+1})/2$, where the value is missing at position *i*. Alternatively, for variables with unknown or nonlinear patterns of change, more sophisticated methods such as multiple imputations using linear regression can be used [28].

After invalid and missing data are handled, categorical values from validated instruments and other self-reports should be transformed to their numeric equivalents. Finally, all data should be scaled. As these steps are not unique to mHealth or app engagement data sets, we refer to studies by García et al [29,30] for further reading.

Establish Participants' Timelines

Next, individual time-stamped data points must be aligned to a standardized study timeline. Researchers often face challenges in recruiting patients with breast cancer to enroll in trials of digital interventions [31] and thus rely on a rolling enrollment

period to increase recruitment over time. As a result, mHealth data sets collected from patients with cancer often have different coverage periods for each patient. Therefore, researchers must convert raw time-stamps to relative time points with respect to the study length and when a participant began the study to establish a standardized timeline for analysis. Consider 2 participants, participant A and participant B. Participant A begins the study on January 1, 2021, and submits a self-report via a mobile app on January 2, 2021. Participant B begins the study later, on January 15, 2021, and submits a self-report on January 20, 2021. Despite their different start and submission dates, both participants were said to have submitted their data during the first week of the study. This is just one example of how time-stamps may be aligned, as researchers may wish to use a different temporal granularity (eg, the day of study).

Consider the Issue of Class Imbalance

For studies involving classification analyses, researchers should address the issue of *class imbalance* in the data set. Class imbalance arises when observations in a small subset of categories dominate the rest [32]. This imbalance can cause problems during the analysis phase of a study by producing classifiers that always predict the dominant class or classes. Consider a study of patients with breast cancer and a simplified binary classification problem. We want to predict whether a participant is depressed given the time and frequency of app use. If most patients are depressed at baseline, the data set is *imbalanced*, and we have an overrepresentation of users with depression. As a result, a machine learning classifier may incorrectly predict that all users are depressed, irrespective of the given data. To handle this class imbalance, researchers can take what Rout et al [33] described as a *data-level approach* and either exclude some of the data of the users with depression or draw from the nondepressed users' data to create new artificial data points. Alternatively, researchers can take an *algorithm-level* approach [33] and select a classifier that will ensure that users with depression do not skew the results. For smaller data sets, we recommend using data-level approaches such as upsampling to generate additional examples of the positive class from which an algorithm can learn. As the literature on class imbalance mitigation is broad, we refer to studies by Yap et al [34] and Rout et al [33] for more targeted reading of data- and algorithm-based techniques and strategies for selecting the most appropriate approach.

Group Participants

Researchers should next decide whether to group participants together or analyze engagement patterns for separate user groups. Methods of grouping participants can be broadly classified as either *theory-driven* or *data-driven*. *Theory-driven* grouping relies heavily on prior literature to categorize participants based on shared characteristics, such as demographics or mental health status. Recent studies that have grouped participants by mental health symptoms (eg, high vs low anxiety and depression) or personality traits (eg, high vs low extraversion) have revealed differences in both social and engagement behaviors between groups [35,36]. Importantly, studies in patients with breast cancer indicate a significant amount of heterogeneity in distress levels and trajectories, such that some patients experience very high levels of distress and mood symptoms, whereas others experience no or relatively low levels of distress throughout treatment [37]. On the basis of this literature, researchers may wish to classify their participants based on their baseline distress and mood scores to understand how these groups engage with mental health apps based on their differences.

Data-driven grouping, or *clustering*, relies on the inherent properties of a data set to identify naturally occurring groups [38]. Clustering is particularly useful for explanatory analysis of medium to large–sized novel data sets when theory-driven grouping may be infeasible. Recent research has applied clustering methods to breast cancer data sets to identify topics of conversation in breast cancer support forums [39] and investigate how depression varies according to adherence to a mood-tracking app [40]. Although outside the scope of this study, researchers seeking to conduct data-driven grouping may wish to start with 1 of the 2 common clustering methods for clinical data: *k-means clustering* or *hierarchical clustering* [41].

Group Apps and Modules

In studies that test >1 app or investigate an app containing multiple distinct modules, researchers must decide whether to analyze engagement in aggregate across all apps or separately for each individual app. Increasingly, researchers are developing suites of related apps that target a general domain of health, such as mental health, but have distinct target goals. In the IntelliCare suite [25], for instance, the Thought Challenger app helps users address negative thoughts, whereas the Daily Feats app helps users track their accomplishments and stay motivated. Women with breast cancer may benefit from multiple apps or a suite of apps, given their unique physical, emotional, and social needs tied to their disease. Multiple apps (or modules within a single app) that independently serve these different needs may be necessary to provide adequate support during treatment.

As with grouping participants, both theory- and data-driven grouping may be useful. For instance, theory-driven grouping can group apps according to health domain (eg, mental health) or subdomain (eg, depression management) or according to a cutoff score for a metric such as use frequency (eg, *highly used apps* are a group containing all apps used ≥6 days per week). Alternatively, data-driven clustering can be used to identify and group similar apps irrespective of the domain. Research should carefully consider the app intervention in question and whether to perform separate analyses for different groupings of apps or intervention components.

Segment Data by Time

Finally, researchers should consider segmenting data into meaningful windows of time or *epochs* [42]. Temporal segmentation has been used to broadly detect human activity and behavioral patterns, including facial behavior, breathing state changes [43], social behavior [35,44], and sleep disruption events [45]. Previous works within mHealth, specifically, have used theory-driven temporal segmentation to examine engagement at hourly intervals, across multihour spans (eg, *morning*, spanning 6 AM to 11:59 AM), and at weekly intervals [35,36,42,46].

When segmenting data into epochs, researchers should weigh the nature of the condition being studied and, in turn, the timescale or timescales along which symptoms and behaviors are likely to vary. Women newly diagnosed with breast cancer may only have sporadic pockets of time throughout the day to engage with a mental health app because of increased time spent attending physician's appointments and managing their illness and sequelae of related factors. In addition, because of the disruptive impact of anxiety, depression, and cancer treatment on daily rhythms [47], patients with breast cancer experiencing mental health challenges may engage with mental health apps at irregular times. Given the stressors that patients with breast cancer face, short and frequent time windows (eg, hours or days) may be most appropriate to capture fluctuations in mood or identify the times at which a participant is most receptive to an intervention.

When segmenting their data, researchers are encouraged to balance temporal granularity against data set size. Larger data sets with more frequent measurements naturally allow for more granular epochs (eg, hourly). Researchers should also take care to ensure that epochs are neither too broad nor too narrow. Epochs that are too broad will fail to capture meaningful patterns, whereas epochs that are too narrow will introduce sparsity into the data set and decrease the effectiveness of the analysis.

Step 2: Extract Engagement Features

After preprocessing and before conducting machine learning classification tasks, researchers must identify the most salient variables (called *features*) within the data set and, when necessary, combine measures into new variables. This process is known as *feature extraction* and should be guided by several key factors, including domain knowledge and the size and overall composition of the data set. Importantly, researchers should avoid creating large, sparse feature sets (FSs), as this can lead to overfitting during the modeling and prediction phases. Feature extraction in small-to-medium–sized data sets, such as those of mood and app engagement, can reasonably be conducted by hand with sufficient knowledge of prior literature and the domain of interest. However, researchers interested in automated methods for high-dimensional data may find tools such as autoencoders useful [48].

Traditionally, researchers have measured engagement with blunt usage metrics such as the total or mean number of app sessions over the course of an intervention or the number of users that fail to complete an intervention [21]. However, with the increasing ubiquity of sensor-equipped smart devices, researchers have been able to derive more granular features of engagement from logs of phone or app use [49]. Several important features have emerged from recent studies, including the frequency of use (eg, number of times per week), number of days of use, duration of use, whether any use occurred in a given period, and the number of self-reports submitted [42,46,50,51]. To summarize these and other *analytic indicators of engagement*, we refer to a study by Pham et al [52].

Step 3: Model Data and Make Predictions

After preprocessing the data and constructing an appropriate set of features, the final step is to model and make predictions using the newly generated features. Several decisions must be made in this step. First, researchers must decide whether an explanatory, predictive, or combined modeling approach is appropriate; that is, whether the goal is to simply identify relationships between measures of engagement and mental health status or to predict one measure from another. Next, researchers must select an appropriate set of models, considering factors such as the overall data set size and structure. mHealth studies are known to have high dropout rates [21], leading to small and sparse data sets. Therefore, it is essential to select modeling techniques that can handle small data sets with a high proportion of missing or imputed data with a reasonable degree of accuracy. Finally, researchers should ensure that modeling and prediction tasks include techniques such as *cross-validation* and *parameter tuning*. Cross-validation is a technique in which random subsets of data (often multiple times) are selected as training and testing sets, which are then used to evaluate the reliability of a machine learning model [53,54]. Meanwhile, parameter tuning is the process of adjusting the model parameters to achieve better model performance metrics (eg, better accuracy and precision) [55]. Both techniques are crucial for ensuring that a machine learning model is well-constructed.

Case Study

Overview

To illustrate the app engagement process, data were extracted from a 7-week trial [56] of a mobile mental health app suite among women newly diagnosed with breast cancer (N=40 participants). IntelliCare is a collection of apps that use an elemental, skills-based approach to improving mental health. In-app exercises are meant to be intuitive, requiring few instructions to complete, and most of these exercises can be found on the first screen presented by the app. Participants used their own personal phones and were recruited from a breast care clinic at a US National Cancer Institute–designated clinical cancer center. A detailed description of the recruitment method, as well as the goals of the IntelliCare apps, can be found in a paper that depicts the primary outcomes of the study [56]. Participants downloaded and tried 1 to 2 apps each week. All participants received light phone coaching that focused on addressing usability issues with the apps, which included an initial 30-minute call at the beginning of the trial, followed by a 10-minute call 3 weeks into the trial. Although 58% (23/40) of participants completed the intervention in the original trial, because of technical issues exporting app use metrics from the system, detailed app engagement data were only available for 35% (14/40) of participants.

Ethics Approval

This study was approved by the institutional review board at the University of Virginia (UVA IRB-HSR#20403).

Participant Demographics

Participants had a mean age of 56.8 (SD 11.6) years; 82% (31/38) of participants who indicated their race were White, 11% (4/38) were Black, 3% (1/38) were Hispanic, 3% (1/38)

were American Indian or Alaska Native, and 3% (1/38) were multiracial.

Measures

The Patient Health Questionnaire-4 (PHQ-4) [57] and Patient-Reported Outcomes Measurement Information System-29 (PROMIS-29) [58] were used to assess the symptoms of depression and anxiety at baseline and after the intervention. To allow for an examination of changes in mood symptoms over the course of the trial, a 2-item measure of symptoms of anxiety and depression was administered once daily during week 1 and at the beginning of weeks 2 to 6 of the trial. The daily measures from week 1 were averaged. This measure comprised questions from the PHQ-4 ("How much did you feel nervous, anxious, or on edge?" and "How much interest or pleasure did you have in doing things?"). Both items were scored on a 5-item Likert scale (1=not at all, 2=a little, 3=somewhat, 4=quite a bit, and 5=a lot or extremely).

Weekly self-reported measures of well-being were also collected. The questions covered topics such as substance use, physical pain, connectedness to others, reception and giving of social support, general activity, and management of negative feelings. Items were scored on a 5-item Likert scale that matched the scale for the PHQ-4 and PROMIS-29 Anxiety (1=not at all, 2=a little, 3=somewhat, 4=quite a bit, and 5=a lot or extremely).

App use data were collected using the IntelliCare platform. These data contained 1 time-stamped entry per participant per app launch. Each entry included information such as the name of the app used and the launch duration in milliseconds.

Missingness

The rate of missing data was 39.6% among all participants (including those who dropped out at any point during the study); this rate is consistent with the often-high dropout rates in mHealth studies [21]. Among patients who completed the baseline survey, the missingness rate was 10%. Only patients who completed the baseline survey and used at least one mobile app in the IntelliCare suite were included in our final analysis (14/40, 35%).

Data Preprocessing and Feature Extraction

We selected 2 time windows for our analysis: the entire 7-week study lifetime and 1-week intervals (eg, week 1 and week 2). Given our overarching goal of examining the interplay between mood and engagement, we selected a theory-driven approach for grouping participants based on a wealth of literature showing that patients with breast cancer vary with regard to their distress levels and trajectory over the course of treatment. Thus, we grouped participants according to their baseline depression and anxiety symptoms and weekly mood [35,36]. For symptoms of anxiety and depression, we segmented users into *high* and *low* groups according to their baseline scores. Cutoff values for determining group placement were identified using the PHQ-4 and PROMIS-29 scoring guidelines. Users who scored ≥3 on the PHQ-4 Anxiety subscale or who scored ≥60 on the PROMIS-29 Anxiety subscale were placed in the *anxious* group, whereas the rest were placed in the group with *low anxiety*. Similarly, users who scored ≥3 on the PHQ-4 Depression subscale or who scored ≥60 on the PROMIS-29 Depression subscale were placed in the group with *high depression*, whereas the rest were placed in the group with *low depression*.

Labeling of weekly mood was conducted in a manner similar to the labeling of depression and anxiety levels at baseline. Participants with scores of ≥4 for weekly anxious mood were labeled *anxious*, and participants with scores of ≤2 for weekly depressed mood were labeled *depressed*. We note that the cutoff score for depression was applied in the inverse direction because of the nature of the question, "How much interest or pleasure did you have in doing things?"; that is, replying 1=not at all or 2=a little indicates a depressed mood.

We conducted feature extraction by hand using domain knowledge and adapting approaches from related studies. Notably, we closely followed the approach of Cheung et al [46] to quantify the metrics of engagement from logs of app use data. For instance, to calculate frequency, we grouped raw app use logs by participant and period (eg, week) and calculated the number of times the app was used during that period. We extracted 3 main measures of engagement from the raw app use data: *frequency* (number of launches), *days of use*, and *duration of use*. Variants of these measures (eg, mean frequency and duration between launches) were also included in our analysis. Table 1 provides an overview of each of the 5 FSs used in the analysis.

Table 1. Feature sets (FSs) used in the analysis.

FS	Description	Example features
FS1	Engagement features for all apps	Frequency of use for all apps combined, days of use, duration of use, and mean duration of use
FS2	Engagement features for only the most frequently used app or apps	Frequency of use for the app "Worry Knot" and days of use for the app "Thought Challenger"
FS3	Self-report features+engagement features for all apps	PROMIS[a] social support score, frequency of use for all apps combined, and days of use
FS4	Self-report features+engagement features for only the most-used app or apps	PROMIS social support score, duration of use for the apps "Thought Challenger" and "Worry Knot"
FS5	Self-report features+engagement features for each individual app	PROMIS physical pain score, frequency of use for the app "Worry Knot," and days of use for the app "Daily Feats"

[a]PROMIS: Patient-Reported Outcomes Measurement Information System.

To prepare the data for both the regression and classification tasks, we conducted multiple imputations [28] to handle missing values in self-reported measures. Class imbalance in the classification tasks was handled using the Synthetic Minority Oversampling Technique (SMOTE) [59], a technique that synthesizes new samples from the minority class feature space.

Modeling and Prediction

Explanatory Analysis of Engagement Across Baseline Affect Groups

For each measure of depression and anxiety, we graphically analyzed the distributions of engagement measures at weekly intervals for both the *low* and *high* groups. Given the size of our data set, we analyzed engagement across all apps rather than by individual or groups of apps to avoid bias because of sparsity. Furthermore, the IntelliCare apps are conceptualized as belonging to the same intervention, and individual apps target related areas of mental health. Graphical analysis revealed notable differences in engagement between the groups with low and high anxiety and between the groups with low and high depression.

Correlation Analysis of App Engagement and Weekly Mood

To study the correlations between app engagement metrics and weekly mood, we fit linear mixed models to account for the repeated measures within each participant, using the subject as a random effect (ie, random intercepts) and different app engagement FSs as fixed effects. Specifically, we fit linear mixed-effects models with the least absolute shrinkage and selection operator with tuned penalty parameter α and weekly anxious mood as the outcome variable on 4 FSs from Table 1 and repeated this process using weekly depressed mood as the outcome variable. Self-reported features were used as control variables.

Predictive Modeling of Weekly Mood

We wanted to investigate whether engagement with mobile apps can be used to predict weekly anxious and depressed moods, as specified in our process. We considered the case of depressed mood and formulated a binary prediction problem as follows: given a vector of a participant's app use activity and survey scores for a given week, we predicted whether the participant was depressed (1) or not depressed (0).

Binary prediction problems are well-handled by tree-based classifiers. These classifiers make decisions by *splitting* into one of several paths at each decision point or *node*. Thus, possible decision paths that can be taken to reach the final prediction are akin to the branches in a tree, with possible final predictions akin to the leaves. Tree-based models are known for their inherent feature selection capabilities and robustness to small sample sizes, which makes them a good fit for our analysis. We selected 2 popular tree-based classifiers, XGBoost (XGB) and random forest (RF), and ran these with leave-one-subject-out cross-validation (LOSOCV) to predict weekly anxious mood and weekly depressed mood separately on the FS3, FS5, and FS4 FSs. LOSOCV is a variant of k-fold cross-validation, a standard technique for evaluating a model's performance, in which the entire data set is randomly split into *k* subsets. A subset was held out for testing, whereas the rest were combined to train the model, and the process was repeated for all *k* subsets. In the same vein, LOSOCV divides the data into subsets based on subjects and follows the k-fold cross-validation process.

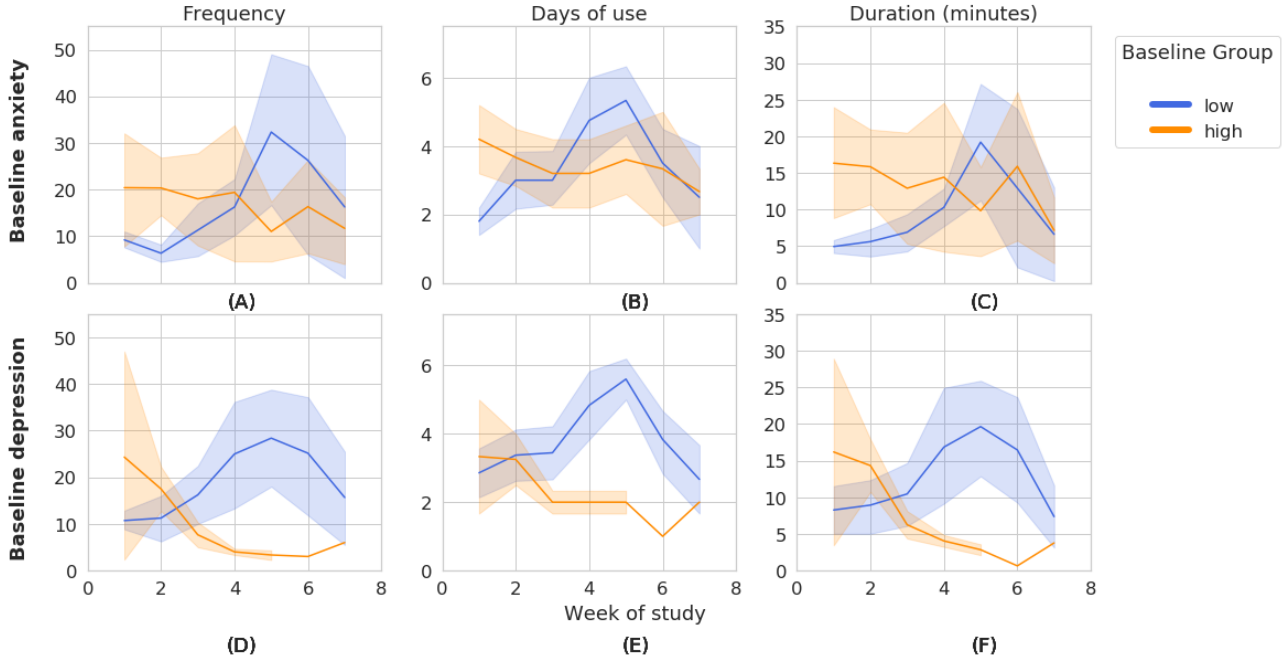
The model hyperparameters were tuned using *gridsearch*, which attempts many combinations of different hyperparameters to find the *optimal* combination (ie, the combination that produces a model with the best performance). In our case, we paired *gridsearch* with a variant of k-fold cross-validation called *stratified group k-fold cross-validation*. This technique is similar to LOSOCV in that it prevents data leakage by ensuring that no subject from the training set also appears in the testing set. It also has the additional benefit of creating stratified splits, such that the balance of positive and negative class labels (1 and 0 seconds) is roughly the same in the training set as in the testing set. This approach, similar to the SMOTE, helps mitigate the effects of class imbalance in smaller data sets.

Results

Explanatory Analysis of Engagement Across Baseline Affect Groups

Both the participant groups with high anxiety and high depression experienced decreases in all 3 engagement measures between week 1 and week 7, as shown in Figure 2. Notably, the groups with high anxiety and high depression started at week 1 with higher group means than their respective *low* group counterpoints but slowly declined across measures over time. In contrast, users with low anxiety and low depression saw gradual rises across all measures, with a sharp peak around weeks 5 to 6, followed by a subsequent decrease. Interestingly, participants with low anxiety and low depression ended the study at week 7 with approximately the same group means as their respective *high* group peers.

Figure 2. Comparison of weekly engagement metric means (with 68% CI) between 8 participants with low anxiety and 6 participants with high anxiety (A-C) and between 10 participants with low depression and 4 participants with high depression (D-F).



Correlation Analysis of App Engagement and Weekly Mood

The correlation analysis results are shown in Table 2. Several features of engagement provided significant correlations with weekly mood at $P<.05$. When engagement features for all apps were used (FS1), anxiety negatively correlated with the minimum duration (−0.0459). When features of only the most-used apps were used (FS2), depression negatively correlated with the week of study (−0.1826) and frequency (−0.1304) and positively correlated with days of use (0.4565), minimum duration (0.0414), and maximum duration (0.0248). The results for FSs FS3 and FS4 show that the inclusion of self-reported features as control variables improves model fit (indicated by root mean square error). When both self-report and engagement features for all apps were used (FS3), depression negatively correlated with frequency (−0.086), mean duration (−0.0637), and maximum duration (−0.0215) and positively correlated with total duration (0.0024), duration SD (0.098), and minimum duration (0.0978). Finally, when both self-report and engagement features for only the most-used apps were used (FS4), depression positively correlated with the minimum duration (0.0917) and maximum duration (0.0386). Interestingly, no significant correlations were observed between the selected app use features on weekly self-reported anxiety levels for FSs FS2, FS3, and FS4. We caution against overinterpreting this finding, given the limited sample size; rather, these results demonstrate the feasibility of identifying correlates with mood from heterogeneous data sets of engagement.

Table 2. Linear mixed model results stratified by feature set (FS) and outcome variable.

Outcome variable	FS1,[a] coefficient (*P* value)		FS2,[b] coefficient (*P* value)		FS3,[c] coefficient (*P* value)		FS4,[d] coefficient (*P* value)	
	Anxiety	Depression	Anxiety	Depression	Anxiety	Depression	Anxiety	Depression
Week of study	0 (—[e])	−0.16 (.14)	−0.0063 (.93)	−0.1826 (<.001)[f]	0.1122 (.22)	0.0659 (.62)	0.0643 (.43)	0.1803 (—)
Frequency	−0.0169 (.55)	−0.0632 (.14)	−0.0976 (.09)	−0.1304 (.004)[f]	−0.0438 (.12)	−0.086 (.004)[f]	−0.1747 (.001)	−0.5962 (—)
Days of use	0.0761 (.53)	−0.0737 (.74)	0.1757 (.08)	0.4565 (<.001)[f]	0.1047 (.38)	0.2374 (.25)	0.2909 (.02)	1.5607 (—)
Total duration	0.0003 (.67)	0.0021 (.12)	0.0011 (.63)	−0.0017 (.17)	0.0009 (.24)	0.0024 (.01)[f]	0.0026 (.24)	0.0009 (.68)
Mean duration	0.0237 (.17)	−0.027 (.24)	0.0071 (.78)	−0.0336 (.12)	0.0007 (.97)	−0.0637 (.03)[f]	−0.0092 (.66)	−0.1536 (—)
Duration SD	−0.0172 (.36)	0.0354 (.45)	0.0055 (.83)	−0.0093 (.66)	−0.0002 (.99)	0.098 (.02)[f]	0.0026 (.91)	0.0901 (—)
Minimum duration	−0.0459 (.02)[f]	0.032 (.37)	−0.0171 (.52)	0.0414 (.03)[f]	−0.0269 (.21)	0.0978 (.01)[f]	−0.0083 (.75)	0.0917 (<.001)[f]
Maximum duration	0.0007 (.92)	−0.0105 (.44)	−0.0047 (.70)	0.0248 (<0.001)[f]	0.0004 (.95)	−0.0215 (.05)[f]	−0.0006 (.96)	0.0386 (<.001)[f]

[a]FS1: anxiety: α=.1, root mean square error 0.7396; depression: α=.1, root mean square error 0.7589.

[b]FS2: anxiety: α=.7, root mean square error 0.8095; depression: α=.1, root mean square error 1.3954.

[c]FS3: anxiety: α=.1, root mean square error 0.5128; depression: α=.1, root mean square error 0.4136.

[d]FS4: anxiety: α=.1, root mean square error 0.5348; depression: α =.1, root mean square error 0.4547.

[e]*P* value was not defined.

[f]Effects with a *P* of <.05.

Predictive Modeling of Weekly Mood

The predictive modeling results are shown in Table 3 below. FS3, which contained survey features and overall app engagement features, achieved the highest predictive accuracy (84.6%) and yielded the best outcome measures when used with an RF classifier to predict depressed mood. FS4, which contained survey features and engagement features only from the most-used apps, achieved the second-best predictive accuracy (81.5%) when used with an XGB classifier. FS5 yielded the worst results overall, likely because of a combination of overfitting and a lack of meaningful information contained in engagement features for individual apps. Overfitting is a common issue for tree-based models applied to small data sets and occurs when the model learns the training set so well that it poorly generalizes when making predictions on the test set. We note that despite using techniques such as the SMOTE and LOSOCV, which are designed to reduce overfitting, we still struggled to mitigate this issue in our predictive task. Further investigation is warranted to determine whether a larger data set might yield better predictive results.

A feature importance graph of Shapley Additive Explanations (SHAP) scores [60] for the top classifier and FS (ie, RF/FS3) for depressed mood prediction is shown in Figure 3. Self-report features such as connectedness to others (feature *Connectedness*) and receiving support from others (feature *Receive support*) were particularly important. Engagement features such as frequency and the mean duration of use were also important. As with the results of our correlation analysis, we caution against overinterpretation of the importance of individual features, given the limited sample size.

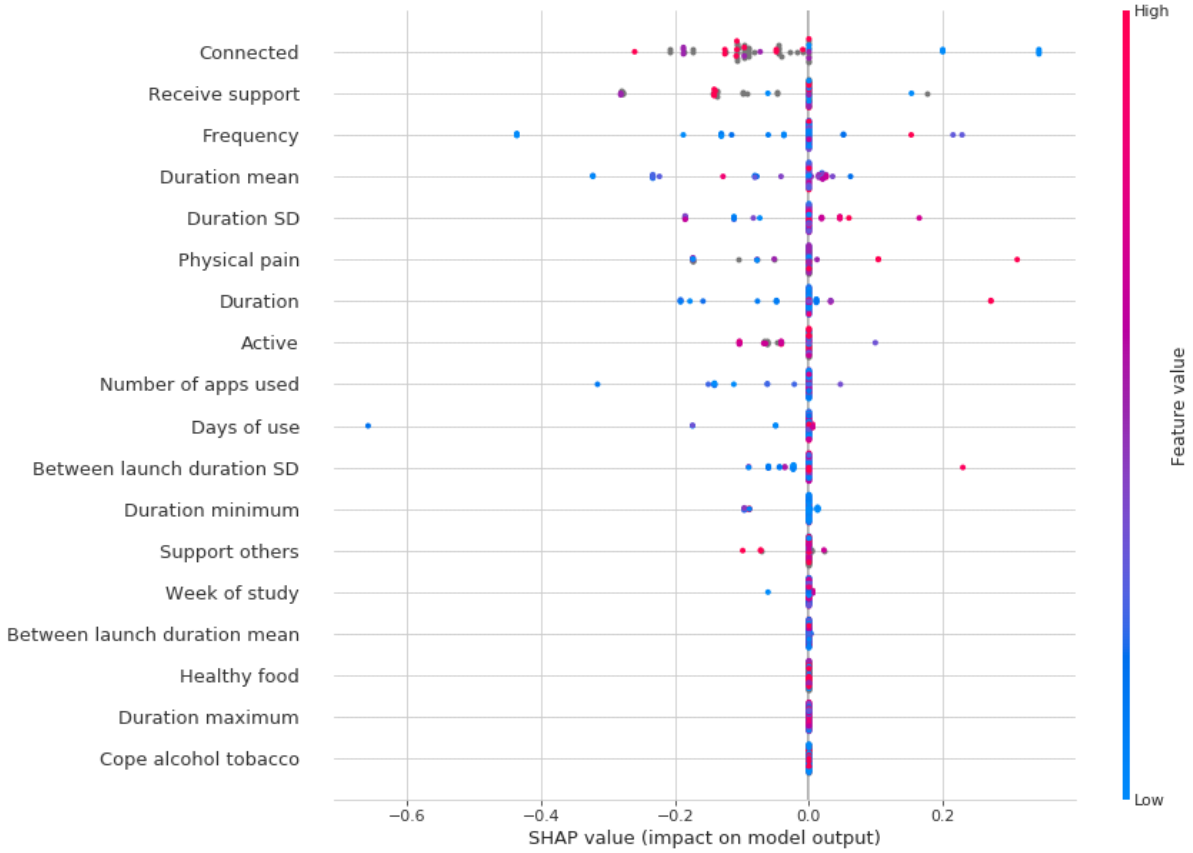
The findings from these exploratory analyses indicate that it may be feasible to identify the weekly moods of patients with breast cancer based on their app use metrics.

Table 3. Weekly depressed mood prediction task results.

Classifier and FS[a]	Accuracy, %	Precision, %	Recall, %	F1, %
Random forest				
FS3	84.61	82.50	64.42	67.75
FS4	83.07	73.50	72.11	72.76
FS5	66.15	50.00	50.00	49.93
XGBoost				
FS3	78.46	67.33	69.23	68.13
FS4	81.53	70.81	62.50	64.54
FS5	67.69	47.95	48.07	48.00

[a]FS: feature set.

Figure 3. Feature importance for the prediction of depressed mood using a random forest classifier on feature set 3. SHAP: Shapley Additive Explanations.



Discussion

Principal Findings

Considering the increased sophistication of mobile devices and app-delivered interventions that can capture minute details of user engagement, there is a need to develop increasingly sophisticated frameworks to make sense of user engagement data. In this study, we proposed a process for understanding the dynamic association between app engagement and mood using machine learning. Importantly, how engagement data are processed differs from study to study. The studies by Cheung et al [46] and Pham et al [52] drew attention to these diverse data-processing approaches and the common features that characterize engagement. Our process attempts to unify the key aspects of these approaches and refocus them on data collected from patients with breast cancer. The application of the proposed process and evaluation of statistical models support the feasibility of predicting mood status based on app engagement. The analyses and results from the case study are meant to demonstrate the potential of this approach; therefore, we caution readers not to overstate the findings of our case study. Replication of the findings in a larger data set is needed to draw more firm and generalizable conclusions.

With this caveat, the application of our process to the case study data yielded some interesting preliminary findings that may be worth pursuing in future studies. The most prominent models

and theories of behavioral change highlight the importance of motivational forces to sustain a behavior [61-63], such as engagement in a mental health app. Individuals with high levels of depression or anxiety symptoms are likely to experience low self-efficacy or a low perceived ability to perform a behavior, which is likely to result in poor engagement. Our results suggest that baseline levels of anxiety and depression affect patterns of engagement among patients with breast cancer, at least in the short term. The findings for the groups with high anxiety and high depression suggest that strong initial engagement does not necessarily lead to long-term engagement *growth*. In addition, the findings for the groups with low anxiety and low depression suggest that engagement may be difficult to sustain in the long term and may reach a point of diminishing returns.

The application of our process that led to the predictive results is promising in that both the RF and XGB classifiers performed well (>60% for all metrics) even with moderate amounts of data when the FS was well-curated (ie, when *FS4* and *FS3* were used). This suggests that heterogeneous FSs comprising both baseline mental health measures and engagement data may be useful for predicting weekly moods when analyzed with robust classifiers. Predictions of weekly mood can, in theory, be used to personalize interventions. A dose-response relationship has been observed in digital health interventions, making it especially important to target patients when they are most open to receiving a dose of an app-delivered intervention. Heterogeneous data sets, along with high-accuracy classifiers, could be used within a just-in-time adaptive intervention (JITAI) [64] to predict the mood of patients with breast cancer. This mood could then be cross-referenced with the patient's schedule to identify the optimal time window for intervention delivery. Studies have also demonstrated that distress tends to spike in women around the time they receive an initial diagnosis [65,66] but that a patient's needs change throughout the course of treatment [67-69]. Such a just-in-time adaptive intervention could be further extended to learn the mood and engagement patterns of a patient with breast cancer over time and adjust the timing of the intervention accordingly. Further research is needed to determine the feasibility of implementing such interventions in vulnerable populations.

Prior studies examining the link between engagement with mHealth tools and symptoms have historically yielded mixed results; some studies have identified a direct relationship [35,70], whereas others have identified an inverse relationship [63,71]. Although we cannot definitively quantify this relationship in our study, both our correlation and predictive analyses suggest that paring down the available features to include only the most relevant engagement data for each individual (eg, features from only the most-used apps) and combining self-report data with passively monitored engagement data may help researchers better identify significant predictors of mood.

Limitations

There are several limitations to this study that should be considered in light of these results. The results from the case study are limited in generalizability because of the small sample size. Data sparsity was a particular challenge when we attempted to break down our time windows of interest into smaller epochs, such as 4-hour windows describing different periods of the day (eg, *morning* and late night); therefore, we had to focus on daily and weekly time windows. Similar issues with sparsity occurred when we attempted to analyze the data for each individual app in the IntelliCare suite. Furthermore, our prediction task experienced overfitting. We recommend that researchers focus particularly on recruitment and retention for similar future studies to ensure that the resultant data set is sufficiently large for granular analyses.

Our study is also limited in scope as we did not account for demographic covariates, such as age, race, or socioeconomic status, in our mixed-effects model. As demographic factors are known to play an impactful role in health outcomes, we encourage researchers to include these factors in future studies on engagement with health apps. Finally, this study focused only on patients with breast cancer; therefore, our results may not be generalizable to other patient populations with cancer or other diseases.

Conclusions

Inspired by existing work, this study introduces a step-by-step process for investigating the relationship between mood and mobile app engagement among patients with breast cancer. We believe our process has important implications for the study of mobile app engagement among patients with breast cancer and for the study of engagement more broadly, given its flexibility and ability to handle large and dense data sets. The results from the case study suggest a need to better tailor interventions according to the baseline symptoms of depression and anxiety of patients with breast cancer. The findings from the case study also support a wider call within the field of digital interventions to advance the understanding of user engagement and attrition to sustain long-term engagement and, hence, more robust outcomes.

Acknowledgments

This work was supported by a grant from the University of Virginia Center for Engineering in Medicine, as well as efforts supported by the University of Virginia Cancer Center. The writing of this manuscript was supported by the National Cancer Institute (R37 CA248434 to PC) and the National Institutes of Health (T32 Training Grant in Biomedical Data Sciences [project number: 5T32LM012416-05] to IP).

Conflicts of Interest

None declared.

References

1. Breast cancer facts and statistics. Breast Cancer. 2021 Feb 4. URL: https://www.breastcancer.org/symptoms/understand_bc/statistics [accessed 2021-05-25]
2. Breast Cancer Statistics. Centers for Disease Control and Prevention. 2021 Apr 27. URL: https://www.cdc.gov/cancer/breast/statistics/index.htm [accessed 2021-05-25]
3. Weaver KE, Forsythe LP, Reeve BB, Alfano CM, Rodriguez JL, Sabatino SA, et al. Mental and physical health-related quality of life among U.S. cancer survivors: population estimates from the 2010 National Health Interview Survey. Cancer Epidemiol Biomarkers Prev 2012 Nov;21(11):2108-2117 [FREE Full text] [doi: 10.1158/1055-9965.EPI-12-0740] [Medline: 23112268]
4. Linden W, Vodermaier A, Mackenzie R, Greig D. Anxiety and depression after cancer diagnosis: prevalence rates by cancer type, gender, and age. J Affect Disord 2012 Dec 10;141(2-3):343-351. [doi: 10.1016/j.jad.2012.03.025] [Medline: 22727334]
5. Reich M, Lesur A, Perdrizet-Chevallier C. Depression, quality of life and breast cancer: a review of the literature. Breast Cancer Res Treat 2008 Jul;110(1):9-17. [doi: 10.1007/s10549-007-9706-5] [Medline: 17674188]
6. Watson M, Haviland JS, Greer S, Davidson J, Bliss JM. Influence of psychological response on survival in breast cancer: a population-based cohort study. Lancet 1999 Oct 16;354(9187):1331-1336. [doi: 10.1016/s0140-6736(98)11392-2] [Medline: 10533861]
7. Carlson LE, Bultz BD. Efficacy and medical cost offset of psychosocial interventions in cancer care: making the case for economic analyses. Psychooncology 2004 Dec;13(12):837-856. [doi: 10.1002/pon.832] [Medline: 15578622]
8. Gudenkauf LM, Antoni MH, Stagl JM, Lechner SC, Jutagir DR, Bouchard LC, et al. Brief cognitive-behavioral and relaxation training interventions for breast cancer: a randomized controlled trial. J Consult Clin Psychol 2015 Aug;83(4):677-688 [FREE Full text] [doi: 10.1037/ccp0000020] [Medline: 25939017]
9. Johnson JA, Rash JA, Campbell TS, Savard J, Gehrman PR, Perlis M, et al. A systematic review and meta-analysis of randomized controlled trials of cognitive behavior therapy for insomnia (CBT-I) in cancer survivors. Sleep Med Rev 2016 Jun;27:20-28. [doi: 10.1016/j.smrv.2015.07.001] [Medline: 26434673]
10. Chi M. The hidden cost of cancer: helping clients cope with financial toxicity. Clin Soc Work J 2019;47(3):249-257. [doi: 10.1007/s10615-017-0640-7]
11. Yabroff KR, Davis WW, Lamont EB, Fahey A, Topor M, Brown ML, et al. Patient time costs associated with cancer care. J Natl Cancer Inst 2007 Jan 03;99(1):14-23. [doi: 10.1093/jnci/djk001] [Medline: 17202109]
12. Holland JC, Kelly BJ, Weinberger MI. Why psychosocial care is difficult to integrate into routine cancer care: stigma is the elephant in the room. J Natl Compr Canc Netw 2010 Apr;8(4):362-366. [doi: 10.6004/jnccn.2010.0028] [Medline: 20410331]
13. Charlton M, Schlichting J, Chioreso C, Ward M, Vikas P. Challenges of rural cancer care in the United States. Oncology (Williston Park) 2015 Sep;29(9):633-640 [FREE Full text] [Medline: 26384798]
14. Martínez Arroyo O, Andreu Vaíllo Y, Martínez López P, Galdón Garrido MJ. Emotional distress and unmet supportive care needs in survivors of breast cancer beyond the end of primary treatment. Support Care Cancer 2019 Mar;27(3):1049-1057. [doi: 10.1007/s00520-018-4394-8] [Medline: 30094729]
15. Muñoz RF, Chavira DA, Himle JA, Koerner K, Muroff J, Reynolds J, et al. Digital apothecaries: a vision for making health care interventions accessible worldwide. Mhealth 2018 Jun 4;4:18 [FREE Full text] [doi: 10.21037/mhealth.2018.05.04] [Medline: 30050914]
16. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298 [FREE Full text] [doi: 10.1002/wps.20472] [Medline: 28941113]
17. Andersson G, Cuijpers P, Carlbring P, Riper H, Hedman E. Guided Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: a systematic review and meta-analysis. World Psychiatry 2014 Oct;13(3):288-295 [FREE Full text] [doi: 10.1002/wps.20151] [Medline: 25273302]
18. Reger MA, Gahm GA. A meta-analysis of the effects of Internet- and computer-based cognitive-behavioral treatments for anxiety. J Clin Psychol 2009 Jan;65(1):53-75. [doi: 10.1002/jclp.20536] [Medline: 19051274]
19. Sucala M, Cuijpers P, Muench F, Cardo R, Soflau R, Dobrean A, et al. Anxiety: there is an app for that. A systematic review of anxiety apps. Depress Anxiety 2017 Jun;34(6):518-525. [doi: 10.1002/da.22654] [Medline: 28504859]
20. Wasil AR, Venturo-Conerly KE, Shingleton RM, Weisz JR. A review of popular smartphone apps for depression and anxiety: assessing the inclusion of evidence-based content. Behav Res Ther 2019 Dec;123:103498. [doi: 10.1016/j.brat.2019.103498] [Medline: 31707224]
21. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
22. Mattila E, Lappalainen R, Välkkynen P, Sairanen E, Lappalainen P, Karhunen L, et al. Usage and dose response of a mobile acceptance and commitment therapy app: secondary analysis of the intervention arm of a randomized controlled trial. JMIR Mhealth Uhealth 2016 Jul 28;4(3):e90 [FREE Full text] [doi: 10.2196/mhealth.5241] [Medline: 27468653]

23. Zhang R, Nicholas J, Knapp AA, Graham AK, Gray E, Kwasny MJ, et al. Clinically meaningful use of mental health apps and its effects on depression: mixed methods study. J Med Internet Res 2019 Dec 20;21(12):e15644 [FREE Full text] [doi: 10.2196/15644] [Medline: 31859682]
24. Low CA. Harnessing consumer smartphone and wearable sensors for clinical cancer research. NPJ Digit Med 2020 Oct 27;3:140 [FREE Full text] [doi: 10.1038/s41746-020-00351-x] [Medline: 33134557]
25. Mohr DC, Tomasino KN, Lattie EG, Palac HL, Kwasny MJ, Weingardt K, et al. IntelliCare: an eclectic, skills-based app suite for the treatment of depression and anxiety. J Med Internet Res 2017 Jan 05;19(1):e10 [FREE Full text] [doi: 10.2196/jmir.6645] [Medline: 28057609]
26. Hamer J, McDonald R, Zhang L, Verma S, Leahey A, Ecclestone C, et al. Quality of life (QOL) and symptom burden (SB) in patients with breast cancer. Support Care Cancer 2017 Feb;25(2):409-419. [doi: 10.1007/s00520-016-3417-6] [Medline: 27696078]
27. Kenne Sarenmalm E, Browall M, Gaston-Johansson F. Symptom burden clusters: a challenge for targeted symptom management. A longitudinal study examining symptom burden clusters in breast cancer. J Pain Symptom Manage 2014 Apr;47(4):731-741 [FREE Full text] [doi: 10.1016/j.jpainsymman.2013.05.012] [Medline: 23916827]
28. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Soft 2011;45(3):1-67. [doi: 10.18637/jss.v045.i03]
29. García S, Luengo J, Herrera F. Data Preprocessing in Data Mining. Cham, Switzerland: Springer International Publishing; 2015:59-105.
30. García S, Ramírez-Gallego S, Luengo J, Benítez JM, Herrera F. Big data preprocessing: methods and prospects. Big Data Anal 2016 Nov 1;1(1):9. [doi: 10.1186/s41044-016-0014-0]
31. Gupta A, Ocker G, Chow PI. Recruiting breast cancer patients for mHealth research: obstacles to clinic-based recruitment for a mobile phone app intervention study. Clin Trials 2020 Dec;17(6):675-683. [doi: 10.1177/1740774520939247] [Medline: 32660354]
32. Sun Y, Wong AK, Kamel MS. Classification of imbalanced data: a review. Int J Patt Recogn Artif Intell 2009;23(04):687-719. [doi: 10.1142/S0218001409007326]
33. Rout N, Mishra D, Mallick MK. Handling imbalanced data: a survey. In: Reddy MS, Viswanath K, KM SP, editors. International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications: ASISA 2016. Singapore, Singapore: Springer; 2018:431-443.
34. Yap BW, Rani KA, Rahman HA, Fong S, Khairudin Z, Abdullah NN. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In: Proceedings of the 1st International Conference on Advanced Data and Information Engineering. 2013 Presented at: DaEng '13; December 16-18, 2013; Kuala Lumpur, Malaysia p. 13-22. [doi: 10.1007/978-981-4585-18-7_2]
35. Saeb S, Lattie EG, Kording KP, Mohr DC. Mobile phone detection of semantic location and its relationship to depression and anxiety. JMIR Mhealth Uhealth 2017 Aug 10;5(8):e112 [FREE Full text] [doi: 10.2196/mhealth.7297] [Medline: 28798010]
36. Mendu S, Baglione A, Baee S, Wu C, Ng B, Shaked A, et al. A framework for understanding the relationship between social media discourse and mental health. Proc ACM Hum-Comput Interact 2020 Oct 14;4(CSCW2):1-23. [doi: 10.1145/3415215]
37. Henselmans I, Helgeson VS, Seltman H, de Vries J, Sanderman R, Ranchor AV. Identification and prediction of distress trajectories in the first year after a breast cancer diagnosis. Health Psychol 2010 Mar;29(2):160-168. [doi: 10.1037/a0017806] [Medline: 20230089]
38. Peach RL, Yaliraki SN, Lefevre D, Barahona M. Data-driven unsupervised clustering of online learner behaviour. NPJ Sci Learn 2019 Sep 3;4:14 [FREE Full text] [doi: 10.1038/s41539-019-0054-0] [Medline: 31508242]
39. Jones J, Pradhan M, Hosseini M, Kulanthaivel A, Hosseini M. Novel approach to cluster patient-generated data into actionable topics: case study of a web-based breast cancer forum. JMIR Med Inform 2018 Nov 29;6(4):e45 [FREE Full text] [doi: 10.2196/medinform.9162] [Medline: 30497991]
40. Kim J, Lim S, Min YH, Shin YW, Lee B, Sohn G, et al. Depression screening using daily mental-health ratings from a smartphone application for breast cancer patients. J Med Internet Res 2016 Aug 04;18(8):e216 [FREE Full text] [doi: 10.2196/jmir.5598] [Medline: 27492880]
41. Horne E, Tibble H, Sheikh A, Tsanas A. Challenges of clustering multimodal clinical data: review of applications in asthma subtyping. JMIR Med Inform 2020 May 28;8(5):e16452 [FREE Full text] [doi: 10.2196/16452] [Medline: 32463370]
42. Doryab A, Villalba DK, Chikersal P, Dutcher JM, Tumminia M, Liu X, et al. Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and Fitbit data. JMIR Mhealth Uhealth 2019 Jul 24;7(7):e13209 [FREE Full text] [doi: 10.2196/13209] [Medline: 31342903]
43. Choudhary T, Sharma LN, Bhuyan MK, Bora K. Identification of human breathing-states using cardiac-vibrational signal for m-health applications. IEEE Sensors J 2021 Feb 1;21(3):3463-3470. [doi: 10.1109/jsen.2020.3025384]
44. Boukhechba M, Daros AR, Fua K, Chow PI, Teachman BA, Barnes LE. DemonicSalmon: monitoring mental health and social interactions of college students using smartphones. Smart Health 2018 Dec;9-10:192-203. [doi: 10.1016/j.smhl.2018.07.005]

45. Bublitz CF, Ribeiro-Teixeira AC, Pianoschi TA, Rochol J, Both CB. Unsupervised segmentation and classification of snoring events for mobile health. In: Proceedings of the 2017 IEEE Global Communications Conference. 2017 Presented at: GLOBECOM '17; December 4-8, 2017; Singapore, Singapore p. 1-6. [doi: 10.1109/GLOCOM.2017.8255031]
46. Cheung K, Ling W, Karr CJ, Weingardt K, Schueller SM, Mohr DC. Evaluation of a recommender app for apps for the treatment of depression and anxiety: an analysis of longitudinal user engagement. J Am Med Inform Assoc 2018 Aug 01;25(8):955-962 [FREE Full text] [doi: 10.1093/jamia/ocy023] [Medline: 29659857]
47. Berger T, Hohl E, Caspar F. Internet-based treatment for social phobia: a randomized controlled trial. J Clin Psychol 2009 Oct;65(10):1021-1035. [doi: 10.1002/jclp.20603] [Medline: 19437505]
48. Meng Q, Catchpoole D, Skillicom D, Kennedy PJ. Relational autoencoder for feature extraction. In: Proceedings of the 2017 International Joint Conference on Neural Networks. 2017 Presented at: IJCNN '17; May 14-19, 2017; Anchorage, AK, USA p. 364-371. [doi: 10.1109/ijcnn.2017.7965877]
49. Van Gemert-Pijnen JE, Kelders SM, Bohlmeijer ET. Understanding the usage of content in a mental health intervention for depression: an analysis of log data. J Med Internet Res 2014 Jan 31;16(1):e27 [FREE Full text] [doi: 10.2196/jmir.2991] [Medline: 24486914]
50. Greer JA, Jacobs JM, Pensak N, Nisotel LE, Fishbein JN, MacDonald JJ, et al. Randomized trial of a smartphone mobile app to improve symptoms and adherence to oral therapy for cancer. J Natl Compr Canc Netw 2020 Feb;18(2):133-141. [doi: 10.6004/jnccn.2019.7354] [Medline: 32023526]
51. Chien I, Enrique A, Palacios J, Regan T, Keegan D, Carter D, et al. A machine learning approach to understanding patterns of engagement with Internet-delivered mental health interventions. JAMA Netw Open 2020 Jul 01;3(7):e2010791 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.10791] [Medline: 32678450]
52. Pham Q, Graham G, Carrion C, Morita PP, Seto E, Stinson JN, et al. A library of analytic indicators to evaluate effective engagement with consumer mHealth apps for chronic conditions: scoping review. JMIR Mhealth Uhealth 2019 Jan 18;7(1):e11941 [FREE Full text] [doi: 10.2196/11941] [Medline: 30664463]
53. Maxwell AE, Warner TA, Fang F. Implementation of machine-learning classification in remote sensing: an applied review. Int J Remote Sens 2018 Feb 02;39(9):2784-2817. [doi: 10.1080/01431161.2018.1433343]
54. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Statist Surv 2010;4:40-79. [doi: 10.1214/09-SS054]
55. Ramezan CA, Warner TA, Maxwell AE. Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. Remote Sens 2019 Jan 18;11(2):185. [doi: 10.3390/rs11020185]
56. Chow PI, Showalter SL, Gerber M, Kennedy EM, Brenin D, Mohr DC, et al. Use of mental health apps by patients with breast cancer in the United States: pilot pre-post study. JMIR Cancer 2020 Apr 15;6(1):e16476 [FREE Full text] [doi: 10.2196/16476] [Medline: 32293570]
57. Kroenke K, Spitzer RL, Williams JB, Löwe B. An ultra-brief screening scale for anxiety and depression: the PHQ-4. Psychosomatics 2009;50(6):613-621. [doi: 10.1176/appi.psy.50.6.613] [Medline: 19996233]
58. Craig BM, Reeve BB, Brown PM, Cella D, Hays RD, Lipscomb J, et al. US valuation of health outcomes measured using the PROMIS-29. Value Health 2014 Dec;17(8):846-853 [FREE Full text] [doi: 10.1016/j.jval.2014.09.005] [Medline: 25498780]
59. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002 Jun 01;16(1):321-357. [doi: 10.1613/jair.953]
60. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA, USA p. 4768-4777 URL: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
61. Bandura A. Human agency in social cognitive theory. Am Psychol 1989 Sep;44(9):1175-1184. [doi: 10.1037/0003-066x.44.9.1175] [Medline: 2782727]
62. 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]
63. Champion VL, Skinner CS. The health belief model. In: Glanz K, Rimer BK, Viswanath K, editors. Health behavior and health education: theory, research, and practice. Hoboken, NJ, USA: Jossey-Bass; 2008:45-65.
64. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
65. Burgess C, Cornelius V, Love S, Graham J, Richards M, Ramirez A. Depression and anxiety in women with early breast cancer: five year observational cohort study. BMJ 2005 Mar 26;330(7493):702 [FREE Full text] [doi: 10.1136/bmj.38343.670868.D3] [Medline: 15695497]
66. Grabsch B, Clarke DM, Love A, McKenzie DP, Snyder RD, Bloch S, et al. Psychological morbidity and quality of life in women with advanced breast cancer: a cross-sectional survey. Palliat Support Care 2006 Mar;4(1):47-56. [doi: 10.1017/s1478951506060068] [Medline: 16889323]

67. Vogel BA, Bengel J, Helmes AW. Information and decision making: patients' needs and experiences in the course of breast cancer treatment. Patient Educ Couns 2008 Apr;71(1):79-85. [doi: 10.1016/j.pec.2007.11.023] [Medline: 18191933]
68. Harrison JD, Young JM, Price MA, Butow PN, Solomon MJ. What are the unmet supportive care needs of people with cancer? A systematic review. Support Care Cancer 2009 Aug;17(8):1117-1128. [doi: 10.1007/s00520-009-0615-5] [Medline: 19319577]
69. Vivar CG, McQueen A. Informational and emotional needs of long-term survivors of breast cancer. J Adv Nurs 2005 Sep;51(5):520-528. [doi: 10.1111/j.1365-2648.2005.03524.x] [Medline: 16098169]
70. Siebenhüner AR, Mikolasek M, Witt CM, Barth J. Improvements in health might contradict adherence to mobile health interventions: findings from a self-care cancer app study. J Altern Complement Med 2021 Mar;27(S1):S115-S123. [doi: 10.1089/acm.2020.0111] [Medline: 33788602]
71. Elhai JD, Levine JC, Dvorak RD, Hall BJ. Non-social features of smartphone use are most related to depression, anxiety and problematic smartphone use. Comput Human Behav 2017 Apr;69:75-82. [doi: 10.1016/j.chb.2016.12.023]

Abbreviations

FS: feature set
LOSOCV: leave-one-subject-out cross-validation
mHealth: mobile health
PHQ-4: Patient Health Questionnaire-4
PROMIS-29: Patient-Reported Outcomes Measurement Information System-29
RF: random forest
SMOTE: Synthetic Minority Oversampling Technique
XGB: XGBoost

Edited by C Lovis; submitted 26.05.21; peer-reviewed by P Laiou, J Kim; comments to author 23.09.21; revised version received 07.02.22; accepted 11.03.22; published 02.06.22.

Please cite as:
Baglione AN, Cai L, Bahrini A, Posey I, Boukhechba M, Chow PI
Understanding the Relationship Between Mood Symptoms and Mobile App Engagement Among Patients With Breast Cancer Using Machine Learning: Case Study
JMIR Med Inform 2022;10(6):e30712
URL: https://medinform.jmir.org/2022/6/e30712
doi:10.2196/30712
PMID:35653183

Original Paper

Noninvasive Screening Tool for Hyperkalemia Using a Single-Lead Electrocardiogram and Deep Learning: Development and Usability Study

Erdenebayar Urtnasan[1,2], PhD; Jung Hun Lee[3], MA; Byungjin Moon[2], MA; Hee Young Lee[2,3], PhD; Kyuhee Lee[1,2], PhD; Hyun Youk[2,4], MD

[1]Artificial Intelligence Big Data Medical Center, Wonju College of Medicine, Yonsei University, Wonju, Republic of Korea

[2]Bigdata Platform Business Group, Yonsei Wonju Health System, Wonju, Republic of Korea

[3]Department of Emergency Medicine, Wonju College of Medicine, Yonsei University, Wonju, Republic of Korea

[4]Center of Regional Trauma, Wonju, Republic of Korea

Corresponding Author:
Hyun Youk, MD
Center of Regional Trauma
20 Ilsan-ro
Wonju Severance Christian Hospital
Wonju, 26426
Republic of Korea
Phone: 82 10 9840 2120
Email: yhmentor@yonsei.ac.kr

Abstract

Background: Hyperkalemia monitoring is very important in patients with chronic kidney disease (CKD) in emergency medicine. Currently, blood testing is regarded as the standard way to diagnose hyperkalemia (ie, using serum potassium levels). Therefore, an alternative and noninvasive method is required for real-time monitoring of hyperkalemia in the emergency medicine department.

Objective: This study aimed to propose a novel method for noninvasive screening of hyperkalemia using a single-lead electrocardiogram (ECG) based on a deep learning model.

Methods: For this study, 2958 patients with hyperkalemia events from July 2009 to June 2019 were enrolled at 1 regional emergency center, of which 1790 were diagnosed with chronic renal failure before hyperkalemic events. Patients who did not have biochemical electrolyte tests corresponding to the original 12-lead ECG signal were excluded. We used data from 855 patients (555 patients with CKD, and 300 patients without CKD). The 12-lead ECG signal was collected at the time of the hyperkalemic event, prior to the event, and after the event for each patient. All 12-lead ECG signals were matched with an electrolyte test within 2 hours of each ECG to form a data set. We then analyzed the ECG signals with a duration of 2 seconds and a segment composed of 1400 samples. The data set was randomly divided into the training set, validation set, and test set according to the ratio of 6:2:2 percent. The proposed noninvasive screening tool used a deep learning model that can express the complex and cyclic rhythm of cardiac activity. The deep learning model consists of convolutional and pooling layers for noninvasive screening of the serum potassium level from an ECG signal. To extract an optimal single-lead ECG, we evaluated the performances of the proposed deep learning model for each lead including lead I, II, and V1-V6.

Results: The proposed noninvasive screening tool using a single-lead ECG shows high performances with F1 scores of 100%, 96%, and 95% for the training set, validation set, and test set, respectively. The lead II signal was shown to have the highest performance among the ECG leads.

Conclusions: We developed a novel method for noninvasive screening of hyperkalemia using a single-lead ECG signal, and it can be used as a helpful tool in emergency medicine.

(JMIR Med Inform 2022;10(6):e34724) doi:10.2196/34724

KEYWORDS

hyperkalemia; ECG; electrocardiogram; deep learning; noninvasive screening; emergency medicine; single-lead ECG

Introduction

Hyperkalemia is a potential life-threatening condition for the general population, and so it can be a clinical and economic burden [1]. Normal levels of potassium are between 3.5 and 5.0 mmol/L with levels above 5.5 mmol/L defined as hyperkalemia. Patients with chronic kidney disease (CKD) are predisposed to hyperkalemia [2], and it is a major risk factor for cardiac arrhythmias and death [3]. According to some clinical studies, serum potassium monitoring can reduce the risk of hyperkalemia in patients with CKD by more than 71% [4]. Therefore, it is very important to frequently check the serum potassium level in patients with CKD.

Potassium is a very important electrolyte for the regulation of the cell membrane potential and nerve conduction, so abnormal levels of potassium are known to be associated with changes in electrocardiogram (ECG) readings. Hyperkalemia is associated with tall, narrow, and symmetrical T waves in an ECG, whereas hypokalemia is associated with flat T waves [5-9]. Monitoring hyperkalemia is very important in patients with CKD, but so far blood testing is the only way to test serum potassium levels. Closer and more reliable monitoring requires the development and verification of noninvasive and continuous monitoring methods.

Electrocardiography is used to detect heart abnormalities in patients with various diseases. The main ECG changes associated with hypokalemia include a decreased T wave amplitude, ST-segment depression, T wave inversion, a prolonged PR interval, and an increased corrected QT interval [10]. The typical ECG findings for hyperkalemia progress from tall, peaked T waves and a shortened QT interval to a lengthened PR interval and a loss of the P wave followed by a widening QRS complex and ultimately a sine wave morphology [11,12]. These morphologic differences of the ECG have been used to detect and diagnose hyperkalemia events urgently in emergency rooms [13]. In addition, there are some studies that have proposed several methods to detect hyperkalemia events using ECG signals. Among them, some researchers have developed ECG quantification algorithms to predict serum potassium concentration based on T wave morphology, mainly using the slope and width of T waves. The algorithms were mostly derived from continuous patient monitoring, such as during hemodialysis, with homogeneous ECG morphologies from a limited set of patients [14,15]. Recently, applying the processing of T wave morphologies manually has been used to improve the diagnosis of hyperkalemia [16]. Nevertheless, using T wave changes alone to detect dyskalemias is less sensitive and specific than a comprehensive ECG interpretation [17]. However, ECG morphology–based methods have shown insufficient performance and require some time to extract the morphologic features required to detect hyperkalemia. Therefore, a more robust and faster method for hyperkalemia detection is needed in the clinical practice of emergency medicine.

With the revolution of artificial intelligence (AI), many deep learning models have been developed that show human-level performance in several clinical fields such as cardiology [18], radiology [19], ophthalmology [20], and pathology [21]. For instance, convolutional neural network (CNN) models have achieved very high performances for abnormal cardiac rhythms such as arrhythmia [22], tachycardia [23], and supraventricular dysfunction [24], among other events [25]. Such diagnostic and prognostic deep learning models could be developed to assist emergency medicine clinicians in recognizing ECG changes associated with diverse diseases. AI algorithms have emerged in clinical decision support systems as “software as a medical device” in a real clinical environment [26]. There are some similar studies conducted by several researchers. Among them, Galloway et al [11] proposed a CNN-based model to screen for hyperkalemia using a multilead ECG signal. They demonstrated a deep CNN model with complex architecture and evaluated its performance using big ECG data sets in a multicenter cohort study. Another study involved the prediction of serum potassium concentration based on an 82-layer CNN model using a 12-lead ECG signal [12]. They achieved robust performance with more than 95.8% hyperkalemia detection. However, all these recent deep learning models are hard to apply to real-time analysis in clinical practices.

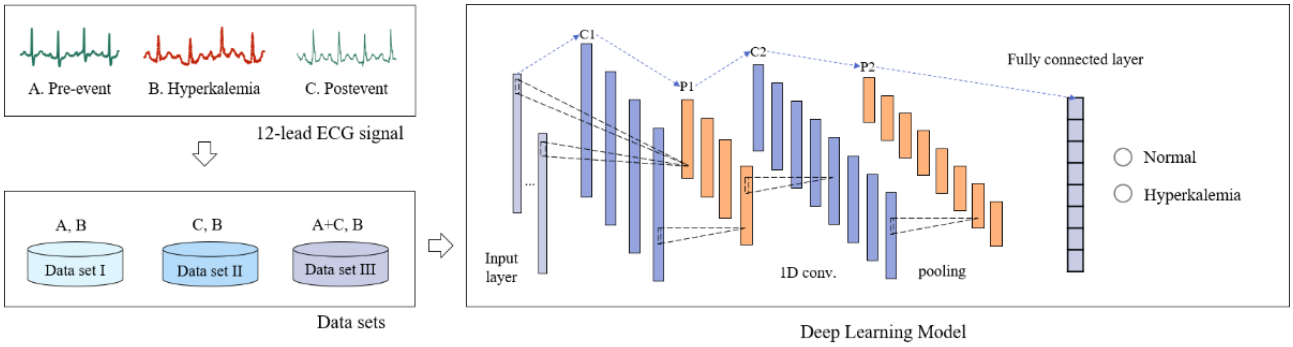
Therefore, this study proposed a novel method for noninvasive screening of hyperkalemia based on a deep learning model using an ECG. For this purpose, we constructed a deep learning structure using a CNN model, and clinical data were used for the training and testing phases. In addition, we conducted several experiments using the different data sets, applying the changes before and after hyperkalemia events. Finally, a simple and accurate deep learning model was designed to implement a noninvasive screening method for hyperkalemia that can be applied to real-time clinical practices.

Methods

Overview

In this study, we proposed a novel method for noninvasive screening of hyperkalemia based on a deep learning model, using ECG. The proposed method consists of the following 3 main parts: 12-lead ECG extraction from the participants, constructing the ECG data sets, and a deep learning model (Figure 1). Each part of the study method is explained in more detail in the following subsections.

Figure 1. Diagram of the proposed method for noninvasive screening of hyperkalemia. ECG: electrocardiogram.



Ethics Approval

This study was approved by the Institutional Review Board (IRB) of the Wonju Severance Christian Hospital (CR320162). Enrolled patients' informed consent was exempted by the IRB due to the retrospective nature of the study that used fully anonymized ECG and health data.

Participants

A total of 2958 patients who have experienced at least 1 hyperkalemia event were enrolled at a single regional emergency center from July 2009 to June 2019. Among them, 1790 patients were diagnosed with chronic renal failure (CRF), and the other 1168 patients did not have CRF. The patients who did not have biochemical electrolyte tests corresponding to the original 12-lead ECG signal were excluded. We then used the data of 855 patients (555 patients with CRF, 300 patients without CRF) (Table 1).

Table 1. Characteristics of the study participants.

Characteristics	Participants (N=2958)		
	Non-CRF[a] (n=1168)	CRF (n=1790)	Total
Gender, n (%)[b]			
Female	496 (39.9)	747 (60.1)	1243 (42)
Male	672 (39.2)	1043 (60.8)	1715 (58)
Total	1168 (39.5)	1790 (60.5)	2958 (100)
Age (years), mean (SD)	70.3 (19.0)	72.6 (13.2)	71.7 (15.8)
Height (cm), mean (SD)	155.5 (26.2)	159.4 (14.4)	158.0 (19.7)
Weight (kg), mean (SD)	58.8 (15.5)	62.2 (12.2)	60.9 (13.6)
Myocardial infarction, n (%)[b]	35 (23)	117 (77)	152 (5.1)
Heart failure, n (%)[b]	116 (30)	271 (70)	387 (13.1)
Angina, n (%)[b]	93 (28.4)	235 (71.6)	328 (11.1)
Diabetes, n (%)[b]	251 (21.6)	912 (78.4)	1163 (39.3)
Hypertension, n (%)[b]	323 (23.8)	1037 (76.3)	1360 (46)

[a]CRF: chronic renal failure.

[b]The denominator used to calculate percentages is the sum of the non-CRF and CRF participants in that category (ie, row).

Data Sets

The 12-lead ECG recordings were collected at the 3 sections of hyperkalemic events: pre-event, event, and postevent. The pre- and postevents were used as normal or control events, and the hyperkalemic event was used as the target or abnormal event. From these 3 sections, the data sets were designed, including data set I (pre-event vs event), data set II (postevent vs event), and data set III (pre-event and postevent vs event). The differences between before- and aftereffects of a hyperkalemia event are presented in Table 2. All 12-lead ECG recordings were matched with an electrolyte test within 2 hours in each section to form a data set. The waveform of the 12-lead ECG signal was then extracted and saved with a sampling frequency of 700 Hz. Finally, an ECG signal segment with a duration of 2 seconds was composed of 1400 samples. For evaluation of the developed AI algorithm, the ECG data set was randomly divided by the ratio of 6:2:2 percent into the training set, validation set, and test set for each data set.

Table 2. Data sets for this study.

Data sets	Data set I, n	Data set II, n	Data set III, n
Training set	1186	879	1426
Validation set	296	220	357
Test set	370	275	446
Total	1852	1374	2229

Deep Learning Model

Deep learning is a method for representation learning that can learn the complex pattern and structure of the input data by high-level data abstraction. It can learn the morphology of the input ECG signal according to the potassium concentrations. A deep learning model was designed based on a 5-layer CNN by using a 1-dimensional convolutional operation, max pooling, and a fully connected layer. The detailed structure of the proposed deep learning model for noninvasive screening of hyperkalemia using ECG signal is shown in Table 3.

Table 3. Architecture of the proposed deep learning model for hyperkalemia screening.

Number and layers	Activation	Filter size	Output shape	Parameter
1				
• batchnorm_1	=	• =	• 1400×1	4
2				
• conv1D_1 • maxpool_1	ReLu	• 100@50×1 • 2×1	• 1351×100 • 675×100	5100
3				
• conv1D_2 • maxpool_2 • dropout_2	ReLu	• 80@50×1 • 2×1 • p=0.25[a]	• 626×80 • 313×80 • 313×80	400,080
4				
• conv1D_3 • maxpool_3 • dropout_3	ReLu	• 60@30×1 • 2×1 • p=0.25	• 284×60 • 142×60 • 142×60	144,060
5				
• conv1D_4 • maxpool_4 • dropout_4	ReLu	• 40@20×1 • 2×1 • p=0.25	• 123×40 • 61×40 • 61×40	48,040
6				
• conv1D_5 • maxpool_5 • dropout_5	ReLu	• 20@10×1 • 2×1 • p=0.25	• 52×20 • 26×20 • 26×20	8020
7				
• flatten_1 • dense_1	Softmax	• 2	• 520×2	1042
Total • 5 conv layers		• 124 filters		606,027

[a]*p*: One of the setting parameters of the dropout technique.

Statistical Analysis

The F1 score was used to evaluate the proposed noninvasive screening method for hyperkalemia; it evaluates the correct classification of each class according to class equality. F1 scores calculated by precision and recall are represented as follows:

The numbers of true positives (TP), false positives (FP), and false negatives (FN) are input into the equations. The F1 score is computed based on the sample proportion of precision and recall as follows:

Results

The results of the proposed novel method for noninvasive screening of hyperkalemia based on deep learning using a 12-lead ECG are shown in Table 4-Table 6 for the test set of each data set. The results of the proposed method showed that lead II achieved the highest performance for hyperkalemia events among other leads for data set I (Table 4).

Table 4. The performance of the proposed method for the test set of data set I.

Index and events	Leads							
	I	II	V1	V2	V3	V4	V5	V6
Precision								
Normal	0.52	0.96	0.47	0.61	0.56	0.66	0.51	0.54
Hyperkalemia	0.61	0.94	0.63	0.70	0.63	0.71	0.64	0.63
Recall								
Normal	0.48	0.93	0.56	0.66	0.51	0.66	0.50	0.60
Hyperkalemia	0.64	0.97	0.54	0.65	0.68	0.71	0.65	0.58
F1 score								
Normal	0.50	0.94	0.51	0.64	0.53	0.66	0.50	0.57
Hyperkalemia	0.62	0.95	0.58	0.68	0.66	0.71	0.65	0.60

Table 5. The performance of the proposed method for the test set of data set II.

Index and events	Leads							
	I	II	V1	V2	V3	V4	V5	V6
Precision								
Normal	0.31	0.88	0.28	0.36	0.28	0.52	0.36	0.51
Hyperkalemia	0.75	0.93	0.74	0.73	0.80	0.79	0.76	0.72
Recall								
Normal	0.22	0.85	0.17	0.27	0.23	0.50	0.28	0.30
Hyperkalemia	0.82	0.95	0.84	0.81	0.84	0.81	0.82	0.87
F1 score								
Normal	0.26	0.87	0.21	0.31	0.25	0.51	0.31	0.38
Hyperkalemia	0.78	0.94	0.79	0.77	0.82	0.80	0.79	0.79

Table 6. The performance of the proposed method for the test set of data set III.

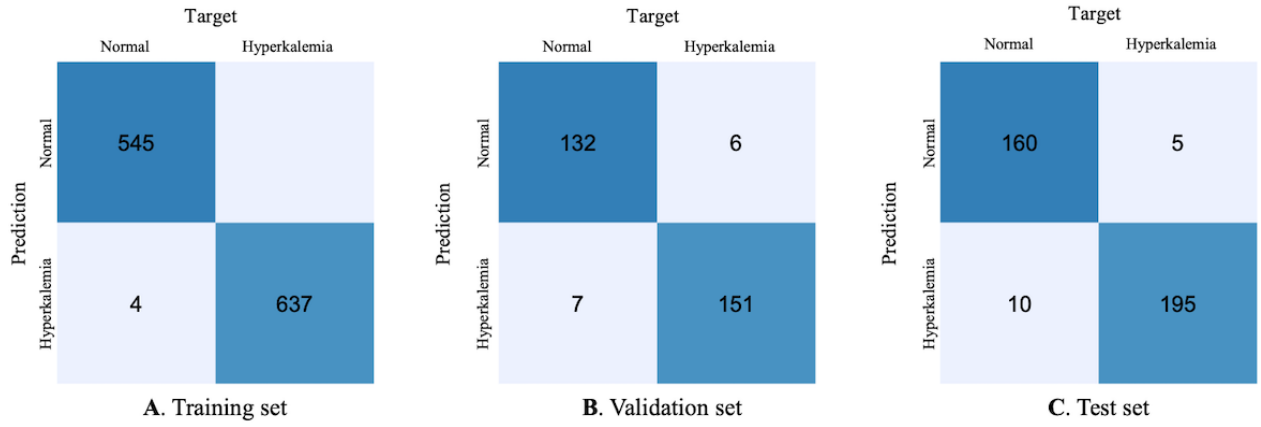
Index and events	Leads							
	I	II	V1	V2	V3	V4	V5	V6
Precision								
Normal	0.56	0.95	0.53	0.68	0.65	0.69	0.57	0.61
Hyperkalemia	0.47	0.94	0.74	0.59	0.57	0.60	0.60	0.51
Recall								
Normal	0.61	0.96	1.00	0.70	0.62	0.63	0.68	0.59
Hyperkalemia	0.42	0.93	0.00	0.57	0.60	0.66	0.48	0.53
F1 score								
Normal	0.58	0.96	0.70	0.69	0.64	0.66	0.62	0.60
Hyperkalemia	0.44	0.94	0.00	0.58	0.59	0.63	0.53	0.52

We also noticed that there are big performance differences between lead II and other leads not only in data set I, but also in data sets II and III. The results showed that V2 achieved the best performance among the V1-V6 leads throughout the 3 different data sets.

We obtained good performances of the proposed deep learning model for noninvasive screening of hyperkalemia using lead II signal, with F1 scores of 95%, 94%, and 94% for data set I, data set II, and data set III for the test set.

For data set I, we presented the confusion matrix of the training set, validation set, and test set for lead II of the ECG. The results showed good performance, with F1scores of 100%, 96%, and 95% for the training set, validation set, and test set, respectively. The confusion matrix showed that the proposed deep learning model gained a high and stable rate for the true positives and false negatives (Figure 2).

Figure 2. Confusion matrix of this study. Confusion matrix of (A) the training set, (B) the validation set, and (C) the test set for the lead II electrocardiogram channel of data set I.



Discussion

In this study, we demonstrated a novel method for noninvasive screening of hyperkalemia based on deep learning using ECG. We designed an optimal and simple architecture of deep learning that can be easily implemented for real clinical applications, especially in emergency medicine. The proposed model achieved good performances based on feature extraction of the characteristics of cardiac activity according to the levels of electrolytes using ECG.

We have tried to see the morphological or rhythmical differences between hyperkalemia pre- and postevents in ECG waveform. To do this, we set up 3 different data sets that selected from the pre-event, postevent, and target event sections: data set I (pre-event vs event), data set II (postevent vs event), and data set III (pre- and postevent vs event). All data sets were applied to the designed deep learning model for hyperkalemia screening in training, validation, and test phases. We trained and tested the 3 different deep learning models using each data set. We also conducted experiments to find the optimal signal of the 12-lead ECG for each data set one by one. We determined that an optimal lead for hyperkalemia screening was lead II for our 12-lead ECG data sets. In general, lead II contains the most information on cardiac activity among the leads, which may have resulted in it having the highest performance for the hyperkalemia screening. In addition, our results support conventional studies on hyperkalemia screening using ECG signals.

There are some similar studies that proposed a screening or detection tool for hyperkalemia using ECG signals. The earliest

one was proposed by Wrenn et al [10] in 1991. This study compared hyperkalemia detection by 2 independent and experienced physicians using ECG signals. The results showed a sensitivity of 62.0% for the first reader and 55.0% for the second reader and the κ value was 0.73. This showed how difficult it was to detect hyperkalemia from the ECG signal without any biochemical electrolyte tests. Another study on the correlation between hyperkalemia and ECG morphologies was conducted by Levis [27] in 2013. They published a clinical case study on ECG signals observed as a hyperkalemia event occurs. The authors demonstrated 2 different cases of older adults with acute renal failure and hyperkalemia. They confirmed the ECG morphology changes with peaked T waves, shortened QT interval, and lengthening PR intervals corresponding to the hyperkalemia or changes of serum potassium levels.

Recently, deep learning models have been applied to studies on detection and screening of hyperkalemia from the 12-lead ECG recordings made in emergency departments. Galloway et al [11] developed and validated a deep learning model to screen for hyperkalemia using ECG. The authors designed an 11-layer deep CNN model, and it was trained and validated with 449,380 patients with CKD. They used 2 data sets for 2 leads (I and II) and 4 leads (I, II, V3, and V5); each patient had a serum potassium count drawn within 4 hours after their ECG was recorded. In this multicenter cohort study, a deep learning model was developed with complex architecture, using big ECG data sets [11]. However, they achieved a good performance, with an area under the curve (AUC) of 88.83% and a sensitivity of 91.3% for the 2 leads' data sets. In contrast, we proposed a deep learning model with a light weight and high performance using a single-lead ECG signal.

Lin et al [12] developed a 12-channel sequence-to-sequence model with an 82-layer CNN structure to predict serum potassium concentration by using a 12-lead ECG signal. They modified DenseNet architecture to read the 12-lead ECG waveforms and detect hypokalemia and hyperkalemia events, and named it ECG12Net. ECG12Net achieved robust performances, with an AUC of 95.8% and 97.6% for hyperkalemia and severe hyperkalemia, respectively. However, ECG12Net requires very high computational power to read the 12-lead ECG signal since it is composed of an 82-layer CNN model. Our method is comparable in performance with this model, and it is well optimized and trained for a 1-channel ECG signal to detect hyperkalemia events. In addition, it is easy to make it into a tool that can be used in the final clinical field because the deep learning engine is relatively lighter.

The proposed new method for noninvasive screening of hyperkalemia based on deep learning using ECG signals surpasses similar previous studies, and the developed model can be applied directly to clinical situations. This noninvasive method does not require any blood test or invasive chemistry diagnosis. In addition, the physicians or clinicians can check the results quickly, within 3 minutes, which is faster than previous invasive diagnostic methods. This is because the deep learning model has a simple structure and is well optimized and trained by the pre- and postevent's ECG signals to screen for hyperkalemia events. In addition, the proposed deep learning model can proceed with feature extraction and classification at once from the input ECG signal for noninvasive screening of hyperkalemia because we do not use any handcrafted preprocessing that extracts input ECG components such as the RR interval or P wave. Finally, we achieved a higher performance of the proposed noninvasive method based on a deep learning model; it can consider the complex and cyclic characteristics of ECG affected by levels of electrolytes.

This study has some limitations, such as the small study population, and there are many comorbidities including heart failure, diabetes, and hypertension in the study groups. In addition, all participants of this study are enrolled at a single regional emergency center, so further study should cover large and diverse populations from multiple centers.

We demonstrated a novel method for noninvasive screening of hyperkalemia based on deep learning using ECG. We obtained high performances with an F1 score of 95% from the ECG signal. In addition, we developed a simple and accurate deep learning model for the noninvasive screening of hyperkalemia that can be used in real-time clinical settings. Therefore, the proposed deep learning model may be appropriate for the noninvasive screening of hyperkalemia using a single-lead ECG signal without any feature extraction (eg, T wave and QT interval). Furthermore, a validation study should be conducted for the proposed deep learning model that uses larger and more diverse data sets based on a single-lead ECG signal.

Acknowledgments

This research was supported by the National Information Society Agency funded by the Ministry of Science and ICT through the Big Data Platform and Center Construction project (2022-Data-W18). It was also partially supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 202011B24, KMDF_PR_20200901_0030-2021-03).

Conflicts of Interest

None declared.

References

1. Fitch K, Woolley JM, Engel T, Blumen H. The clinical and economic burden of hyperkalemia on medicare and commercial payers. Am Health Drug Benefits 2017 Jun;10(4):202-210 [FREE Full text] [Medline: 28794824]

2. Luo J, Brunelli SM, Jensen DE, Yang A. Association between serum potassium and outcomes in patients with reduced kidney function. CJASN 2015 Oct 23;11(1):90-100. [doi: 10.2215/cjn.01730215]
3. Einhorn LM, Zhan M, Hsu VD, Walker LD, Moen MF, Seliger SL, et al. The frequency of hyperkalemia and its significance in chronic kidney disease. Arch Intern Med 2009 Jun 22;169(12):1156-1162 [FREE Full text] [doi: 10.1001/archinternmed.2009.132] [Medline: 19546417]
4. Raebel MA, Ross C, Xu S, Roblin DW, Cheetham C, Blanchette CM, et al. Diabetes and drug-associated hyperkalemia: effect of potassium monitoring. J Gen Intern Med 2010 Apr 20;25(4):326-333 [FREE Full text] [doi: 10.1007/s11606-009-1228-x] [Medline: 20087674]
5. Levey AS, Rocco MV, Anderson S, Andreoli SP, Bailie GR, Bakris GL. K/DOQI clinical practice guidelines on hypertension and antihypertensive agents in chronic kidney disease. AJKD 2004 Mar;43:S1-S290. [doi: 10.1053/j.ajkd.2004.03.003]
6. Jacobs AK, Kushner FG, Ettinger SM, Guyton RA, Anderson JL, Ohman EM, et al. ACCF/AHA Clinical practice guideline methodology summit report. Circulation 2013 Jan 15;127(2):268-310. [doi: 10.1161/cir.0b013e31827e8e5f]
7. Schmidt M, Mansfield KE, Bhaskaran K, Nitsch D, Sørensen HT, Smeeth L, et al. Adherence to guidelines for creatinine and potassium monitoring and discontinuation following renin-angiotensin system blockade: a UK general practice-based cohort study. BMJ Open 2017 Jan 09;7(1):e012818 [FREE Full text] [doi: 10.1136/bmjopen-2016-012818] [Medline: 28069618]
8. Cooper LB, Hammill BG, Peterson ED, Pitt B, Maciejewski ML, Curtis LH, et al. Consistency of laboratory monitoring during initiation of mineralocorticoid receptor antagonist therapy in patients with heart failure. JAMA 2015 Nov 10;314(18):1973. [doi: 10.1001/jama.2015.11904]
9. Surawicz B. Relationship between electrocardiogram and electrolytes. Am Heart J 1967 Jun;73(6):814-834. [doi: 10.1016/0002-8703(67)90233-5]
10. Wrenn KD, Slovis CM, Slovis BS. The ability of physicians to predict hyperkalemia from the ECG. Ann Emerg Med 1991 Nov;20(11):1229-1232. [doi: 10.1016/s0196-0644(05)81476-3]
11. Galloway CD, Valys AV, Shreibati JB, Treiman DL, Petterson FL, Gundotra VP, et al. Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. JAMA Cardiol 2019 May 01;4(5):428. [doi: 10.1001/jamacardio.2019.0640]
12. Lin C, Lin C, Fang W, Hsu C, Chen S, Huang K, et al. A deep-learning algorithm (ECG12Net) for detecting hypokalemia and hyperkalemia by electrocardiography: algorithm development. JMIR Med Inform 2020 Mar 05;8(3):e15931 [FREE Full text] [doi: 10.2196/15931] [Medline: 32134388]
13. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017 Dec 12;318(22):2211-2223 [FREE Full text] [doi: 10.1001/jama.2017.18152] [Medline: 29234807]
14. Varga C, Kálmán Z, Szakáll A, Drubits K, Koch M, Bánhegyi R, et al. ECG alterations suggestive of hyperkalemia in normokalemic versus hyperkalemic patients. BMC Emerg Med 2019 May 31;19(1):33 [FREE Full text] [doi: 10.1186/s12873-019-0247-0] [Medline: 31151388]
15. Rafique Z, Aceves J, Espina I, Peacock F, Sheikh-Hamad D, Kuo D. Can physicians detect hyperkalemia based on the electrocardiogram? Am J Emerg Med 2020 Jan;38(1):105-108. [doi: 10.1016/j.ajem.2019.04.036] [Medline: 31047740]
16. Galloway CD, Valys AV, Petterson FL, Gundotra VP, Treiman DL, Albert DE, et al. Non-invasive detection of hyperkalemia with a smartphone electrocardiogram and artificial intelligence. J Am Coll Cardiol 2018 Mar;71(11):A272. [doi: 10.1016/s0735-1097(18)30813-1]
17. Teplitzky BA, McRoberts M, Ghanbari H. Deep learning for comprehensive ECG annotation. Heart Rhythm 2020 May;17(5 Pt B):881-888 [FREE Full text] [doi: 10.1016/j.hrthm.2020.02.015] [Medline: 32354454]
18. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. J Cardiovasc Magn Reson 2018 Sep 14;20(1):65 [FREE Full text] [doi: 10.1186/s12968-018-0471-x] [Medline: 30217194]
19. Nam JG, Park S, Hwang EJ, Lee JH, Jin K, Lim KY, et al. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology 2019 Jan;290(1):218-228. [doi: 10.1148/radiol.2018180237] [Medline: 30251934]
20. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
21. Chang HY, Jung CK, Woo JI, Lee S, Cho J, Kim SW, et al. Artificial intelligence in pathology. J Pathol Transl Med 2019 Jan;53(1):1-12 [FREE Full text] [doi: 10.4132/jptm.2018.12.16] [Medline: 30599506]
22. Erdenebayar U, Kim H, Park J, Kang D, Lee K. Automatic prediction of atrial fibrillation based on convolutional neural network using a short-term normal electrocardiogram signal. J Korean Med Sci 2019 Feb 25;34(7):e64 [FREE Full text] [doi: 10.3346/jkms.2019.34.e64] [Medline: 30804732]
23. Madias JE. Apparent electrocardiogram left ventricular hypertrophy during tachycardia. J Electrocardiol 2021 Mar;65:3-7. [doi: 10.1016/j.jelectrocard.2021.01.001] [Medline: 33460860]

24. Lih OS, Jahmunah V, San TR, Ciaccio EJ, Yamakawa T, Tanabe M, et al. Comprehensive electrocardiographic diagnosis based on deep learning. Artif Intell Med 2020 Mar;103:101789. [doi: 10.1016/j.artmed.2019.101789] [Medline: 32143796]
25. Ebrahimi Z, Loni M, Daneshtalab M, Gharehbaghi A. A review on deep learning methods for ECG arrhythmia classification. Expert Syst Appl 2020 Sep;7:100033. [doi: 10.1016/j.eswax.2020.100033]
26. Mincholé A, Rodriguez B. Artificial intelligence for the electrocardiogram. Nat Med 2019 Jan 7;25(1):22-23. [doi: 10.1038/s41591-018-0306-1] [Medline: 30617324]
27. Levis JT. ECG diagnosis: hyperkalemia. Perm J 2013;17(1):69 [FREE Full text] [doi: 10.7812/TPP/12-088] [Medline: 23596374]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CKD: chronic kidney disease
CNN: convolutional neural network
ECG: electrocardiogram
IRB: Institutional Review Board

Edited by C Lovis; submitted 05.11.21; peer-reviewed by H Turbe, H Kim; comments to author 20.12.21; revised version received 21.03.22; accepted 11.04.22; published 03.06.22.

Please cite as:
Urtnasan E, Lee JH, Moon B, Lee HY, Lee K, Youk H
Noninvasive Screening Tool for Hyperkalemia Using a Single-Lead Electrocardiogram and Deep Learning: Development and Usability Study
JMIR Med Inform 2022;10(6):e34724
URL: https://medinform.jmir.org/2022/6/e34724
doi:10.2196/34724
PMID:35657658

Original Paper

Predicting Abnormal Laboratory Blood Test Results in the Intensive Care Unit Using Novel Features Based on Information Theory and Historical Conditional Probability: Observational Study

Camilo E Valderrama[1,2,3], PhD; Daniel J Niven[4], MD, PhD; Henry T Stelfox[3,4], MD, PhD; Joon Lee[1,2,3,5], PhD

[1]Data Intelligence for Health Lab, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[2]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[3]O'Brien Institute for Public Health, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[4]Department of Critical Care Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[5]Department of Cardiac Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

Corresponding Author:
Camilo E Valderrama, PhD
Data Intelligence for Health Lab
Cumming School of Medicine
University of Calgary
3280 Hospital Dr NW
Calgary, AB, T2N 4Z6
Canada
Phone: 1 403 220 8230
Email: camilo.valderramacua@ucalgary.ca

Abstract

Background: Redundancy in laboratory blood tests is common in intensive care units (ICUs), affecting patients' health and increasing health care expenses. Medical communities have made recommendations to order laboratory tests more judiciously. Wise selection can rely on modern data-driven approaches that have been shown to help identify low-yield laboratory blood tests in ICUs. However, although conditional entropy and conditional probability distribution have shown the potential to measure the uncertainty of yielding an abnormal test, no previous studies have adapted these techniques to include them in machine learning models for predicting abnormal laboratory test results.

Objective: This study aimed to address the limitations of previous reports by adapting conditional entropy and conditional probability to extract features for predicting abnormal laboratory blood test results.

Methods: We used an ICU data set collected across Alberta, Canada, which included 55,689 ICU admissions from 48,672 patients. We investigated the features of conditional entropy and conditional probability by comparing the performances of 2 machine learning approaches for predicting normal and abnormal results for 18 blood laboratory tests. Approach 1 used patients' vitals, age, sex, and admission diagnosis as features. Approach 2 used the same features plus the new conditional entropy–based and conditional probability–based features. Both approaches used 4 different machine learning models (fuzzy model, logistic regression, random forest, and gradient boosting trees) and 10 metrics (sensitivity, specificity, accuracy, precision, negative predictive value [NPV], F_1 score, area under the curve [AUC], precision-recall AUC, mean G, and index balanced accuracy) to assess the performance of the approaches.

Results: Approach 1 achieved an average AUC of 0.86 for all 18 laboratory tests across the 4 models (sensitivity 78%, specificity 84%, precision 82%, NPV 75%, F_1 score 79%, and mean G 81%), whereas approach 2 achieved an average AUC of 0.89 (sensitivity 84%, specificity 84%, precision 83%, NPV 81%, F_1 score 83%, and mean G 84%). We found that the inclusion of the new features resulted in significant differences for most of the metrics in favor of approach 2. Sensitivity significantly improved for 8 and 15 laboratory tests across the different classifiers (minimum $P<.001$ and maximum $P=.04$). Mean G and index balanced accuracy, which are balanced performance metrics, also improved significantly across the classifiers for 6 to 10 and 6 to 11 laboratory tests. The most relevant feature was the pretest probability feature, which is the probability that a test result was normal when a certain number of consecutive prior tests was already normal.

Conclusions: The findings suggest that conditional entropy–based features and pretest probability improve the capacity to discriminate between normal and abnormal laboratory test results. Detecting the next laboratory test result is an intermediate step toward developing guidelines for reducing overtesting in the ICU.

(JMIR Med Inform 2022;10(6):e35250) doi:10.2196/35250

KEYWORDS

blood laboratory test redundancy; intensive care unit; electronic medical records; machine learning; fuzzy modeling

Introduction

Background

Redundancy in laboratory blood tests is common in health care [1]. Laboratory blood test redundancy increases health care expenses and reduces health care resources for future patients [1,2]. Moreover, overtesting in the intensive care unit (ICU) can harm patients' health by causing anemia, the need for transfusion, discomfort, and poor sleep quality [3-8].

One of the areas greatly experiencing laboratory blood test redundancy is ICUs, in which daily blood tests are performed to monitor physiological functions and define clinical management strategies. Previous reports have underscored overtesting in ICUs. In a study conducted in an ICU of a tertiary hospital in Ontario, Canada, physicians retrospectively analyzed 694 blood tests performed over 4 weeks and concluded that only 48.7% of those tests were essential [9]. A similar pattern was found in a Brazilian ICU, in which approximately half (1768/3622, 48.81%) blood tests performed over 2 months resulted in normal values [10].

To reduce redundancy in the ICU, the Choosing Wisely campaign has made recommendations to order laboratory tests judiciously [11]. These recommendations have been introduced in the ICU via strategies such as education, audits and feedback, and computerized physician order entry systems [12-14]. However, these recommendations require accurate identification of laboratory tests that can be reduced without compromising the quality of patient care.

Modern data-driven approaches can help identify redundant laboratory blood tests in ICUs [15]. A study by Lee and Maslove [16] used entropy, conditional entropy, and mutual information to measure redundancy in 11 blood tests performed during the first 3 days in the ICU. They found a decreasing trend in the novelty of information throughout the ICU stay, showing that performing additional laboratory tests does not necessarily result in the gain of information. Roy et al [17] used laboratory blood test data from a tertiary academic hospital to calculate the conditional probability of a test yielding a normal result when a certain number of consecutive prior tests were already normal (pretest probability). They reported that common laboratory tests, such as those for creatinine, potassium, and sodium, had high chances of yielding normal results (>80%) when preceded by a small number (3-5) of consecutive normal results.

In addition to using data-driven approaches to describe redundancy in the ICU, other reports have used electronic medical record (EMR) data collected during the ICU stay to predict whether ordering a new blood test would provide new information. Cismondi et al [18] used heart rate, blood pressure, temperature, pulse oximeter, respiratory rate, 4 transfusion quantities, and the value of the first laboratory test performed in the day to classify redundancy among 8 different types of laboratory blood tests, with an average redundancy rate of 53%, provided to 746 patients with gastrointestinal bleeding in an ICU. To this end, they used a supervised machine learning approach with a fuzzy model, achieving an average accuracy of 79.5% for detecting redundant tests. Mahani and Pajoohan [19] followed a similar approach to predict the values of calcium and hematocrit blood tests in the same type of patients, achieving a mean absolute error of 0.03 mg/dL and 2.60%, respectively. A study by Roy et al [17] reported a maximum area under the curve (AUC) value of 0.88 for predicting low information laboratory diagnostic tests using a random forest (RF) on an extensive feature set comprising patients' demographics, vitals, and descriptive statistics of 12 additional laboratory tests. This study was further extended by Xu et al [20], who used between 600 and 870 raw features from EMRs to predict normal laboratory results collected from 3 tertiary hospitals, achieving an area under the receiver operating characteristic curve of ≥0.90 for 12 laboratory tests.

More complex models based on deep learning have also been used to recommend laboratory reduction strategies. Yu et al [21] developed a spatial-temporal deep learning model using patients' laboratory tests, time differences between adjacent visits, and demographics to predict the following four outputs: (1) the necessity of ordering a new laboratory test, (2) test values, (3) abnormalities (based on normal reference ranges), and (4) transitions (normal to abnormal or abnormal to normal from the latest laboratory test). By assessing different thresholds for their estimated necessity of a new test, the authors achieved a reduction rate of 20.26%, with an average abnormality or normality accuracy rate of 98.27% for 12 standard laboratory tests.

Although previous reports have shown to be effective in identifying unnecessary blood tests, none have used conditional entropy and pretest probability [16,17] to predict abnormal laboratory test results. However, as conditional entropy and pretest probability can measure the uncertainty of yielding an abnormal test for patients with different diagnoses, we hypothesize that performing feature engineering on these techniques can improve normal or abnormal laboratory test results. Feature engineering is not a common trend in data-driven approaches because of the capacity of deep learning models to learn complex and robust features from raw data. However, feature engineering is still necessary as using large amounts of raw data as input could also be a drawback as it is not always easy to obtain, clean, and process biomedical data [22]. Moreover, using additional laboratory tests as features can be

counterproductive if the goal is to reduce the number of laboratory tests.

Objectives

In this study, we adapted conditional entropy and pretest probability techniques to derive features to predict normal and abnormal laboratory test results. Our rationale is that by identifying whether the next laboratory test would yield a normal or abnormal result, medical professionals could decide on the necessity of such a test based on their experience and the patient's diagnosis and disease severity. To evaluate the effect of the inclusion of new types of features, we compared the performance of 2 machine learning approaches for predicting normal or abnormal laboratory test results on large-scale ICU data from Alberta, Canada. The difference between the 2 approaches was that only the second approach included new features based on conditional entropy and conditional probability.

Methods

Alberta ICU Database

This retrospective study was conducted using the Alberta ICU data set collected from 17 ICUs, comprising 55,689 ICU admissions from 48,672 deidentified unique patients admitted between February 2012 and December 2019. The primary data source was eCritical, an EMR-based data repository containing the device and laboratory data in use in all ICUs across Alberta.

Ethics Approval

The use of the ICU data set was approved by the Conjoint Health Research Ethics Board at the University of Calgary (reference number REB17-0389).

Selected Laboratory Blood Tests

We focused on 18 laboratory blood tests that are common and critical in the ICU (Table 1). The reference range to determine normality was determined using the Alberta Health Services guideline [23].

Table 1. Blood laboratory tests and reference normal ranges for tests selected for analysis.[a]

Laboratory test	Normal range	Total records, N
Potential of hydrogen: arterial (pH)	7.20-7.40	668,388
PaO_2[b] (mm Hg)	70-90	668,130
PCO_2[c] (mm Hg)	35-45	667,889
Blood potassium (mmol/L)	3.5-5.0	400,306
Hemoglobin (g/L)		
If male	140-175	398,436
If female	123-153	398,436
Blood sodium (mmol/L)		
If age (years) <90	136-145	396,431
If age (years) ≥90	132-146	396,431
Hematocrit (%)		
If male	0.42-0.50	395,046
If female	0.36-0.45	395,046
White blood cells (E+9 units/L)	4.5-11.0	394,809
Blood carbon dioxide content (mmol/L)		
If age (years) ≤60	23-29	390,906
If age (years) >60 and ≤90	23-31	390,906
If age (years) >90	20-29	390,906
Blood creatinine (μmol/L)		
If male and age (years) <60	80-115	370,361
If male and age (years) ≥60	71-115	370,361
If female and age (years) <60	53-97	370,361
If female and age (years) ≥60	53-106	370,361
Blood urea (μmol/L)		
If male and age (years) ≤55	3.0-9.0	295,445
If male and age (years) >55	3.0-8.0	295,445
If female and age (years) ≤55	3.0-8.0	295,445
If female and age (years) >55	2.0-7.0	295,445
Random glucose (mmol/L)	3.3-11.0	225,627
Alanine transaminase (U/L)		
If male	0-60	136,552
If female	0-40	136,552
Total bilirubin (μmol/L)	1.71-20.5	133,806
Alkaline phosphatase (U/L)	40-120	128,773
Blood albumin (g/L)	30.0-45.0	102,923
Aspartate aminotransferase (U/L)		
If male	10-40	98,399
If female	9-32	98,399
Gamma-glutamyl transferase (U/L)		
If male	0-80	36,095
If female	0-50	36,095

[a]For some laboratory tests, the reference values depend on patients' sex and age [23].

[b]PaO_2: partial pressure of oxygen (arterial).

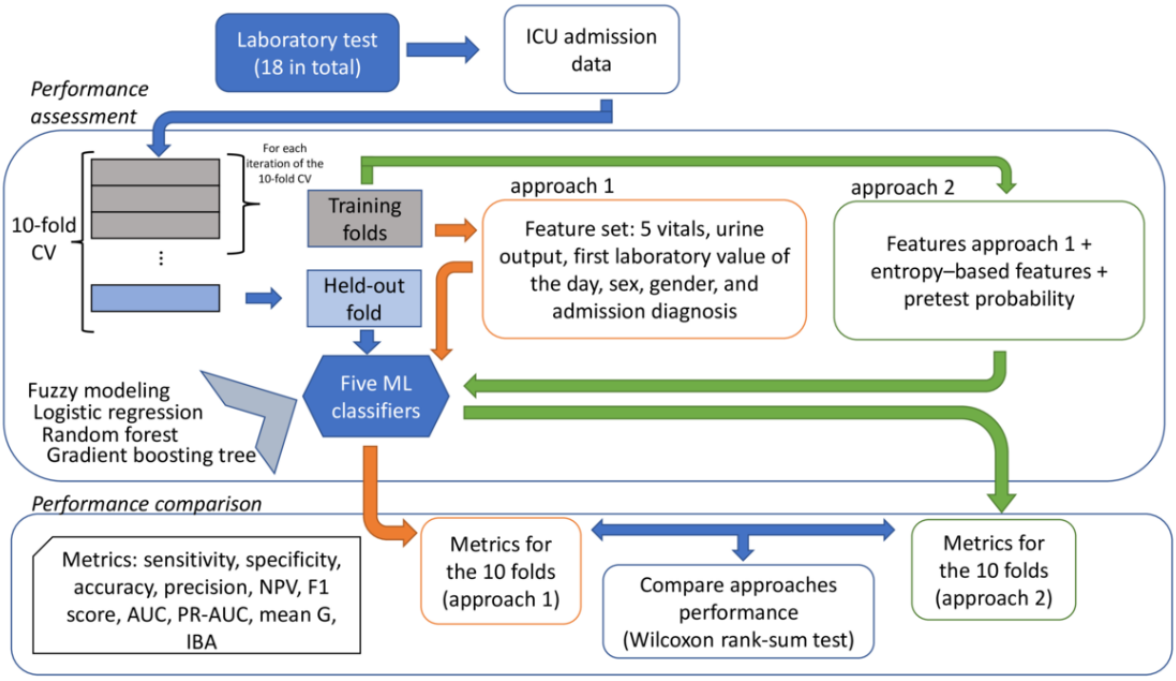
[c]PCO_2: partial pressure of carbon dioxide (arterial).

Framework Overview

This study compared 2 approaches to predict normal and abnormal blood laboratory tests performed in the ICU. The prediction was performed for all laboratory tests except for those first performed on the day, whose value was used as a feature in both approaches.

Figure 1 shows an overview of the framework used to compare the 2 different approaches. Inspired by the study by Cismeondi et al [18], approach 1 used heart rate, respiration rate, heart rate, blood pressure, temperature, pulse oximeter, respiratory rate, and urine output to perform the classification. We also included additional features, namely, sex, age, and admission diagnosis. In addition to all the features from approach 1, approach 2 included the adaptation of conditional entropy and pretest probability. The following sections explain in detail the different stages of these 2 approaches. The code used for performing the comparison between the approaches is publicly available in a public repository [24]; however, our data cannot be shared because of health care regulations.

Figure 1. A framework to compare the two redundancy detection approaches. AUC: area under the curve; CV: cross-validation; IBA: index balanced accuracy; ICU: intensive care unit; ML: machine learning; NPV: negative predictive value; PR-AUC: precision-recall area under the curve.



Inclusion Criteria

ICU admissions that meet the following inclusion criteria were included in the study: aged >18 years; at least one measurement of each of heart rate, respiration rate, blood pressure, temperature, oxygen saturation, and urine output; and ≥2 orders for at least one of the 18 laboratory blood tests in Table 1.

ICU admissions satisfying these inclusion criteria were split into 10 folds. Laboratory tests from the same admission were assigned to the same fold, ensuring that ICU admissions were mutually exclusive among the folds.

Features for Approach 1

Patients' Vitals, Demographics, and Admission Diagnoses

In approach 1, the variables used to predict the abnormal results of the next laboratory blood test were heart rate (beats per minute), oxygen saturation (%), respiration rate (breaths per minute), temperature (°C), blood pressure (mm Hg), and total amount of urine void (mL). These measurements were selected as bedside monitors commonly collect large quantities of these vitals regardless of patients' admission diagnosis. We also included patients' sex, age, and admission diagnosis. Age and sex were included as they affect the normality of the laboratory test results (Table 1). Age corresponded to the patient's age in years at ICU admission. The admission diagnosis was also included as patients in the ICU have a diverse set of underlying diagnoses; therefore, such a feature may affect laboratory test results. Categorical variables (sex and admission diagnosis) were coded using an approach that maps categories into numeric data using entropy, as presented in the study by Lopez-Arevalo et al [25].

Preprocessing Patients' Vitals

Owing to different sampling rates, laboratory blood tests and patients' vital measurements do not always occur simultaneously. We corrected the misalignment between

laboratory tests and patients' vitals following the steps in the study by Cismondi et al [26]. Specifically, for each admission, we fitted a cubic spline interpolation for each of the 6 patient variables (heart rate, oxygen saturation, respiration rate, temperature, blood pressure, and urine output). The patients' vitals were then estimated at the time of the laboratory tests. The interpolation procedure used neither the laboratory test values nor their class (normal or abnormal), thus avoiding any data leakage caused by using the target predictors to preprocess the data. Moreover, as the imputation was performed per ICU admission and the 10 folds of the cross-validation procedure were mutually exclusive, imputed data were not shared between the training and test sets.

Features for Approach 2

Pretest Probability

The pretest probability was calculated as the conditional probability of yielding a normal value, given a specific number of previous consecutive laboratory tests were normal. This probability was calculated on the training admissions by following the procedure presented by Roy et al [17]. Specifically, for each admission, we counted the number of consecutive normal laboratory tests before performing a new test and noted whether the new test yielded a normal result. Then, the information across the admissions was summed up, and the pretest probability distribution for each laboratory test was calculated as follows:

Here, *countNormalTests* is a function that returns the total cases of laboratory tests yielding normal when *M* previous tests were already normal, and *countTests* is the total number of laboratory tests performed when *M* previous consecutive tests were normal.

The pretest probability distribution was calculated using only ICU admissions from the training set. The feature values for the held-out fold were calculated using the pretest probability distribution obtained with the training folds.

Conditional Entropy for Abnormal Laboratory Tests

Entropy measures the expected amount of information. The conditional entropy also measures the expected amount of information of a random variable, given the occurrence of a value of secondary random variables, described as follows:

Here, $P(Y_i, Z_j)$ is the probability of value Y_i occurring while value Z_j occurs, and $P(Z_j)$ is the probability of Z resulting in the possible value Z_j.

We adapted conditional entropy to measure the expected amount of information of a test result if a patient's features were already known. This conditional entropy was calculated for all the features of approach 1. The conditional entropy for each feature was calculated as follows:

Here, Z is any variable of the patient's vitals or age, z is a possible value for such a variable, and Y=normal and Y=abnormal indicate laboratory blood tests that yielded normal or abnormal results, respectively. The values most associated with a certain result (normal or abnormal) had lower entropy (ie, number of bits), whereas those associated with a more uncertain result had higher entropy.

To estimate the conditional probability distribution for each patient's feature, we grouped each feature into a histogram with a bin width defined by the Freedman-Diaconis rule [27] as follows:

Here, IQR(f) is the IQR for feature f, and N_z is the number of observations in feature f.

Similar to the pretest probability, the conditional entropy distribution was calculated using only ICU admissions from the training folds. For the held-out fold, values were obtained from the distribution derived from the training folds.

Classifiers

Overview

We used four different classifiers to perform the comparison between approaches 1 and 2: (1) fuzzy modeling, (2) logistic regression (LR), (3) RF, and (4) gradient boosting (GB) trees.

For all classifiers, the features of the training folds and the held-out fold set were standardized before training the models using minimum-maximum normalization to avoid any feature scale impact on the performance. Normalization was performed using the maximum and minimum values from the training set as a reference.

Fuzzy Model

Fuzzy models are classifiers that define rules to establish nonlinear relationships between a set of features and a response variable. In this study, we used the Takagi-Sugeno model [28], which defines rules composed of antecedents and consequences on the features as follows:

Here, x_p is the pth feature of sample x; A_{kp}^{C} is the membership function for the kth rule, the pth feature, and class C; and $d_k^C(x)$ and $f_k^C(x)$ are the discriminant and consequent for the kth rule and class C. The advantage of these rules is that they establish connectivity between the features to derive the target output. For example, a rule can state that if the heart rate is high and the first laboratory in the morning is low, the next laboratory test will be abnormal.

As multiple rules are derived for the data, they are aggregated for the final output using their degree of activation. The degree of activation of the kth rule for class C is given by the following equation:

Here,  is the membership function of the fuzzy set  . The final discriminant for each class is as follows:

More details about fuzzy modeling can be found in the study by Takagi and Sugeno [28].

The number of features included in each rule was selected using a wrapper feature selection method that iteratively evaluated whether adding a new feature improved the model classification performance [29]. The consequence of each rule was defined using the probabilistic approach presented in the study by van den Berg et al [30] as follows:

Here,  is the summation of the degree of activation of all the samples in the training set, and  is the summation of the degree of activation of only the samples belonging to class *C*. The fuzzy model was implemented using the Python libraries *scikit-fuzzy* [31] and *pyFume* [32].

Machine Learning Models

Machine learning classifiers included the LR, RF, and GB tree models. The model parameters were tuned using nested cross-validation on the grid search and defined as follows:

- For LR, the grid search for the inverse of regularization strength (*C*) was defined as {0.1, 1.0, 10.0}.
- For RF, the grid search for the number of trees was defined as {300, 500, 800}, the number of maximum splits (tree height) was defined as {8, 15, 25}, the number of minimum samples to split was defined as {5, 10}, the number of maximum samples in leaves was defined as {2, 5}, and the number of maximum features was defined as {*sqrt*, log_2, None}.
- For the GB tree, the grid search for the learning rate was defined as {0.01, 0.05, 0.10}, the number of trees was defined as {300, 500, 800}, and the number of maximum features was defined as {*sqrt*, log_2, None}.

The best parameters were used to retrain a model using all data from the training folds and then test the held-out fold. The models were implemented using the *sklearn* Python library [33].

Measuring Performance

Table 2 shows the metrics used for assessing the performance of approaches 1 and 2. A total of 10 metrics were included to measure the different aspects of the approaches. Specificity, sensitivity, accuracy, and AUC measured the raw performance without considering class imbalances. In contrast, F_1 score, AUC, precision-recall AUC, mean G, and index balanced accuracy (IBA) are less sensitive to class imbalance, thereby providing a less biased performance for assessing the approaches.

The metrics also allow the comparison of the approaches from a medical perspective. Sensitivity indicates the proportion of actual abnormal laboratory tests that were correctly classified, whereas specificity indicates the proportion of actual normal laboratory tests that were correctly classified. These 2 metrics are related to precision (positive predictive value) and negative predictive value. When the number of false positives (normal tests predicted as abnormal) increases, the specificity and precision metrics decrease. The same occurs with the sensitivity and negative predictive value metrics when the number of false negatives increases.

Table 2. Metrics used to measure the performance of approaches 1 and 2.

Metric	Equation	Description
Specificity	TN[a]/(FP[b] + TN)	The proportion of actual normal laboratory tests that were correctly classified
Sensitivity (or recall)	TP[c]/(FN[d] + TP)	The proportion of actual abnormal laboratory tests that were correctly classified
Accuracy	(TP + TN)/(FN + FP + TP + TN)	The proportion of laboratory tests that were correctly classified
Precision (positive predictive value)	TP/(FP + TP)	The proportion of laboratory tests predicted as abnormal that, in fact, were abnormal
Negative predictive value	TN/(FN + TN)	The proportion of laboratory tests predicted as normal that, in fact, were normal
F_1 score	2 × (precision × recall)/(precision + recall)	Weighted mean of precision and recall
Area under the receiver operating characteristic curve		The balance between the true positive rate and true negative rate of the predictions
Area under the precision-recall curve		The balance between the precision and recall of the predictions
Mean G	√(sensitivity × specificity)	The balance between the performance of majority and minority classes
Index balanced accuracy [34]	(mean G)^2 × (1 + [sensitivity – specificity])	Imbalanced index of the overall accuracy

[a]TN: true negative.

[b]FP: false positive.

[c]TP: true positive.

[d]FN: false negative.

Comparing the 2 Approaches

The sets of metrics for each approach were compared pairwise using a 2-sided Wilcoxon rank-sum hypothesis test. The null hypothesis was that there was no difference between the metrics obtained using the 2 approaches, whereas the alternative hypothesis was that there was a difference. As 720 comparisons were conducted for the 18 laboratory tests, 4 classifiers, and 10 metrics, we used Benjamin-Hochberg correction with the false-positive rate set at 0.05.

Relevant Features

In addition to comparing the performances of the approaches, we explored the most relevant features for classification. For each iteration of the 10-fold cross-validation, we stored the relevance of each feature for the trained model.

For each classifier, features were ranked based on their relevance values. For the fuzzy model, relevance was given by the wrapper feature selection method used to derive the antecedent of the fuzzy rules. For LR, the relevance was given by the absolute value of the coefficient associated with each feature. For the RF and GB tree, the relevance was calculated using the mean of the impurity reduction within each tree of the fitted models.

After performing the 10-fold cross-validation, a total of 10 ranking feature sets were obtained for each laboratory blood test and each classifier. We aggregated these ranking feature sets by averaging the rank of each feature, which is an aggregation strategy used in the medical domain [35,36]. Specifically, we first averaged the rank of each feature across the folds. We then aggregated the ranking features by averaging the feature rank over the classifiers. As a result, we obtained an aggregated ranking feature set for each laboratory blood test.

Comparison Using Individual Features

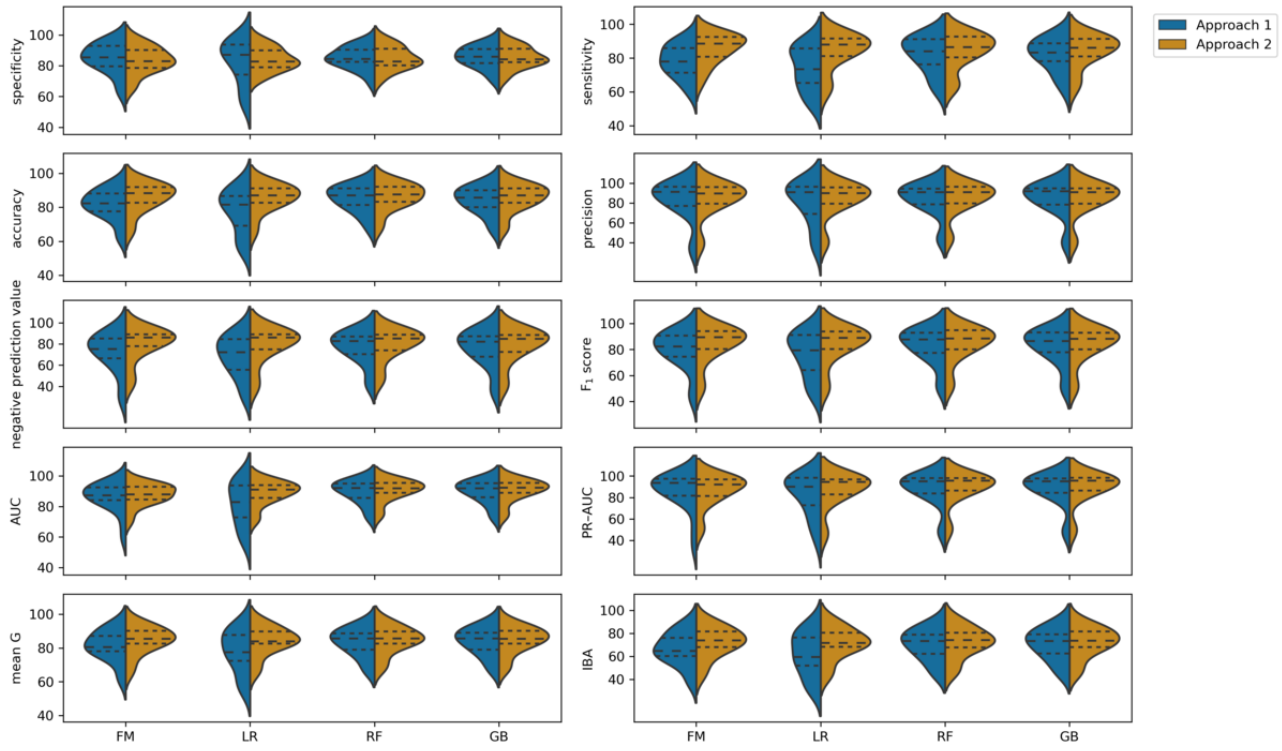
To compare the performance obtained with each new feature, we compared approach 1 with the 2 alternative approaches. The first alternative approach used the features of approach 1 plus the pretest probability features, whereas the second alternative approach used the features of approach 1 plus the entropy-based features. These alternative approaches were trained and compared with the same methodology used for approaches 1 and 2.

Results

Performance of the Approaches

Figure 2 shows the average performance of approaches 1 and 2 across the 18 laboratory tests. Both approaches suitably predicted the laboratory blood test results. Approach 1 achieved a median performance of at most 80% for all classifiers in 5 out of the 10 metrics (specificity, accuracy, precision, AUC, and precision-recall AUC), whereas approach 2 achieved a median performance >80% for all classifiers in all metrics except IBA. Notably, higher values (>80%) in the average performance for the F_1 score, mean G, and AUC suggest that approach 2 led to a more accurate prediction of both normal and abnormal results for most of the 18 blood laboratory tests.

Figure 2. Performance distribution of approaches 1 and 2 across the laboratory blood tests. The first quantile, median, and third quantile are displayed inside each distribution (dashed lines). AUC: area under the curve; FM: fuzzy model; GB: gradient boosting; IBA: index balanced accuracy; LR: logistic regression; PR-AUC: precision-recall area under the curve; RF: random forest.



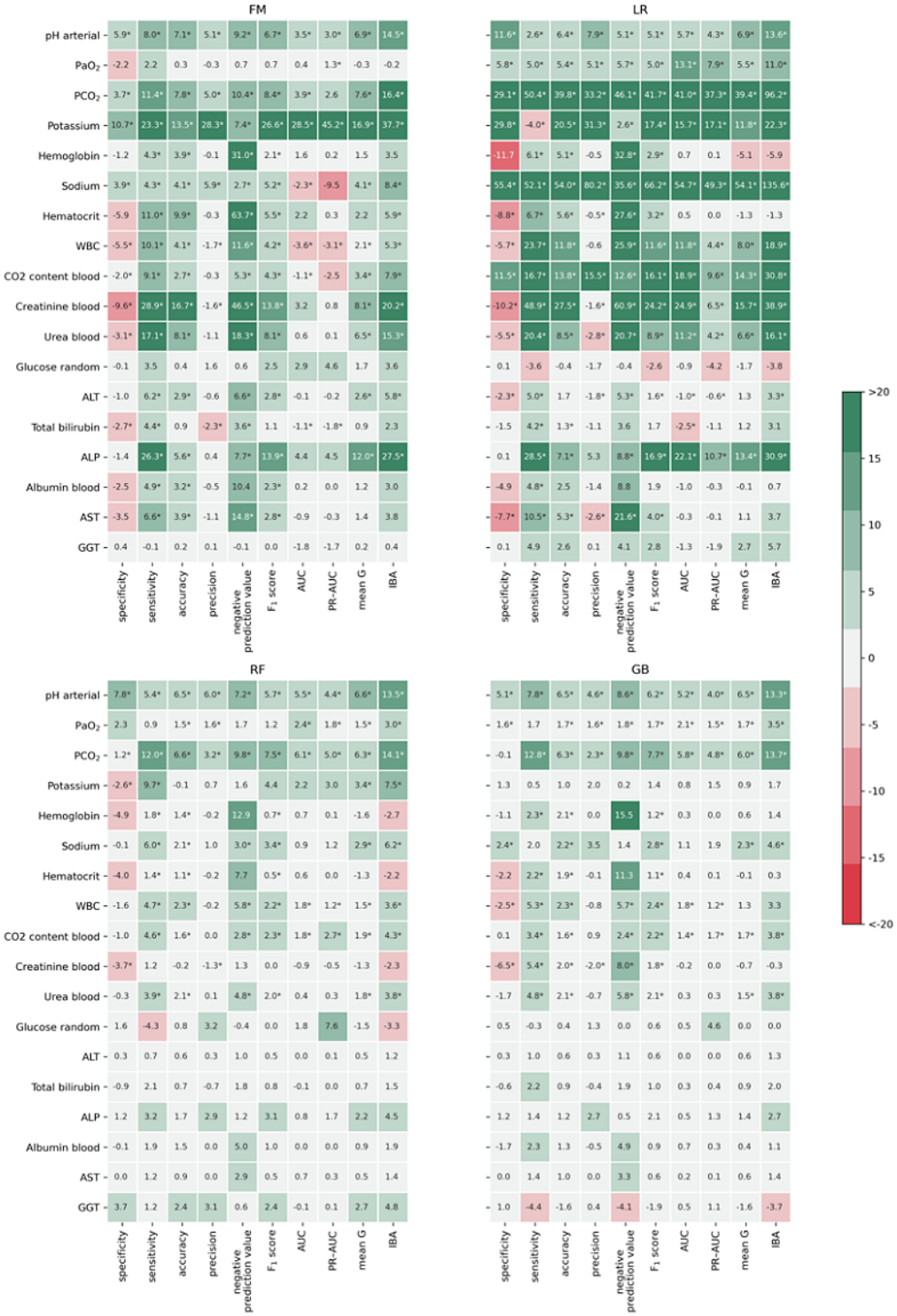
The detailed performance of the approaches for each laboratory test, metric, and machine learning classifier is presented in Multimedia Appendices 1 and 2. For both approaches, the machine learning classifiers achieved similar performance. The ensemble classifiers (ie, RF and GB) achieved the best overall performance across laboratory blood tests.

Comparison Between Old and New Features

Figure 3 shows the percentage change between approaches 1 and 2 for the 10-fold mean of each metric. The inclusion of the new features resulted in significant differences for most of the metrics in favor of approach 2. The metric that improved the most was sensitivity, achieving a significant improvement between 8 and 15 laboratory tests for the different classifiers. Specificity, in contrast, was the metric with the lowest improvement, with a significant reduction between 2 and 5 for the different classifiers. The F_1 score, mean G, and IBA, which are balanced performance metrics, significantly improved across the classifiers, for 8 to 14, 6 to 10, and 6 to 11 laboratory tests, respectively. A detailed comparison of approaches 1 and 2 for each blood laboratory test, metric, and classifier is presented in Multimedia Appendix 3.

Among the classifiers, LR benefited the most from the inclusion of the new features, achieving an improvement of at least eight metrics for 10 out of the 18 laboratory blood tests. The RF and GB obtained less significant improvements for the different metrics than the fuzzy and LR models.

Figure 3. Percentage change for the 10-fold mean metric values between approaches 1 and 2. The asterisk indicates a statistically significant difference (2-sided Wilcoxon rank-sum hypothesis tests adjusted via Benjamin-Hochberg correction using a false-positive rate set at 0.05). ALP: alkaline phosphatase; ALT: alanine transaminase; AST: aspartate aminotransferase; AUC: area under the curve; FM: fuzzy model; GB: gradient boosting; GGT: gamma-glutamyl transferase; IBA: index balanced accuracy; LR: logistic regression; PaO_2: partial pressure of oxygen (arterial); PCO_2: partial pressure of carbon dioxide (arterial); PR-AUC: precision-recall area under the curve; RF: random forest; WBC: white blood cell.



Most Relevant Features

Figure 4 shows the top 5 features selected across the classifiers to discriminate between abnormal and normal laboratory blood tests for approach 2. The most common relevant feature across the laboratory tests was the pretest probability, which ranked in the first 2 positions in 17 of the 18 laboratory tests. The first value of the day was also relevant for classification, ranking in the first 2 places for half of the blood laboratory tests. The conditional entropy variant of the features, such as diagnosis, urine output entropy, respiratory entropy, and heart rate entropy,

appeared more frequently in the top 5 ranking than their base forms.

Finally, to visualize how the features relate to the prediction of abnormal test results, the fuzzy predictive rules obtained by retraining a fuzzy model on the data set are presented in Multimedia Appendix 4.

Figure 4. The top 5 ranking of the features selected across the machine learning classifiers for each of the laboratory tests for approach 2. Light blue- and light red boxes correspond to the vital features and diagnoses, respectively, shared with approach 1. The light green boxes correspond to the pretest probability feature, and the light gray boxes correspond to the entropy-based features. ALP: alkaline phosphatase; ALT: alanine transaminase; AST: aspartate aminotransferase; GGT: gamma-glutamyl transferase; PaO_2: partial pressure of oxygen (arterial); PCO_2: partial pressure of carbon dioxide (arterial); SPO2: oxygen saturation; WBC: white blood cell.

Comparison Using Individual Features

Figure 5, Figure 6, Figure 7, and Figure 8 show the cubic root of the percentage change between the 10-fold means of approaches 1 and 2, approach 1 plus the pretest probability feature, and approach 1 plus the entropy-based features for the fuzzy modeling, LR, RF, and GB tree, respectively. For most of the laboratory tests, the percentage change was more consistent between approach 2 and approach 1 plus the pretest probability feature. Indeed, approach 1 plus the pretest probability feature obtained the same significant improvement that was achieved with approach 2 for almost all the laboratory tests. In contrast, approach 1 plus the entropy-based features showed a negative percentage change, particularly for fuzzy and logistic models.

Figure 5. Cubic root of the percentage change between the 10-fold means of approaches 1 and 2 (blue bars), approach 1 plus the pretest probability (yellow bars), and approach 1 plus the entropy-based features for the fuzzy model. The asterisk indicates a statistically significant difference (2-sided Wilcoxon rank-sum hypothesis tests adjusted via Benjamin-Hochberg correction using a false-positive rate set at 0.05). ALP: alkaline phosphatase; ALT: alanine transaminase; AST: aspartate aminotransferase; AUC: area under the curve; GGT: gamma-glutamyl transferase; IBA: index balanced accuracy; PaO_2: partial pressure of oxygen (arterial); PCO_2: partial pressure of carbon dioxide (arterial); PR-AUC: precision-recall area under the curve; WBC: white blood cell.

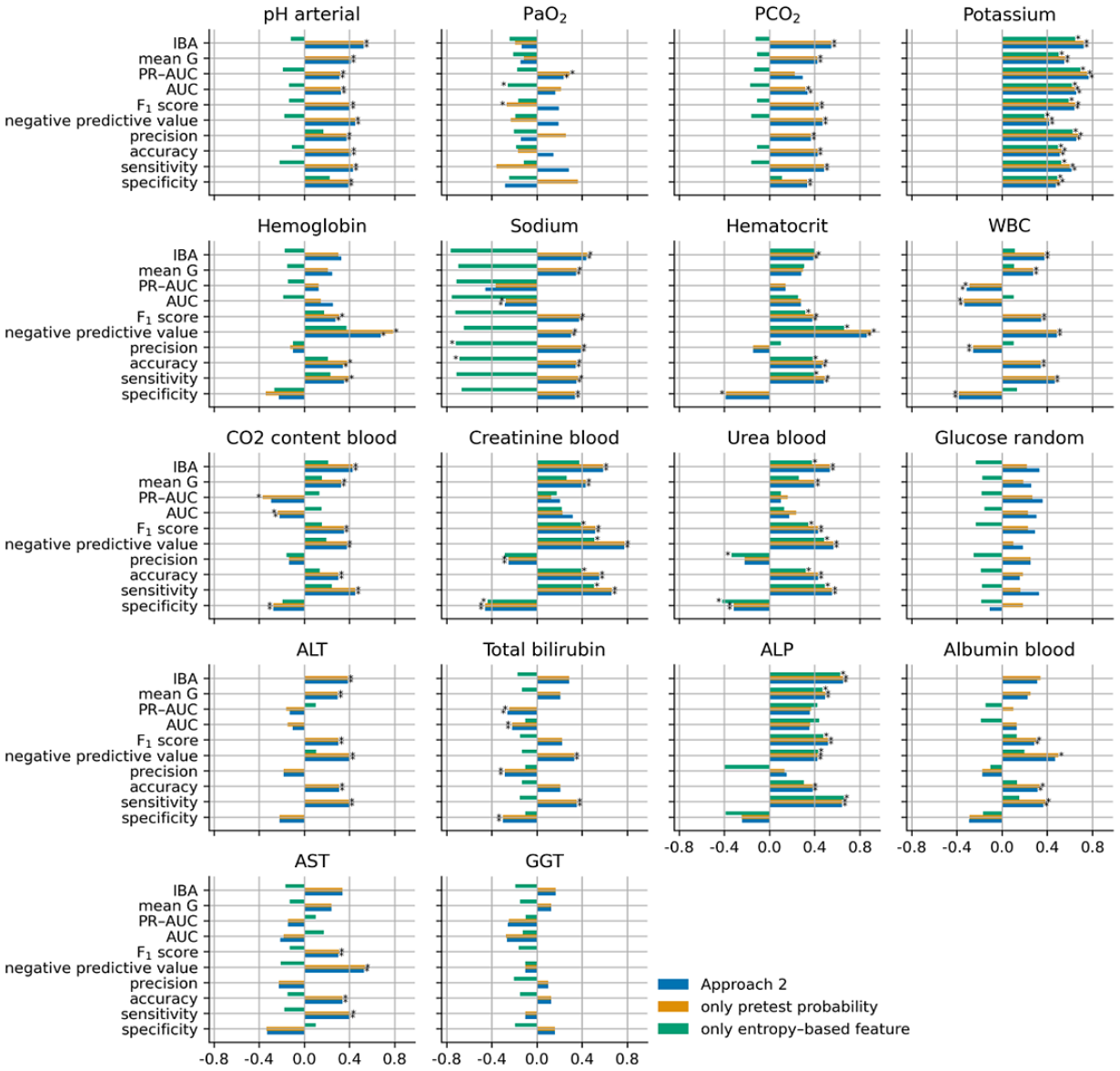


Figure 6. Cubic root of the percentage change between the 10-fold means of approaches 1 and 2 (blue bars), approach 1 plus the pretest probability (yellow bars), and approach 1 plus the entropy-based features for the logistic regression. The asterisk indicates a statistically significant difference (2-sided Wilcoxon rank-sum hypothesis tests adjusted via Benjamin-Hochberg correction using a false-positive rate set at 0.05). ALP: alkaline phosphatase; ALT: alanine transaminase; AST: aspartate aminotransferase; AUC: area under the curve; GGT: gamma-glutamyl transferase; IBA: index balanced accuracy; PaO_2: partial pressure of oxygen (arterial); PCO_2: partial pressure of carbon dioxide (arterial); PR-AUC: precision-recall area under the curve; WBC: white blood cell.

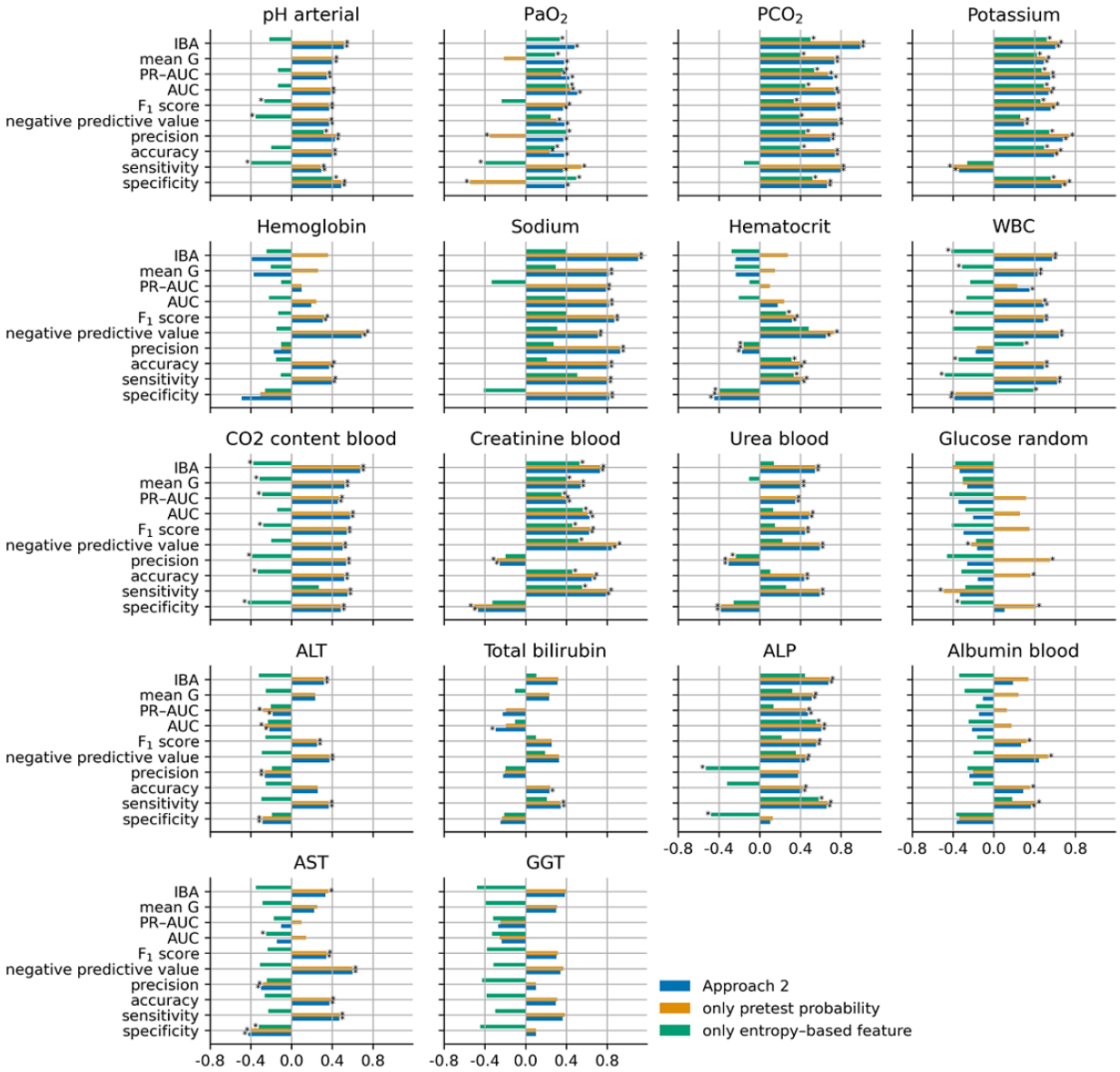


Figure 7. Cubic root of the percentage change between the 10-fold means of approaches 1 and 2 (blue bars), approach 1 plus the pretest probability (yellow bars), and approach 1 plus the entropy-based features for the random forest model. The asterisk indicates a statistically significant difference (2-sided Wilcoxon rank-sum hypothesis tests adjusted via Benjamin-Hochberg correction using a false-positive rate set at 0.05). ALP: alkaline phosphatase; ALT: alanine transaminase; AST: aspartate aminotransferase; AUC: area under the curve; GGT: gamma-glutamyl transferase; IBA: index balanced accuracy; PaO_2: partial pressure of oxygen (arterial); PCO_2: partial pressure of carbon dioxide (arterial); PR-AUC: precision-recall area under the curve; WBC: white blood cell.

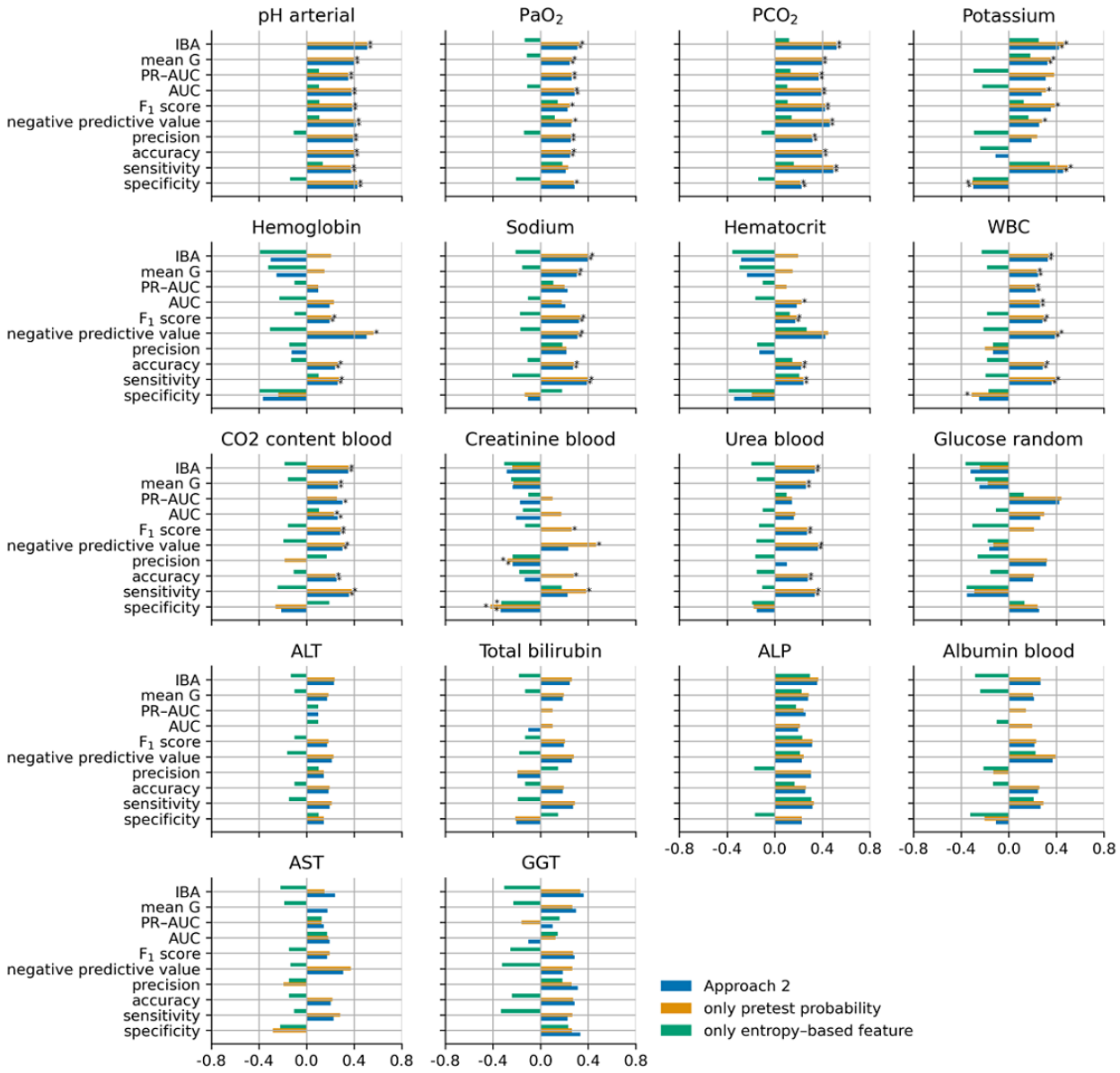
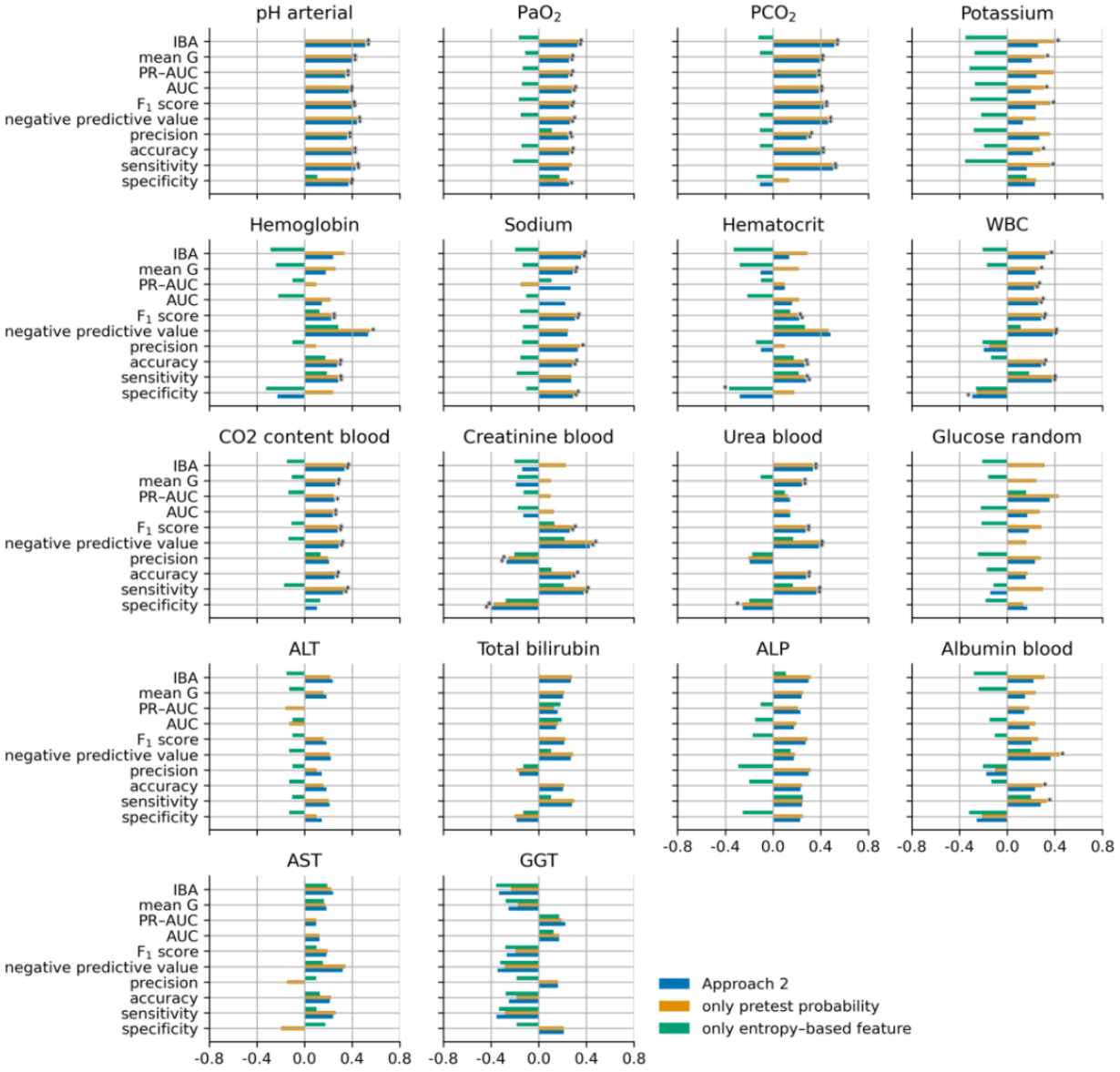


Figure 8. Cubic root of the percentage change between the 10-fold means of approaches 1 and 2 (blue bars), approach 1 plus the pretest probability (yellow bars), and approach 1 plus the entropy-based features for the gradient boosting model. The asterisk indicates a statistically significant difference (2-sided Wilcoxon rank-sum hypothesis tests adjusted via Benjamin-Hochberg correction using a false-positive rate set at 0.05). ALP: alkaline phosphatase; ALT: alanine transaminase; AST: aspartate aminotransferase; AUC: area under the curve; GGT: gamma-glutamyl transferase; IBA: index balanced accuracy; PaO_2: partial pressure of oxygen (arterial); PCO_2: partial pressure of carbon dioxide (arterial); PR-AUC: precision-recall area under the curve; WBC: white blood cell.



Discussion

Principal Findings

We found that the inclusion of the conditional entropy–based features and pretest probability significantly improved the capacity to predict abnormal results of a new laboratory test. Notably, the inclusion of these features improved the detection of actual abnormal tests (sensitivity) for half or more than half of the laboratory blood tests across the 4 classifiers (Figure 3).

The most relevant feature analysis revealed that the pretest probability feature was the most relevant among the new 2 types of features. In fact, the models strongly relied on the pretest probability to discriminate between normal and abnormal laboratory blood tests (Figure 4). A comparison of the performance of adding individual features further supports this fact by showing that approach 1 plus the pretest probability feature can achieve results comparable with those of approach 2.

The classifiers that improved the most were the LR and fuzzy models. A possible reason for this difference is that the LR and fuzzy models used all the features to fit their model. Instead, the ensemble models built individual trees by randomly selecting a subset of the total features, thereby excluding the pretest probabilities or entropy-based features for some trees. Nevertheless, the RF and GB tree also improved for approach

2, achieving significant improvements in the sensitivity, F_1 score, and IBA metrics.

The inclusion of the new features improved sensitivity and negative predictive value but decreased specificity and precision. This trade-off is beneficial for the medical context because although ordering extra blood tests when it may not be necessary (higher false positives) can raise the medical expenses, patients' safety is preserved (lower false negatives). The new features also improved balanced metrics such as F_1 score, AUC, mean G, and IBA, thus showing the benefit of the inclusion of such types of features to improve the capacity for discriminating between normal and abnormal test results. For instance, approach 2 improved the aforementioned metrics for blood gas tests (potential of hydrogen and PCO_2), which are among the most expensive laboratory blood tests ordered in the ICU [9].

However, we note that predicting normal and abnormal blood test results is an intermediate step toward detecting redundant tests. Deciding whether to order a new test should be based on more than predicting a normal laboratory result, as the situation and severity of each patient in the ICU are different. We included vitals and admission diagnosis to mitigate these factors; however, human interpretation still plays a crucial role in deciding whether ordering a new laboratory test is clinically meaningful. Normal laboratory test results can help measure trends, validate the required thresholds, and assess treatments. Therefore, predicting the result of a new test as normal does not imply its relevance or redundancy. However, redundancy guidelines can be established by analyzing predictions using prior consecutive results. For instance, if ≥1 previous result has yielded normal results and the prediction of the new test is again normal, the new laboratory blood test may be redundant. In contrast, if the prediction is abnormal, the new test may be relevant as it can inform medical decisions.

Relationship With Prior Reports

The relevance of the new features is consistent with prior literature [16,17], in which entropy and conditional probability were used to describe the high redundancy that exists in ICUs. In our work, we went further by adapting these to predict the abnormal results of performing a new test. Notably, to the best of our knowledge, no previous study has used these features to predict laboratory blood test results. This study also supports the work by Cismondi et al [18], showing that patients' vitals and the value of the first laboratory test performed in the day could guide the detection of abnormal results. However, unlike their study, we did not include their proposed blood transfusions to predict normal or abnormal laboratory test results as such transfusions targeted patients with gastrointestinal bleeding. In contrast, we extended the scope to different types of diagnoses by the inclusion of conditional entropy and pretest probability based on historical data.

In comparison with the studies by Roy et al [16] and Xu et al [20], who used machine learning to predict laboratory abnormal or normal results but did not include the pretest probability as a feature, approach 2 achieved comparable results using a smaller feature set (21 features vs 600 raw features). Specifically, the RF and GB tree achieved a mean AUC >0.89 for 13 out of the 18 laboratory tests (Multimedia Appendix 2 and Figure 2). This AUC improvement again shows the relevance of the inclusion of pretest probability as a feature in the predictive models.

Limitations

We note that this study used an ICU data set collected in Alberta, Canada. As ethnical and racial subgroups have different distributions for laboratory tests [37], ICU data sets collected in other countries may lead to different results, particularly in low- and middle-income countries whose populations deal with economic and cultural barriers that exacerbate their health challenges. However, this study introduced new features that rely on historical data, making these features flexible and applicable to different populations. Therefore, using historical data from a different population, the conditional entropy and pretest probability distributions can be derived to calculate the uncertainty of a new test that yields an abnormal result.

We also noted that our exclusion criteria excluded patients who did not have >1 sample of the target laboratory blood test or did not have any measurements for heart rate, respiration rate, temperature, oxygen saturation, blood pressure, or urine output. This condition limits the applicability of our work as it was not designed to predict abnormal results of the first laboratory test provided in the day or when the patient's vitals are missing. Future work should explore how to predict abnormal results of a new test in such cases.

Conclusions

This study introduced new types of features to predict abnormal or normal results in laboratory blood tests in the ICU. The new features were extracted from historical data to describe the chances of yielding a normal test if previous sequential tests were normal (pretest probability) and the expected uncertainty of an abnormal yield if a patient's vitals were already known (conditional entropy). These historical data combined with patients' data are suitable indicators to predict the abnormal results of performing an additional laboratory blood test. Therefore, this study provides tools that can help develop guidelines to reduce overtesting in the ICU.

Authors' Contributions

CEV and JL contributed to the development of the techniques and analysis tools. JL, HTS, and DJN coordinated the data collection process of the study and acquired funding. CEV implemented the methods presented herein and prepared the manuscript. CEV, DJN, HTS, and JL provided oversight throughout the study and proofread the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Performance of approach 1 using the 10-fold cross-validation for each blood laboratory test and machine learning classifier. For each classifier, the means and SDs of metrics across the 10 folds are presented. The best result for each metric and laboratory test is in bold.
[DOCX File , 35 KB - medinform_v10i6e35250_app1.docx]

Multimedia Appendix 2
Performance of approach 2 using the 10-fold cross-validation for each blood laboratory test and machine learning classifier. For each classifier, the means and SDs of metrics across the 10 folds are presented. The best result for each metric and laboratory test is in bold.
[DOCX File , 37 KB - medinform_v10i6e35250_app2.docx]

Multimedia Appendix 3
Percentage change for the 10-fold mean metric values between approach 1 and approach 2. The asterisk indicates a statistically significant difference (2-sided Wilcoxon rank-sum hypothesis tests adjusted with Benjamini-Hochberg using a false-positive rate set at 0.05). The difference was conducted for each classifier and for each metric.
[DOCX File , 36 KB - medinform_v10i6e35250_app3.docx]

Multimedia Appendix 4
Predictive rules for the fuzzy model. Entropy means that the feature is the conditional-based version.
[DOCX File , 15 KB - medinform_v10i6e35250_app4.docx]

References

1. Levinson W, Born K, Wolfson D. Choosing wisely campaigns: a work in progress. JAMA 2018 May 15;319(19):1975-1976. [doi: 10.1001/jama.2018.2202] [Medline: 29710232]
2. Kumwilaisak K, Noto A, Schmidt UH, Beck CI, Crimi C, Lewandrowski K, et al. Effect of laboratory testing guidelines on the utilization of tests and order entries in a surgical intensive care unit. Crit Care Med 2008 Nov;36(11):2993-2999. [doi: 10.1097/CCM.0b013e31818b3a9d] [Medline: 18824907]
3. Valentine SL, Bateman ST. Identifying factors to minimize phlebotomy-induced blood loss in the pediatric intensive care unit. Pediatr Crit Care Med 2012 Jan;13(1):22-27. [doi: 10.1097/PCC.0b013e318219681d] [Medline: 21499175]
4. McEvoy MT, Shander A. Anemia, bleeding, and blood transfusion in the intensive care unit: causes, risks, costs, and new strategies. Am J Crit Care 2013 Nov;22(6 Suppl):eS1-e14. [doi: 10.4037/ajcc2013729] [Medline: 24186829]
5. Salisbury AC, Reid KJ, Alexander KP, Masoudi FA, Lai SM, Chan PS, et al. Diagnostic blood loss from phlebotomy and hospital-acquired anemia during acute myocardial infarction. Arch Intern Med 2011 Oct 10;171(18):1646-1653. [doi: 10.1001/archinternmed.2011.361] [Medline: 21824940]
6. Thavendiranathan P, Bagai A, Ebidia A, Detsky AS, Choudhry NK. Do blood tests cause anemia in hospitalized patients? The effect of diagnostic phlebotomy on hemoglobin and hematocrit levels. J Gen Intern Med 2005 Jun;20(6):520-524 [FREE Full text] [doi: 10.1111/j.1525-1497.2005.0094.x] [Medline: 15987327]
7. Flabouris A, Bishop G, Williams L, Cunningham M. Routine blood test ordering for patients in intensive care. Anaesth Intensive Care 2000 Oct;28(5):562-565 [FREE Full text] [doi: 10.1177/0310057X0002800515] [Medline: 11094676]
8. Bates DW, Goldman L, Lee TH. Contaminant blood cultures and resource utilization. The true consequences of false-positive results. JAMA 1991 Jan 16;265(3):365-369. [Medline: 1984535]
9. Mikhaeil M, Day AG, Ilan R. Non-essential blood tests in the intensive care unit: a prospective observational study. Can J Anaesth 2017 Mar;64(3):290-295. [doi: 10.1007/s12630-016-0793-9] [Medline: 28000153]
10. Peixoto Jr AA, Meneses FA, Barbosa BP, Pessoa LF, Melo RH, Fideles GM. Laboratory routine in the ICU: a practice to be abolished? Crit Care 2013 Jun 19;17(S3):P12. [doi: 10.1186/cc12628]
11. Critical Care Societies Collaborative – Critical Care: Five Things Physicians and Patients Should Question. Choose Wisely. 2014 Jan 28. URL: https://www.choosingwisely.org/societies/critical-care-societies-collaborative-critical-care/ [accessed 2021-11-30]
12. Kleinpell RM, Farmer JC, Pastores SM. Reducing unnecessary testing in the intensive care unit by choosing wisely. Acute Crit Care 2018 Feb;33(1):1-6 [FREE Full text] [doi: 10.4266/acc.2018.00052] [Medline: 31723853]
13. Jalbert R, Gob A, Chin-Yee I. Decreasing daily blood work in hospitals: what works and what doesn't. Int J Lab Hematol 2019 May;41 Suppl 1:151-161. [doi: 10.1111/ijlh.13015] [Medline: 31069984]
14. Faisal A, Andres K, Rind JA, Das A, Alter D, Subramanian J, et al. Reducing the number of unnecessary routine laboratory tests through education of internal medicine residents. Postgrad Med J 2018 Dec;94(1118):716-719. [doi: 10.1136/postgradmedj-2018-135784] [Medline: 30670487]

15. Gruson D, Helleputte T, Rousseau P, Gruson D. Data science, artificial intelligence, and machine learning: opportunities for laboratory medicine and the value of positive regulation. Clin Biochem 2019 Jul;69:1-7. [doi: 10.1016/j.clinbiochem.2019.04.013] [Medline: 31022391]
16. Lee J, Maslove DM. Using information theory to identify redundancy in common laboratory tests in the intensive care unit. BMC Med Inform Decis Mak 2015 Jul 31;15:59 [FREE Full text] [doi: 10.1186/s12911-015-0187-x] [Medline: 26227625]
17. Roy SK, Hom J, Mackey L, Shah N, Chen JH. Predicting low information laboratory diagnostic tests. AMIA Jt Summits Transl Sci Proc 2018 May 18;2017:217-226. [Medline: 29888076]
18. Cismondi F, Celi LA, Fialho AS, Vieira SM, Reti SR, Sousa JM, et al. Reducing unnecessary lab testing in the ICU with artificial intelligence. Int J Med Inform 2013 May;82(5):345-358 [FREE Full text] [doi: 10.1016/j.ijmedinf.2012.11.017] [Medline: 23273628]
19. Mahani GK, Pajoohan M. Predicting lab values for gastrointestinal bleeding patients in the intensive care unit: a comparative study on the impact of comorbidities and medications. Artif Intell Med 2019 Mar;94:79-87. [doi: 10.1016/j.artmed.2019.01.004] [Medline: 30871685]
20. Xu S, Hom J, Balasubramanian S, Schroeder LF, Najafi N, Roy S, et al. Prevalence and predictability of low-yield inpatient laboratory diagnostic tests. JAMA Netw Open 2019 Sep 04;2(9):e1910967 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.10967] [Medline: 31509205]
21. Yu L, Li L, Bernstam E, Jiang X. A deep learning solution to recommend laboratory reduction strategies in ICU. Int J Med Inform 2020 Dec;144:104282. [doi: 10.1016/j.ijmedinf.2020.104282] [Medline: 33010730]
22. Yang S, Zhu F, Ling X, Liu Q, Zhao P. Intelligent health care: applications of deep learning in computational medicine. Front Genet 2021 Apr 12;12:607471 [FREE Full text] [doi: 10.3389/fgene.2021.607471] [Medline: 33912213]
23. DTH Meditech HCIS Reference Intervals. Alberta Health Services. 2018. URL: https://www.albertahealthservices.ca/assets/wf/lab/wf-lab-dth-meditech-hcis-reference-intervals.pdf [accessed 2021-11-30]
24. data-intelligence-for-health-lab/BloodLaboratoryTestRedundancy: Redundancy in Blood Laboratory Tests. Zenodo. 2022. URL: https://zenodo.org/record/6383240#.Ym79XvNBxpQ [accessed 2022-03-20]
25. Lopez-Arevalo I, Aldana-Bobadilla E, Molina-Villegas A, Galeana-Zapién H, Muñiz-Sanchez V, Gausin-Valle S. A memory-efficient encoding method for processing mixed-type data on machine learning. Entropy (Basel) 2020 Dec 09;22(12):1391 [FREE Full text] [doi: 10.3390/e22121391] [Medline: 33316972]
26. Cismondi F, Fialho AS, Vieira SM, Sousa JM, Reti SR, Howell MD, et al. Computational intelligence methods for processing misaligned, unevenly sampled time series containing missing data. In: Proceedings of the 2011 IEEE Symposium on Computational Intelligence and Data Mining. 2011 Presented at: CIDM '11; April 11-15, 2011; Paris, France p. 224-231. [doi: 10.1109/CIDM.2011.5949447]
27. Freedman D, Diaconis P. On the histogram as a density estimator:L2 theory. Z. Wahrscheinlichkeitstheorie verw Gebiete 1981 Dec;57(4):453-476. [doi: 10.1007/BF01025868]
28. Takagi T, Sugeno M. Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Man Cybern 1985 Jan;SMC-15(1):116-132. [doi: 10.1109/tsmc.1985.6313399]
29. Mendonça LF, Vieira SM, Sousa JM. Decision tree search methods in fuzzy modeling and classification. Int J Approx Reason 2007 Feb;44(2):106-123. [doi: 10.1016/j.ijar.2006.07.004]
30. van den Berg J, Kaymak U, van den Bergh WM. Fuzzy classification using probability-based rule weighting. In: Proceedings of the 2002 IEEE World Congress on Computational Intelligence. 2002 IEEE International Conference on Fuzzy Systems. 2002 Presented at: FUZZ-IEEE '02; May 12-17, 2002; Honolulu, HI, USA p. 991-996. [doi: 10.1109/fuzz.2002.1006639]
31. Warner J, Sexauer J, scikit-fuzzy, twmeggs, alexsavio, Unnikrishnan A, p2df, laurazh, alexbuy, et al. JDWarner/scikit-fuzzy: Scikit-Fuzzy version 0.4.2. Zenodo. 2019 Nov 14. URL: https://zenodo.org/record/3541386#.Ym7_oPNBxpQ [accessed 2022-05-13]
32. Fuchs C, Spolaor S, Nobile MS, Kaymak U. pyFUME: a Python package for fuzzy model estimation. In: Proceedings of the 2020 IEEE International Conference on Fuzzy Systems. 2020 Presented at: FUZZ-IEEE '20; July 19-24, 2020; Glasgow, UK p. 1-8. [doi: 10.1109/FUZZ48607.2020.9177565]
33. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
34. García V, Mollineda RA, Sánchez JS. Index of balanced accuracy: a performance measure for skewed class distributions. In: Proceedings of the 4th Iberian Conference on Pattern Recognition and Image Analysis. 2009 Presented at: IbPRIA '09; June 10-12, 2009; Póvoa de Varzim, Portugal p. 441-448. [doi: 10.1007/978-3-642-02172-5_57]
35. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007 Oct 01;23(19):2507-2517. [doi: 10.1093/bioinformatics/btm344] [Medline: 17720704]
36. Wu L, Hu Y, Liu X, Zhang X, Chen W, Yu AS, et al. Feature ranking in predictive models for hospital-acquired acute kidney injury. Sci Rep 2018 Nov 23;8(1):17298 [FREE Full text] [doi: 10.1038/s41598-018-35487-0] [Medline: 30470779]
37. Lim E, Miyamura J, Chen JJ. Racial/ethnic-specific reference intervals for common laboratory tests: a comparison among Asians, Blacks, Hispanics, and White. Hawaii J Med Public Health 2015 Sep;74(9):302-310 [FREE Full text] [Medline: 26468426]

Abbreviations

AUC: area under the curve
EMR: electronic medical record
GB: gradient boosting
IBA: index balanced accuracy
ICU: intensive care unit
LR: logistic regression
RF: random forest

Edited by C Lovis; submitted 28.11.21; peer-reviewed by W Zhang, A Banerjee, V Rajan, M Burns, J Chen; comments to author 13.02.22; revised version received 24.03.22; accepted 21.04.22; published 03.06.22.

Please cite as:
Valderrama CE, Niven DJ, Stelfox HT, Lee J
Predicting Abnormal Laboratory Blood Test Results in the Intensive Care Unit Using Novel Features Based on Information Theory and Historical Conditional Probability: Observational Study
JMIR Med Inform 2022;10(6):e35250
URL: https://medinform.jmir.org/2022/6/e35250
doi:10.2196/35250
PMID:35657648

Original Paper

Noninvasive Diagnosis of Nonalcoholic Steatohepatitis and Advanced Liver Fibrosis Using Machine Learning Methods: Comparative Study With Existing Quantitative Risk Scores

Yonghui Wu[1*], PhD; Xi Yang[1*], PhD; Heather L Morris[2], PhD; Matthew J Gurka[1], PhD; Elizabeth A Shenkman[1], PhD; Kenneth Cusi[3], MD; Fernando Bril[4], MD; William T Donahoo[3], MD

[1]Department of Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, United States

[2]Target RWE Health Evidence Solutions, Durham, NC, United States

[3]Department of Medicine, College of Medicine, University of Florida, Gainesville, FL, United States

[4]Division of Endocrinology, Diabetes and Metabolism, Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, United States

[*]these authors contributed equally

Corresponding Author:
William T Donahoo, MD
Department of Medicine
College of Medicine
University of Florida
1600 SW Archer Rd
Gainesville, FL, 32610
United States
Phone: 1 (352) 273 8655
Email: Troy.Donahoo@medicine.ufl.edu

Abstract

Background: Nonalcoholic steatohepatitis (NASH), advanced fibrosis, and subsequent cirrhosis and hepatocellular carcinoma are becoming the most common etiology for liver failure and liver transplantation; however, they can only be diagnosed at these potentially reversible stages with a liver biopsy, which is associated with various complications and high expenses. Knowing the difference between the more benign isolated steatosis and the more severe NASH and cirrhosis informs the physician regarding the need for more aggressive management.

Objective: We intend to explore the feasibility of using machine learning methods for noninvasive diagnosis of NASH and advanced liver fibrosis and compare machine learning methods with existing quantitative risk scores.

Methods: We conducted a retrospective analysis of clinical data from a cohort of 492 patients with biopsy-proven nonalcoholic fatty liver disease (NAFLD), NASH, or advanced fibrosis. We systematically compared 5 widely used machine learning algorithms for the prediction of NAFLD, NASH, and fibrosis using 2 variable encoding strategies. Then, we compared the machine learning methods with 3 existing quantitative scores and identified the important features for prediction using the SHapley Additive exPlanations method.

Results: The best machine learning method, gradient boosting (GB), achieved the best area under the curve scores of 0.9043, 0.8166, and 0.8360 for NAFLD, NASH, and advanced fibrosis, respectively. GB also outperformed 3 existing risk scores for fibrosis. Among the variables, alanine aminotransferase (ALT), triglyceride (TG), and BMI were the important risk factors for the prediction of NAFLD, whereas aspartate transaminase (AST), ALT, and TG were the important variables for the prediction of NASH, and AST, hyperglycemia (A_{1c}), and high-density lipoprotein were the important variables for predicting advanced fibrosis.

Conclusions: It is feasible to use machine learning methods for predicting NAFLD, NASH, and advanced fibrosis using routine clinical data, which potentially can be used to better identify patients who still need liver biopsy. Additionally, understanding the relative importance and differences in predictors could lead to improved understanding of the disease process as well as support for identifying novel treatment options.

(JMIR Med Inform 2022;10(6):e36997) doi:10.2196/36997

KEYWORDS

machine learning; nonalcoholic fatty liver disease; nonalcoholic steatohepatitis; fatty liver; liver fibrosis

Introduction

Obesity, metabolic syndrome, and type 2 diabetes have reached epidemic proportions, and these conditions are strongly associated with nonalcoholic fatty liver disease (NAFLD) [1]. Consequently, NAFLD has become the most common type of chronic liver disease in both adults and children [2,3]. Data from the National Health and Nutrition Examination Survey showed that the prevalence of NAFLD has increased from 20% in 1988-1994 to 28.3% in 1999-2004 to 33% in 2009-2012 and leveled off at 32% in 2013-2016 [4]. Although NAFLD as well as nonalcoholic steatohepatitis (NASH) and fibrosis can be reversed in many cases with weight loss, these diseases remain significantly underdiagnosed; a recent electronic health record analysis of almost 18 million adults in Europe found the prevalence of NAFLD and NASH to be only 1.85% [5]. NAFLD ranges from isolated steatosis to NASH and cirrhosis. Knowing the difference between the more benign isolated steatosis and the more severe NASH and cirrhosis informs the physician regarding the need for more aggressive management. Unfortunately, these can only be distinguished through an invasive liver biopsy. As liver biopsies are associated with various complications and high expenses, there is an increasing interest in developing noninvasive methods to determine the stage of NAFLD [6].

Previous studies have explored several biomarkers as noninvasive surrogates, including markers of apoptosis [7], oxidative stress [8,9], and inflammation [10,11]. Several quantitative risk score calculators, such as the US Fatty Liver Index (US FLI) [12], aspartate aminotransferase-to-platelet ratio index (APRI) [13], and Fibrosis-4 (FIB-4) score [14], have been proposed and applied in clinical studies. These scores are easy and straightforward to calculate, yet they use data that are not routinely collected in the clinic (eg, the US FLI includes the waist circumference) or only use a limited number of variables (eg, APRI uses lab values for aspartate transaminase [AST] and platelets).

With the recent development of machine learning algorithms, we are now able to use clinical data in much more sophisticated ways. Perveen et al [15] applied a decision tree (DT) method to evaluate the risk of developing NAFLD in a Canadian population, where the onset of NAFLD is determined according to the clinical criteria, namely Adult Treatment Panel III. Islam et al [16] compared logistic regression (LR), random forests (RFs), and support vector machines (SVMs) for the prediction of fatty liver disease using gender, age, and 8 other variables from lab tests. Yip et al [17] compared LR, ridge regression, AdaBoost, and DT for NAFLD prediction using 6 predictors from routine clinical and laboratory variables.

Although machine learning methods have been applied to predict NAFLD, previous studies only focused on detecting NAFLD without discriminating between isolated steatosis and NASH, or advanced fibrosis. In addition, it is not clear how machine learning methods perform compared to existing quantitative calculators (eg, APRI) in predicting NASH or advanced fibrosis. Therefore, the aim of this project was to determine if machine learning algorithms could identify NASH or advanced liver fibrosis using commonly available clinical and biochemical data.

Methods

Data Set

Deidentified data from a NASH research database (KC) were used. Baseline data from a total of 492 participants who had been recruited from the general population as well as the hepatology and endocrinology clinics at the University of Florida in Gainesville, Florida, and the University of Texas Health Science Center at San Antonio in San Antonio, Texas, were included. Patients participating in this study were screened for NAFLD by routine chemistries and liver magnetic resonance spectroscopy. The final diagnoses of NASH and fibrosis staging were determined via a percutaneous liver biopsy. For collecting lab test data, the measurements were conducted at 1 point for each patient. All patients signed the informed consent form before participating in the study.

Variable Encoding

To use the clinical and laboratory variables in machine learning algorithms, we compared 2 encoding methods including (1) categorical encoding, where the continuous lab values were converted into clinically meaningful categories according to domain experts; and (2) continuous encoding, where the continuous values were directly used without categorization. The categorical variables (eg, gender) were directly used in both encoding methods.

Machine Learning Methods

We compared LR, DTs, RFs, SVMs, and gradient boosting (GB), 5 widely used machine learning algorithms, for the prediction of NAFLD, NASH, and advanced fibrosis. LR is a widely used statistical model that applies a logistic function to determine model dependency among variables. LR has been widely used in a number of clinical studies to assess associations or predict outcomes. In this study, we used LR as the baseline and compared it with other machine learning methods. DT and RFs are 2 tree-based machine learning methods that are widely used in data mining and machine learning. An SVM is a typical machine learning algorithm based on the large margin theory and has been applied to various prediction tasks. GB is a machine learning technology that produces a strong predictive model through ensembles of a number of weak models such as DTs. We implemented LR, DT, RFs and SVMs using the sciki-learn library [18] and implemented GB using the official XGBoost package.

Feature Importance Analysis Using SHAP (SHapley Additive ExPlanations)

We also evaluated the important variables contributing to the prediction to examine how machine learning methods work using the SHAP method [19]. We used the feature importance, summary plot, and decision plot in SHAP to examine these variables. SHAP feature importance is a global importance score derived from the averaged absolute Shapley values per feature across the data set. Features with high SHAP importance are more influential for model prediction. The SHAP summary plot combines feature importance with feature effects. In a summary plot, each point is a Shapley value for a feature and an instance. The position on the y-axis is determined by the feature (ranked by the feature importance) and that on the x-axis by the Shapley value (positive or negative impact on model prediction). The color represents the feature value from low to high (red for high and blue for low). The summary plot is typically used to interpret the feature-model prediction association (positive or negative). The SHAP decision plot is used to show how features influence the models' decision-making for individual samples. In a typical decision plot, there is a straight gray line indicating the model's base value (starting point) and a colored line indicating prediction. Starting at the bottom of the plot, the prediction line shows how the SHAP values (ie, feature effects) accumulate from the base value to arrive at the model's final score at the top of the plot. Thus, we can interpret which sets of features determine the model prediction results quantitatively. In this study, we adopted the decision plots for misclassification analysis.

Existing NAFLD Risk Score Calculators

We examined 3 existing risk score calculators for the staging of liver fibrosis, including APRI [13] ([AST / 40] / platelets × 100), FIB-4 score [14] ([age × AST] / [platelets × √ALT]), and NAFLD fibrosis score (NFS) [20] (–1.675 + [0.037 × age] + [0.094 × BMI] + [1.13 × diabetes] + [0.99 × AST/ALT ratio] – [0.013 × platelets] – [0.66 × albumin]). We excluded the US FLI [12], as the waist circumference is not routinely measured in clinical practice.

Experiments and Evaluation

For machine learning methods, we used 5-fold cross-validation and determined the area under the receiver operating characteristic curve (AUC or AUC-ROC) as the evaluation metric. In the 5-fold cross-validation, the 492 patients were divided into 5 equal groups. We trained the machine learning model using 5 groups and used the remaining group as the test set for prediction. We repeated this training/prediction procedure 5 times and shuffled the groups so that each group could get a chance to serve as the test set. The parameters of the machine learning methods were optimized according to the 5-fold cross-validation result (training curves shown in Figures S2 through S6 in Multimedia Appendix 1). Then, we calculated the specificity and sensitivity based on the Youden's J statistic (Youden index) [21,22] determined from the ROC curve along with the AUC using the prediction from the 5-fold cross-validation. To reduce the bias of random grouping, for each machine learning method, we repeated the 5-fold cross-validation 20 times using different random seeds and calculated the mean specificity, mean sensitivity, mean AUC, and 95% CI. For existing scoring algorithms (APRI, FIB-4, and NFS), we used the bootstrapping strategy 100 times (80% data each time) to calculate the mean specificity, sensitivity, and AUC. Then, we selected the best machine learning method and compared it with existing scoring algorithms for the prediction of fibrosis. The mean AUC was used as the primary score for evaluation. All statistically significant parameters were identified by conducting 2-tailed *t* tests.

Ethics Approval

This study was approved by the Institutional Review Board of the University of Florida (reference number: IRB201800923).

Results

Baseline characteristics are presented in Table 1, separating patients based on the presence or absence of NASH. Tables S1and S2 (see Multimedia Appendix 1) present the baseline characteristics based on the presence or absence of advanced fibrosis and NAFLD, respectively.

Table 2 shows the performance of the machine learning methods for NAFLD prediction. The GB model with continuous encoding of variables achieved the best mean AUC score of 0.9043 (derived by performing the 5-fold cross-validation 20 times). The RF model with the continuous encoding method also achieved a comparable mean AUC score of 0.9020. Subsequent statistical analysis showed no significant difference (P=.61) between RFs and GB. Both GB and RFs outperformed the LR with P<.001 indicating statistical significance.

Table 1. Baseline characteristics of patients with and without nonalcoholic steatohepatitis (N=492).

Characteristic	Patients with NASH[a] (n=198)	Patients without NASH (n=294)	*P* value[b]
Age, years, mean (SD)	55 ± 10	54 ± 11	.22
Males, n (%)	142 (72)	214 (73)	.88
Ethnicity, n (%)			<.001
Caucasian	109 (55)	126 (43)	
Hispanic	73 (37)	107 (36)	
African American	11 (5.5)	55 (19)	
Asian	3 (1.5)	4 (1)	
Indian	0 (0)	2 (1)	
Pacific Islander	2(1)	0(0)	
BMI, kg/m^2, mean (SD)	34.1 (4.7)	33 (5.5)	.02
SBP[c], mmHg, mean (SD)	134 (16)	134 (17)	.93
DBP[d], mmHg, mean (SD)	79 (10)	78 (10)	.57
Total cholesterol, mg/dL, mean (SD)	183 (44)	168 (38)	<.001
TG[e], mg/dL, mean (SD)	202 (148)	137 (85)	<.001
LDL-C[f], mg/dL, mean (SD)	106 (36)	98 (34)	.03
HDL-C[g], mg/dL, mean (SD)	39 (11)	43 (13)	<.001
A_{1c}[h], %	6.8 (1.3)	6.5 (1.2)	.004
AST[i], IU[j]/L, mean (SD)	47 (26)	28 (14)	<.001
ALT[k], IU/L, mean (SD)	64 (37)	37 (27)	<.001
Bilirubin, mg/dL, mean (SD)	0.9 (0.5)	0.8 (0.4)	.003
Platelets, 10^9/L, mean (SD)	257 (84)	237 (63)	.006
Albumin, g/L, mean (SD)	4.2 (0.3)	4.1 (0.4)	.005
TSH[l], mIU/L, mean (SD)	2.31 (1.51)	2.05 (2.41)	.14
FPG[m], mg/dL, mean (SD)	136 (39)	127 (40)	.01
Glucose tolerance (n, %)			<.001
Type 2 diabetes	144 (73)	181 (62)	
Impaired glucose tolerance	41 (21)	48 (16)	
Impaired fasting glucose	7 (3)	36 (12)	
Normal glucose tolerance	6 (3)	29 (10)	
Presence of metabolic syndrome, n (%)	191 (96)	247 (84)	<.001
Presence of dyslipidemia, n (%)	180 (91)	206 (70)	<.001
Use of blood pressure medications, n (%)	159 (80)	181 (62)	<.001
Use of statins, n (%)	103 (52)	154 (52)	.99
Use of metformin, n (%)	92 (46)	119 (40)	.22
Use of sulfonylurea, n (%)	45 (23)	65 (22)	.96

[a]NASH: nonalcoholic steatohepatitis.

[b]For continuous variables, the *P* values were calculated by the 2-sided *t* test using 2 independent variables with unequal population variances. For categorical variables, the *P* values were calculated using the chi-square test.

[c]SBP: systolic blood pressure.

[d]DBP: diastolic blood pressure.

[e]TG: triglyceride.
[f]LDL-C: low-density lipoprotein-cholesterol.
[g]HDL-C: high-density lipoprotein-cholesterol.
[h]A_{1c}: hyperglycemia
[i]AST: aspartate transaminase.
[j]IU: international units.
[k]ALT: alanine aminotransferase.
[l]TSH: thyroid-stimulating hormone.
[m]FPG: fasting plasma glucose.

Table 2. Performance of machine learning methods for prediction of nonalcoholic fatty liver disease.

Method and feature encoding	Mean sensitivity	Mean specificity	Mean AUC[a] (95% CI)
Logistic regression			
Categorical	0.7631	0.8557	0.8632 (0.8560-0.8704)
Continuous	0.8232	0.8452	0.8786 (0.8716-0.8855)
Support vector machines			
Categorical	0.8013	0.8112	0.8599 (0.8523-0.8676)
Continuous	0.7773	0.8245	0.8524 (0.8455-0.8594)
Decision tree			
Categorical	0.7297	0.7796	0.7932 (0.7835-0.8029)
Continuous	0.7888	0.7809	0.8078 (0.7974-0.8183)
Random forests			
Categorical	0.7811	0.8602	0.8782 (0.8717-0.8848)
Continuous	0.8250	0.8595	0.9020 (0.8957-0.9083)
Gradient boosting			
Categorical	0.7895	0.8380	0.8686 (0.8615-0.8756)
Continuous	0.8343	0.8694	*0.9043 (0.8979-0.9107)*

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

Table 3 compares the performance of the machine learning models in the prediction of NASH. The GB model with continuous encoding achieved the best mean AUC of 0.8166. The RF model with the continuous encoding method achieved a similar mean AUC score of 0.8119. Statistical comparisons between GB and RFs showed that P=.42, indicating no significant difference. Again, both GB and RFs significantly outperformed LR with P<.001 and P=.007, respectively.

Table 4 summarizes the performance of the machine learning methods in the prediction of advanced fibrosis. GB with the continuous encoding method achieved the best mean AUC of 0.8360. RFs with the continuous encoding method achieved a comparable mean AUC score of 0.8337, which is not significantly different from that of GB (P=.76). Although both GB and RFs outperformed LR in terms of the mean AUC score, subsequent statistical tests showed no significant difference between them (P=.29 between GB and LR; P=.46 between RFs and LR).

Next, we compared the best machine learning method (GB with continuous variable) with existing scoring algorithms in predicting advanced fibrosis. Table 5 shows the comparison results. The GB model outperformed the 3 existing scoring algorithms with an averaged AUC of 0.8360 for advanced fibrosis with significant P values. Among the 3 existing scoring algorithms, APRI achieved the best performance with an averaged AUC of 0.7890 in predicting the outcome. The AUC-ROC curves are provided in Figure S1 of Multimedia Appendix 1.

Finally, we examined the importance scores of the top 10 variables for the disease states based on the SHAP values (see Table S2 in Multimedia Appendix 1). The top important variables for each condition were determined by the SHAP importance feature, which is defined as the mean absolution SHAP value. Figure 1 graphically demonstrates these results. For NAFLD, ALT was the most important variable (SHAP importance =1.02) followed by TG and BMI. For NASH, AST was the most important factor (SHAP importance=0.5) followed by ALT and TG. For advanced fibrosis, AST was the most important risk factor (SHAP importance=0.91) followed by hyperglycemia (A_{1c}) and HDL.

Table 3. Performance of machine learning methods in prediction of nonalcoholic steatohepatitis.

Method and feature encoding	Mean sensitivity	Mean specificity	Mean AUC[a] (95% CI)
Logistic regression			
Categorical	0.7244	0.7523	0.7858 (0.7769-0.7948)
Continuous	0.7070	0.7903	0.7956 (0.7871-0.8041)
Support vector machines			
Categorical	0.7383	0.7480	0.7924 (0.7813-0.7983)
Continuous	0.6836	0.8256	0.7968 (0.7886-0.8050)
Decision trees			
Categorical	0.7064	0.6693	0.7201 (0.7098-0.7304)
Continuous	0.6937	0.6881	0.7305 (0.7210-0.7401)
Random forests			
Categorical	0.6979	0.8041	0.7910 (0.7819-0.8001)
Continuous	0.7582	0.7691	0.8119 (0.8036-0.8215)
Gradient boosting			
Categorical	0.7226	0.7600	0.7914 (0.7827-0.8001)
Continuous	0.7525	0.7836	*0.8166 (0.8083-0.8249)*

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

Table 4. Performance of machine learning methods in prediction of advanced fibrosis.

Method and feature encoding	Mean sensitivity	Mean specificity	Mean AUC[a] (95% CI)
Logistic regression			
Categorical	0.7683	0.7730	0.7950 (0.7837-0.8063)
Continuous	0.8500	0.7428	0.8278 (0.8172-0.8392)
Support vector machines			
Categorical	0.7367	0.7587	0.7628 (0.7489-0.7767)
Continuous	0.8242	0.7320	0.8122 (0.8002-0.8233)
Decision tree			
Categorical	0.7467	0.8010	0.7844 (0.7651-0.8037)
Continuous	0.6667	0.7379	0.6947 (0.6740-0.7153)
Random forests			
Categorical	0.7425	0.8529	0.8118 (0.7985-0.8251)
Continuous	0.8325	0.7757	0.8337 (0.8227-0.8447)
Gradient boosting			
Categorical	0.7492	0.8361	0.8115 (0.7977-0.8253)
Continuous	0.8083	0.8074	*0.8360 (0.8254-0.8467)*

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

Table 5. Comparison of gradient boosting (the best machine learning method) with existing scoring algorithms for prediction of advanced fibrosis[a].

Method	Mean sensitivity	Mean specificity	Mean AUC[b] (95% CI)	*P* value
GB[c]	0.8083	0.8074	0.8360 (0.8254-0.8467)	N/A[d]
APRI[e]	0.7424	0.7606	0.7984 (0.7964-0.8004)	<.001
FIB-4[f]	0.7176	0.6674	0.7394 (0.7371-0.7417)	<.001
NFS[g]	0.7506	0.5673	0.6843 (0.6777-0.6909)	<.001

[a]The scores for APRI, FIB-4, and NFS were calculated by bootstrapping 80% of the data from all 492 patients 100 times.

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

[c]GB: gradient boosting.

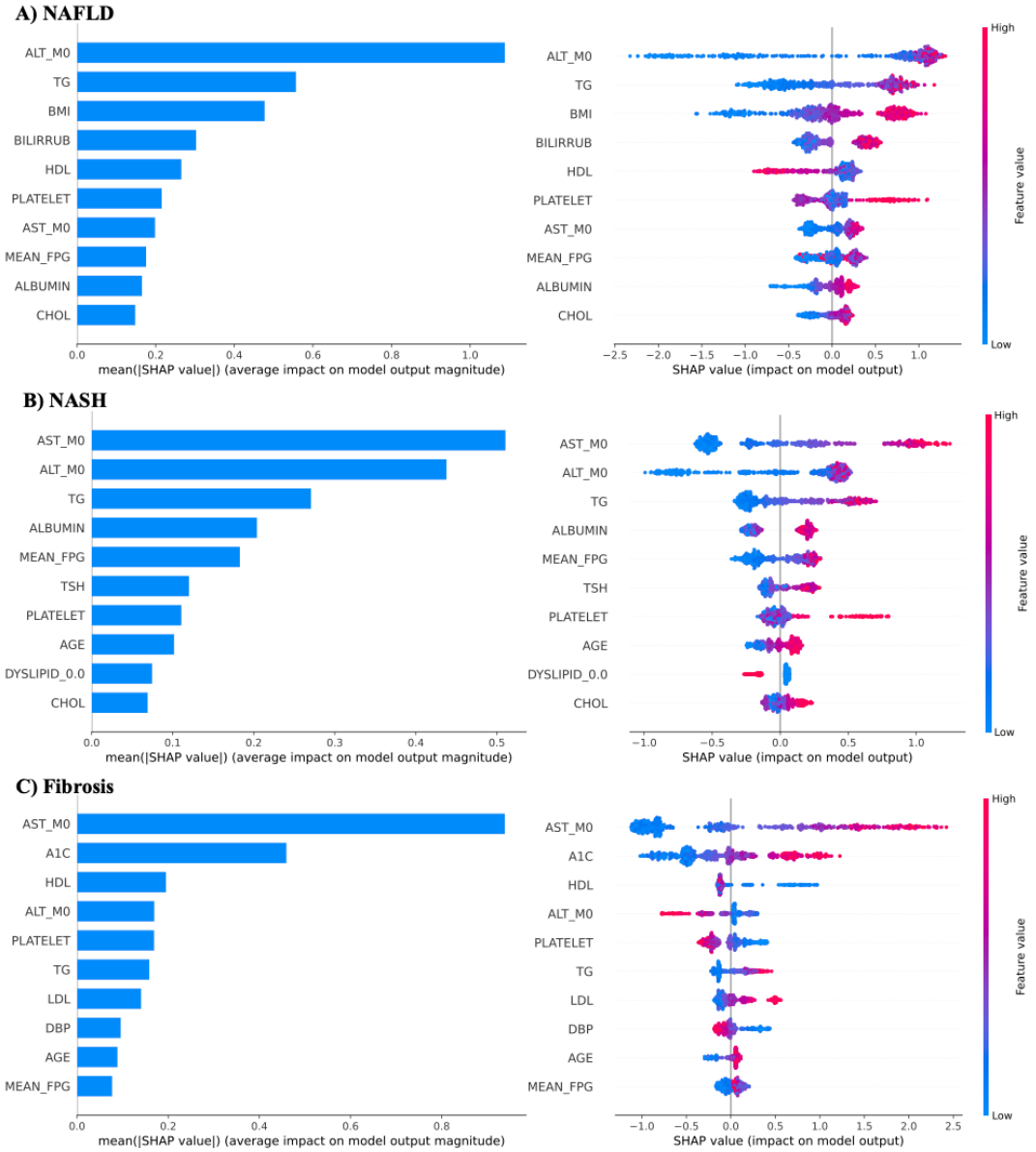
[d]N/A: not applicable.

[e]APRI: aspartate aminotransferase-to-platelet ratio index.

[f]FIB-4: Fibrosis-4.

[g]NFS: Nonalcoholic Fatty Liver Disease Fibrosis Score.

Figure 1. Top 10 important risk factors for prediction of NAFLD, NASH, and fibrosis based on SHAP importance calculated using the GB models with the continuous feature encoding method. (SHAP importance was derived from the averaged absolute SHAP values). A_{1c}: hyperglycemia; ALT: alanine aminotransferase; AST: aspartate transaminase; BILIRRUB: bilirubin; CHOL: cholesterol; DBP: diastolic blood pressure; DYSLIPID: dyslipidemia; FPG: fasting plasma glucose; GB: gradient boosting; HDL: high-density lipoprotein; LDL: low-density lipoprotein; NAFLD: nonalcoholic fatty liver disease; NASH: nonalcoholic steatohepatitis; TG: triglyceride; TSH: thyroid-stimulating hormone; SHAP: SHapley Additive exPlanations.



Discussion

Principal Findings

In this study, we systematically compared 5 machine learning algorithms for prediction of NAFLD, NASH, and advanced fibrosis using variables from routine lab tests and patients' demographics. We collected 33 variables from a total of 492 patients with NAFLD, NASH, and advanced fibrosis verified by liver biopsy. The experimental results show that the GB model achieved the best mean AUC scores of 0.9040, 0.8135, and 0.8360 for the prediction of NAFLD, NASH, and advanced fibrosis, respectively. This study demonstrated that it is feasible to use machine learning methods for noninvasive diagnosis of NAFLD, NASH, and advanced fibrosis.

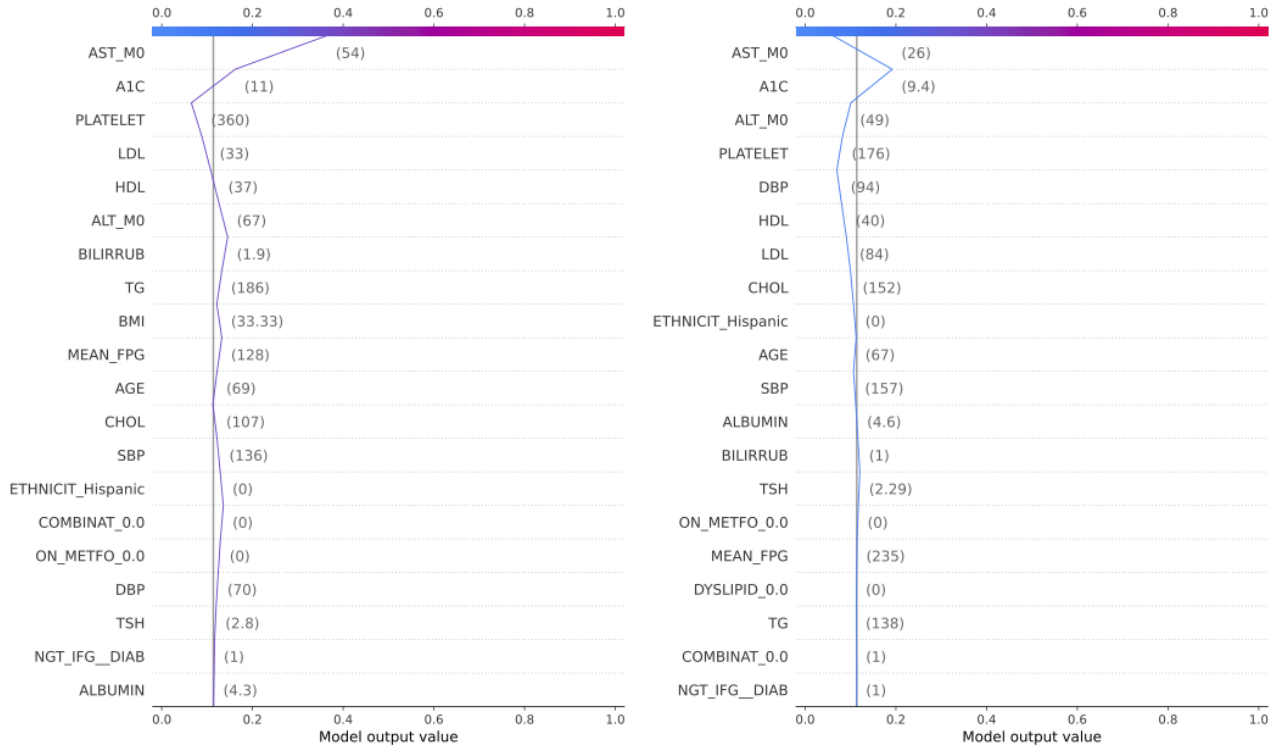
We compared the best machine learning model, GB, with 3 existing risk score calculators (APRI, FIB-4, and NFS) and the comparison results showed that GB significantly outperformed the existing calculators in identifying fibrosis by leveraging more patient variables. Even though APRI is a simple calculator defined using only AST and Platelet, it achieved a decent performance in identifying fibrosis cases with a relatively small margin (~4%) compared to GB. Existing risk score calculators are defined using a limited number of variables; therefore, they are straightforward to calculate and easy to use in clinical settings. On the other hand, machine learning methods can achieve better performance by leveraging more variables from patients. The GB model significantly outperformed FIB-4 and NFS recommended in recent guidelines, indicating the potential use of machine learning models as screening tools for improved identification of advanced fibrosis in clinics.

To use the variables in machine learning methods, we compared 2 encoding methods including continuous encoding and categorical encoding. Categorical encoding used domain expert knowledge to categorize the continuous lab test values into different clinically meaningful categories (eg, low, normal, and high). In contrast, continuous encoding is purely a data-driven approach, using the lab values as they are and leaving the machine learning models to learn the cutoffs. The experimental results show that continuous encoding is better for representing lab values in machine learning methods.

To understand how the GB model predicts NAFLD, NASH, and advanced fibrosis, we examined the top 10 important features, as shown in Figure 1. For NAFLD (Figure 1A), the findings make clinical sense with ALT as the most important risk factor, followed by obesity (BMI) and an indirect measure of steatosis such as TG and HDL, which are inversely related to NAFLD in the SHAP summary plot (Figure 1A right). As expected, other risk factors were also positively associated with NAFLD. For example, a high ALT indicates a high probability of NAFLD. This is consistent with clinical practice. For NASH (Figure 1B), AST is the most important feature followed by ALT with a SHAP importance value comparable to that of AST, which is also consistent with clinical practice. However, when compared to NAFLD, we identified 3 novel features in the top 10, including atherogenic dyslipidemia (TG), hyperglycemia (fasting plasma glucose), and thyroid hormone status (thyroid-stimulating hormone). Abnormalities in the hepatic thyroid hormone metabolism are gaining momentum as conditions that may be linked with the development of steatohepatitis [23]. Similar to NAFLD, many features (Figure 1B right) have positive associations with NASH. As anticipated, AST was the most important feature for advanced fibrosis (Figure 1C); however, A_{1c} was a novel factor related to the development of advanced liver fibrosis and the second most important one. Some studies have suggested a link between A_{1c} and diabetes and NASH [24,25], but the relationship of diabetes with the severity of steatohepatitis and fibrosis remains controversial [26]. Their relevance can be best appreciated in the summary plot (Figure 1C right). The order of these variables only provides correlative evidence and certainly not cause and effect; however, data such as these can also lead to the generation of hypotheses pertaining to the relative role of adiposity vs insulin resistance vs hyperglycemia in the progression of liver disease from NAFLD to NASH, and then to advanced fibrosis, and offer insights into the opportunities for future targeted therapies.

Figure 2 presents 2 error cases of the GB model in predicting advanced fibrosis. As for the false positive case (Figure 2A), this patient had no fibrosis according to the biopsy result (has NASH), but the model predicted fibrosis. The decision plot shows that the HDL (37 mg/dL), low-density lipoprotein (33 mg/dL), and Platelet (360K) of this patient are within the normal range, thus decreasing the SHAP value for fibrosis. However, the A_{1c} (11%) and AST (56 units per liter) of this patient are significantly higher than the normal range, which increased the SHAP value for fibrosis and led to the final predicted positive outcome. This observation is consistent with the feature importance analysis (Figure 1C) showing that A_{1c} and AST are strongly positively associated with the risk of advanced fibrosis. As for the false negative case (Figure 2B), the patient had advanced fibrosis determined from biopsy, but the GB model provided a negative prediction (no fibrosis). Although this patient has an A_{1c} of 9.4%, which increases the SHAP value, a normal AST (26 units per liter) significantly reduced the SHAP value and led to the final negative prediction outcome.

Figure 2. Decision plots for false positive and false negative prediction cases using the gradient boosting model with the continuous feature encoding method on advanced fibrosis. A_{1c}: hyperglycemia; AST: aspartate transaminase; ALT: alanine aminotransferase; BILIRRUB: bilirubin; CHOL: cholesterol; DBP: diastolic blood pressure; DIAB: diabetes; DYSLIPID: dyslipidemia; FPG: fasting plasma glucose; HDL: high-density lipoprotein; IFG: impaired fasting glucose; LDL: low-density lipoprotein; METFO: metformin; NGT: narrow gastric tube; SBP: systolic blood pressure; TG: triglyceride; TSH: thyroid-stimulating hormone.



A) False positive case

B) False negative case

Limitations

This study has limitations. First, the cohort in this study had 492 patients who were recruited at the University of Florida and the University of Texas Health Science Center. Future studies should examine our model using cohorts from different regions. Second, this study focused on 4 types or groups of medications as blood pressure medications, including statins, metformin, and sulfonylurea identified by the domain experts (physicians at the University of Florida). We plan to extend the data set and examine more medications (eg, obeticholic acid, pentoxifylline). Recent studies [27-30] showed that social determinants of health and environmental exposure are associated with the risk of liver diseases, which could be further explored.

Conclusions

This study shows that it is feasible to use machine learning algorithms to identify NAFLD, NASH, and advanced fibrosis using common clinically available data. Further validation using larger and more clinically diverse data sets is required. Using only clinically available data, this method can effectively target individuals most likely to benefit from a liver biopsy to diagnose advanced liver disease. Additionally, understanding the relative importance of and differences in predictors could lead to improved understanding of the disease process and provide better support for identifying novel treatment options.

Acknowledgments

The research reported in this paper was supported in part by a Patient-Centered Outcomes Research Institute (PCORI) Award (grant ME-2018C3-14754), the OneFlorida+ Clinical Research Network, the Patient-Centered Outcomes Research Institute (grants CDRN-1501-26692 and RI-CRN-2020-005); in part by the OneFlorida+ Cancer Control Alliance, funded by the Florida Department of Health's James and Esther King Biomedical Research Program (grant 4KB16); and in part by the University of Florida Clinical and Translational Science Institute, which is supported in part by the NIH National Center for Advancing Translational Sciences (grants UL1TR001427 and UL1TR000064). The content is solely the responsibility of the authors and does not necessarily represent the official views of the PCORI and its Board of Governors or Methodology, the OneFlorida+ Clinical Research Network, the University of Florida-Florida State University Clinical and Translational Science Institute, the Florida Department of Health, or the National Institutes of Health.

Authors' Contributions

YW, WTD, FB, HLM, and MJG were responsible for the overall design, development, and data evaluation of this study. FB collected the data for this study. XY, YW, FB, HLM, and WTD contributed to data analysis. YW, WTD, and FB did most of the writing. MJG, EAS, and KC were also involved in the writing and editing of this manuscript. All authors reviewed the manuscript critically for scientific content and gave final approval of the manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables and figures.
[PDF File (Adobe PDF File), 1863 KB - medinform_v10i6e36997_app1.pdf]

References

1. Angulo P, Lindor KD. Non-alcoholic fatty liver disease. J Gastroenterol Hepatol 2002 Feb;17 Suppl:S186-S190. [doi: 10.1046/j.1440-1746.17.s1.10.x] [Medline: 12000605]
2. Browning JD, Szczepaniak LS, Dobbins R, Nuremberg P, Horton JD, Cohen JC, et al. Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity. Hepatology 2004 Dec;40(6):1387-1395. [doi: 10.1002/hep.20466] [Medline: 15565570]
3. Adams LA, Angulo P, Lindor KD. Nonalcoholic fatty liver disease. CMAJ 2005 Mar;172(7):899-905 [FREE Full text] [doi: 10.1503/cmaj.045232] [Medline: 15795412]
4. Younossi ZM, Stepanova M, Younossi Y, Golabi P, Mishra A, Rafiq N, et al. Epidemiology of chronic liver diseases in the USA in the past three decades. Gut 2020 Mar;69(3):564-568. [doi: 10.1136/gutjnl-2019-318813] [Medline: 31366455]
5. Alexander M, Loomis AK, Fairburn-Beech J, van der Lei J, Duarte-Salles T, Prieto-Alhambra D, et al. Real-world data reveal a diagnostic gap in non-alcoholic fatty liver disease. BMC Med 2018 Aug;16(1):130 [FREE Full text] [doi: 10.1186/s12916-018-1103-x] [Medline: 30099968]
6. Alkhouri N, McCullough AJ. Noninvasive diagnosis of NASH and liver fibrosis within the spectrum of NAFLD. Gastroenterol Hepatol (N Y) 2012 Oct;8(10):661-668 [FREE Full text] [Medline: 24683373]
7. Alkhouri N, Carter-Kent C, Feldstein AE. Apoptosis in nonalcoholic fatty liver disease: diagnostic and therapeutic implications. Expert Rev Gastroenterol Hepatol 2011 Apr;5(2):201-212 [FREE Full text] [doi: 10.1586/egh.11.6] [Medline: 21476915]
8. Oliveira CPMS, da Costa Gayotto LC, Tatai C, Della Bina BI, Janiszewski M, Lima ES, et al. Oxidative stress in the pathogenesis of nonalcoholic fatty liver disease, in rats fed with a choline-deficient diet. J Cell Mol Med 2002 Jul;6(3):399-406 [FREE Full text] [doi: 10.1111/j.1582-4934.2002.tb00518.x] [Medline: 12417056]
9. Roskams T, Yang SQ, Koteish A, Durnez A, DeVos R, Huang X, et al. Oxidative stress and oval cell accumulation in mice and humans with alcoholic and nonalcoholic fatty liver disease. Am J Pathol 2003 Oct;163(4):1301-1311 [FREE Full text] [doi: 10.1016/S0002-9440(10)63489-X] [Medline: 14507639]
10. Wieckowska A, Papouchado BG, Li Z, Lopez R, Zein NN, Feldstein AE. Increased hepatic and circulating interleukin-6 levels in human nonalcoholic steatohepatitis. Am J Gastroenterol 2008 Jun;103(6):1372-1379. [doi: 10.1111/j.1572-0241.2007.01774.x] [Medline: 18510618]
11. Abiru S, Migita K, Maeda Y, Daikoku M, Ito M, Ohata K, et al. Serum cytokine and soluble cytokine receptor levels in patients with non-alcoholic steatohepatitis. Liver Int 2006 Feb;26(1):39-45. [doi: 10.1111/j.1478-3231.2005.01191.x] [Medline: 16420507]
12. Ruhl CE, Everhart JE. Fatty liver indices in the multiethnic United States National Health and Nutrition Examination Survey. Aliment Pharmacol Ther 2015 Jan;41(1):65-76 [FREE Full text] [doi: 10.1111/apt.13012] [Medline: 25376360]
13. Lin ZH, Xin YN, Dong QJ, Wang Q, Jiang XJ, Zhan SH, et al. Performance of the aspartate aminotransferase-to-platelet ratio index for the staging of hepatitis C-related fibrosis: an updated meta-analysis. Hepatology 2011 Mar;53(3):726-736. [doi: 10.1002/hep.24105] [Medline: 21319189]
14. Sterling RK, Lissen E, Clumeck N, Sola R, Correa MC, Montaner J, APRICOT Clinical Investigators. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology 2006 Jun;43(6):1317-1325. [doi: 10.1002/hep.21178] [Medline: 16729309]
15. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. A systematic machine learning based approach for the diagnosis of non-alcoholic fatty liver disease risk and progression. Sci Rep 2018 Feb;8(1):2112 [FREE Full text] [doi: 10.1038/s41598-018-20166-x] [Medline: 29391513]
16. Islam MM, Wu CC, Poly TN, Yang HC, Li YCJ. Applications of machine learning in fatty liver disease prediction. Stud Health Technol Inform 2018;247:166-170. [Medline: 29677944]

17. Yip TCF, Ma AJ, Wong VWS, Tse YK, Chan HLY, Yuen PC, et al. Laboratory parameter-based machine learning model for excluding non-alcoholic fatty liver disease (NAFLD) in the general population. Aliment Pharmacol Ther 2017 Aug;46(4):447-456. [doi: 10.1111/apt.14172] [Medline: 28585725]
18. Sciki-learn library. http://scikit-learn.org. URL: https://scikit-learn.org/stable/ [accessed 2022-05-27]
19. Lundberg S, Lee S. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. United States: Curran Associates Inc; 2017 Dec Presented at: 31st International Conference on Neural Information Processing Systems; Dec 4, 2017; Red Hook, NY p. 4768-4777.
20. Angulo P, Bugianesi E, Bjornsson ES, Charatcharoenwitthaya P, Mills PR, Barrera F, et al. Simple noninvasive systems predict long-term outcomes of patients with nonalcoholic fatty liver disease. Gastroenterology 2013 Oct;145(4):782-789.e4 [FREE Full text] [doi: 10.1053/j.gastro.2013.06.057] [Medline: 23860502]
21. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biom. J 2005 Aug;47(4):458-472. [doi: 10.1002/bimj.200410135] [Medline: 16161804]
22. Yuden WJ. Index for rating diagnostic tests. Cancer 1950 Jan;3(1):32-35. [doi: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3] [Medline: 15405679]
23. Mantovani A, Nascimbeni F, Lonardo A, Zoppini G, Bonora E, Mantzoros CS, et al. Association between primary hypothyroidism and nonalcoholic fatty liver disease: a systematic review and meta-analysis. Thyroid 2018 Oct;28(10):1270-1284. [doi: 10.1089/thy.2018.0257] [Medline: 30084737]
24. Portillo-Sanchez P, Bril F, Maximos M, Lomonaco R, Biernacki D, Orsak B, et al. High prevalence of nonalcoholic fatty liver disease in patients with type 2 diabetes mellitus and normal plasma aminotransferase levels. J Clin Endocrinol Metab 2015 Jun;100(6):2231-2238 [FREE Full text] [doi: 10.1210/jc.2015-1966] [Medline: 25885947]
25. Barb D, Repetto EM, Stokes ME, Shankar SS, Cusi K. Type 2 diabetes mellitus increases the risk of hepatic fibrosis in individuals with obesity and nonalcoholic fatty liver disease. Obesity (Silver Spring) 2021 Nov;29(11):1950-1960. [doi: 10.1002/oby.23263] [Medline: 34553836]
26. Gastaldelli A, Cusi K. From NASH to diabetes and from diabetes to NASH: mechanisms and treatment options. JHEP Rep 2019 Oct;1(4):312-328 [FREE Full text] [doi: 10.1016/j.jhepr.2019.07.002] [Medline: 32039382]
27. Kardashian A, Wilder J, Terrault NA, Price JC. Addressing social determinants of liver disease during the COVID-19 pandemic and beyond: a call to action. Hepatology 2021 Feb;73(2):811-820. [doi: 10.1002/hep.31605] [Medline: 33150599]
28. Spearman CW, Afihene M, Betiku O, Bobat B, Cunha L, Kassianides C, Gastroenterology and Hepatology Association of sub-Saharan Africa (GHASSA). Epidemiology, risk factors, social determinants of health, and current management for non-alcoholic fatty liver disease in sub-Saharan Africa. Lancet Gastroenterol Hepatol 2021 Dec;6(12):1036-1046. [doi: 10.1016/S2468-1253(21)00275-2] [Medline: 34508671]
29. Golovaty I, Tien PC, Price JC, Sheira L, Seligman H, Weiser SD. Food insecurity may be an independent risk factor associated with nonalcoholic fatty liver disease among low-income adults in the United States. J Nutr 2020 Jan;150(1):91-98 [FREE Full text] [doi: 10.1093/jn/nxz212] [Medline: 31504710]
30. Nobili V, Alkhouri N, Alisi A, Della Corte C, Fitzpatrick E, Raponi M, et al. Nonalcoholic fatty liver disease: a challenge for pediatricians. JAMA Pediatr 2015 Feb;169(2):170-176. [doi: 10.1001/jamapediatrics.2014.2702] [Medline: 25506780]

Abbreviations

ALT: alanine aminotransferase
APRI: aspartate aminotransferase- to-platelet ratio index
AST: aspartate aminotransferase
AUC: area under the receiver operating characteristic curve
DT: decision tree
FIB-4: Fibrosis-4
GB: gradient boosting
HDL: high-density lipoprotein
LR: logistic regression
NAFLD: nonalcoholic fatty liver disease
NASH: nonalcoholic steatohepatitis
RF: random forest
ROC: receiver operating characteristic
SHAP: SHapley Additive exPlanations
SVM: support vector machine
TG: triglyceride
US FLI: US Fatty Liver Index

Edited by C Lovis; submitted 02.02.22; peer-reviewed by G Nneji, H Monday, Y Fan, S Kandaswamy; comments to author 27.03.22; accepted 22.04.22; published 06.06.22.

Please cite as:
Wu Y, Yang X, Morris HL, Gurka MJ, Shenkman EA, Cusi K, Bril F, Donahoo WT
Noninvasive Diagnosis of Nonalcoholic Steatohepatitis and Advanced Liver Fibrosis Using Machine Learning Methods: Comparative Study With Existing Quantitative Risk Scores
JMIR Med Inform 2022;10(6):e36997
URL: https://medinform.jmir.org/2022/6/e36997
doi:10.2196/36997
PMID:35666557

Original Paper

Medication-Wide Association Study Using Electronic Health Record Data of Prescription Medication Exposure and Multifetal Pregnancies: Retrospective Study

Lena Davidson[1], MA; Silvia P Canelón[1], PhD; Mary Regina Boland[1], BSc, MA, MPhil, DPhil

Biostatistics, Epidemiology & Informatics, University of Pennsylvania, Philadelphia, PA, United States

Corresponding Author:
Mary Regina Boland, BSc, MA, MPhil, DPhil
Biostatistics, Epidemiology & Informatics
University of Pennsylvania
423 Guardian Drive
421 Blockley Hall
Philadelphia, PA, 19104
United States
Phone: 1 215 573 7394
Email: bolandm@upenn.edu

Abstract

Background: Medication-wide association studies (MWAS) have been applied to assess the risk of individual prescription use and a wide range of health outcomes, including cancer, acute myocardial infarction, acute liver failure, acute renal failure, and upper gastrointestinal ulcers. Current literature on the use of preconception and periconception medication and its association with the risk of multiple gestation pregnancies (eg, monozygotic and dizygotic) is largely based on assisted reproductive technology (ART) cohorts. However, among non-ART pregnancies, it is unknown whether other medications increase the risk of multifetal pregnancies.

Objective: This study aimed to investigate the risk of multiple gestational births (eg, *twins* and *triplets*) following preconception and periconception exposure to prescription medications in patients who delivered at Penn Medicine.

Methods: We used electronic health record data between 2010 and 2017 on patients who delivered babies at Penn Medicine, a health care system in the Greater Philadelphia area. We explored 3 logistic regression models: model 1 (no adjustment); model 2 (adjustment for maternal age); and model 3—our final logistic regression model (adjustment for maternal age, ART use, and infertility diagnosis). In all models, multiple births (MBs) were our outcome of interest (binary outcome), and each medication was assessed separately as a binary variable. To assess our MWAS model performance, we defined ART medications as our gold standard, given that these medications are known to increase the risk of MB.

Results: Of the 63,334 distinct deliveries in our cohort, only 1877 pregnancies (2.96%) were prescribed any medication during the preconception and first trimester period. Of the 123 medications prescribed, we found 26 (21.1%) medications associated with MB (using nominal *P* values) and 10 (8.1%) medications associated with MB (using Bonferroni adjustment) in fully adjusted model 3. We found that our model 3 algorithm had an accuracy of 85% (using nominal *P* values) and 89% (using Bonferroni-adjusted *P* values).

Conclusions: Our work demonstrates the opportunities in applying the MWAS approach with electronic health record data to explore associations between preconception and periconception medication exposure and the risk of MB while identifying novel candidate medications for further study. Overall, we found 3 novel medications linked with MB that could be explored in further work; this demonstrates the potential of our method to be used for hypothesis generation.

(JMIR Med Inform 2022;10(6):e32229) doi:10.2196/32229

KEYWORDS

pregnancy; pregnancy, multiple; assisted reproductive technique; electronic health record

Introduction

Multifetal pregnancies are at a high risk for obstetric complications, including anemia, preterm labor, pregnancy-induced hypertension, placental previa, and fetal malformations [1,2]. These pregnancies pose a risk of adverse fetal and infant outcomes and an increased risk of maternal morbidity and mortality [3,4]. Multifetal pregnancy can occur due to genetic and environmental factors, with higher maternal age, advanced parity, and use of assisted reproductive technology (ART) as established factors in multifetal pregnancy [5]. Although the etiology of dizygotic twins is in most cases straightforward (eg, increase in the number of embryo transfers and medications that increase oocyte release), the etiology of increased monozygotic twinning is less well characterized outside of ART use and fertility treatments [6].

ART is a widely accepted treatment for infertile couples, referring to all treatments that include the handling of eggs, sperm, and embryos. Outside the scope of ART, hormonal medications for the purpose of facilitating a successful pregnancy are referred to as fertility treatment. Increased rates of monozygotic twinning have been observed in pregnancies due to ART use (ie, in vitro fertilization [IVF], micromanipulation, multiple embryo transfer, and gonadotrophin treatment) [6-8]. Ovulation induction (eg, gonadotrophin treatment) therapy may predispose to monozygotic twinning or greater survival of monozygotic twins after their formation [6]. An estimated 1.8% of births in the United States in 2016 were conceived with ART, of which approximately 30.4% were twins and 1.1% were triplets. In animal models, mitotic inhibitors and teratogenic agents were observed to induce monozygotic twinning [9]. In humans, the mechanism of induction of spontaneous twinning remains unknown; twinning-inducing factors outside of ART are thought to involve an environmental exposure (eg, medications and teratogenic agents) during a critical window of pregnancy [9].

The wealth of information from electronic health record (EHR) data can allow for hypothesis-driven research on the associations between medications and pregnancy outcomes. Ryan et al [10] proposed a medication-wide association study (MWAS) approach, in which an outcome of interest is compared with all drugs available for comparison. This approach has been applied to a variety of health outcomes, including cancer risk [11,12]; spontaneous preterm birth [13]; acute myocardial infarction [10,14,15]; and acute liver failure, acute renal failure, and upper gastrointestinal ulcers [10].

Except for research conducted on nationwide health care data registries [11,12,14], MWAS approaches often depend on administrative claims data [10,13,15]. We aimed to present a methodology to systematically explore potential associations between the medications prescribed during the preconception and first trimester period and the occurrence of multiple birth (MB) in patients who delivered at Penn Medicine. Existing screening tools for multifetal pregnancies aim to characterize perinatal morbidity and mortality [16,17], observe noninvasive prenatal screening techniques [18], detect twin-twin transfusion syndrome [19], determine intertwin weight discordance [20], predict MB risk [21], and discover other associations with pregnancy complications [22]. A multitude of these studies depend on IVF clinical data [21], involve increased fetal monitoring [19,20,22], concern twin pregnancy management [18], or are focused on pregnancy complications associated with MB [16,17,22]. Our literature review found no prior research that observes prescription medication use during the periconceptual and first trimester period and its association with MB, let alone using EHR data.

This study illustrates a proof-of-concept MWAS approach for hypothesis-driven pharmacovigilance research on EHR data, with a particular focus on MB.

Methods

Data Source and Identification of MB (Outcome)

We used EHR data obtained from 4 different hospitals within the Penn Medicine system: the Hospital of the University of Pennsylvania, the Pennsylvania Hospital, Penn Presbyterian Hospital, and Chester County Hospital. The deliveries were identified using a previously developed algorithm called Method to Acquire Delivery Date Information from Electronic Health Records (MADDIE) [23]. The MADDIE identified deliveries occurred between 2010 and 2017. The outcome of interest was MB as determined by the International Classification of Diseases, 9th Revision, Clinical Modification and International Classification of Diseases (ICD), 10th Revision, billing codes. We used only the MB codes assigned at the time of delivery (ie, we did not include MB if coded during the pregnancy and not at the time of birth). The total list of codes used to define our outcome is provided in Multimedia Appendix 1. MB differs slightly from multifetal pregnancies in that MB indicates that at the time of birth, the pregnancy consisted of multiple fetuses. Therefore, vanishing twin syndrome and other pregnancy conditions or procedures that reduce the number of fetuses before birth were not assessed in this study [24,25]. We obtained a waiver of consent, as this study included retrospective EHR data analysis without further contact with patients.

Ethics Approval

This study was approved by the institutional review board of the University of Pennsylvania (#828000).

Adjustment for Known Associations of MB

Although a majority of twin births result from natural conception, the incidence of twins and other higher order multifetal pregnancy resulting from superovulation and ART is 20 times greater than the incidence from natural conception [26]. Therefore, we adjusted for ICD, 9th Revision and ICD, 10th Revision billing codes for ART-resulting pregnancy and infertility diagnoses (Multimedia Appendix 2). As ART and infertility diagnoses would likely be assigned both before pregnancy and during pregnancy, we assigned patients as having ART and infertility if they received any of the corresponding ICD codes between 315 days before delivery and the expected date of delivery.

Drug Classification (Exposure Classification)

We mapped all inpatient and outpatient medications from Epic and other EHR systems to RxNorm using a previously described method [27]. In short, medications are mapped to the best match to RxNorm, which is limited to the granularity of the ingredient concept. We defined a *preconception/first trimester exposure* as any medication prescription occurring from 275 days before delivery to 215 days before delivery to capture medications slightly before conception and the first trimester of pregnancy. As ART and fertility medications are often prescribed around the time of conception, we chose this window. Most multifetal pregnancies result in preterm birth and are often completed in <270 days after conception. Therefore, we chose the window of 275-215 days before birth to capture the preconception and periconception window where ART and fertility medications are likely to be used.

We manually annotated the complete list of medications, adding the following elements: generic name, medication type, specific medication type, US Federal Drug Agency pregnancy category, associated comorbidities, and associations with pregnancy outcome treatment. We manually annotated this list because many drugs used in fertility treatments are used off-label; therefore, standardized medical terminology systems would be ineffective in capturing those use cases [28]. We referred to the database [29], RxNav, and a reference guide to fetal and neonatal risk [30] to assign medication use categories to each medication as appropriate. The database [29] is sourced from several medication information suppliers, including Wolters Kluwer Health, the American Society of Health System Pharmacists, Cerner Multum, IBM Watson Micromedex, and Mayo Clinic. Medications used for ART and infertility treatment were defined by the Society for Assisted Reproductive Technology (SART) consumer information and practice guidelines [31]. We grouped medications that were generic and brand names into one to evaluate the effect of the primary ingredient on the birth outcome. Next, we limited the medication list to medications prescribed to at least five patients during the defined exposure time.

Statistical Analysis: MWAS of MB

We constructed 3 logistic regression models with MB as our outcome of interest (binary outcome, 0 or 1), and the effect of each medication on the outcome was assessed separately (each medication exposure was a binary variable, coded 0 or 1). The analysis was performed using the general linear model function in R. The control group for each medication comprised all patients without exposure to the target medication (coded as 0), including patients who had no exposure to medications in the EHR data. Consequently, each target medication had its own control group. We adjusted for 3 known confounders of MB: maternal age (encounter age), ART-resulting pregnancy diagnoses (0 or 1), and infertility diagnoses (0 or 1; Multimedia Appendix 2). A total of three models were constructed: (1) model 1 (no adjustment), (2) model 2 (adjustment for maternal age), and (3) model 3 (adjustment for maternal age, ART-resulting pregnancy, and infertility diagnosis). Diagnoses for ART-resulting pregnancy and infertility were considered in model 3 to account for potential missing prescription data for fertility medical treatment. We reported significant medications (nominal $P<.05$; Bonferroni-adjusted $P<.05$) given the multiple testing that we were performing and calculated odds ratios (ORs) with 95% CIs.

Validation of MWAS and Determining Novel Medications Associated With MB

Significant medications ($P<.05$) with nominal P values and Bonferroni-adjusted P values were evaluated on performance to capture medications used in ART and infertility treatment with binary classification. As previously stated, ART use is an established factor in multifetal pregnancy; therefore, these medications are likely to be associated with MB. The analysis was limited to the medications captured within the defined medication exposure window. Using confusion matrices, we calculated precision, sensitivity, specificity, accuracy, and F1 score (Multimedia Appendix 3).

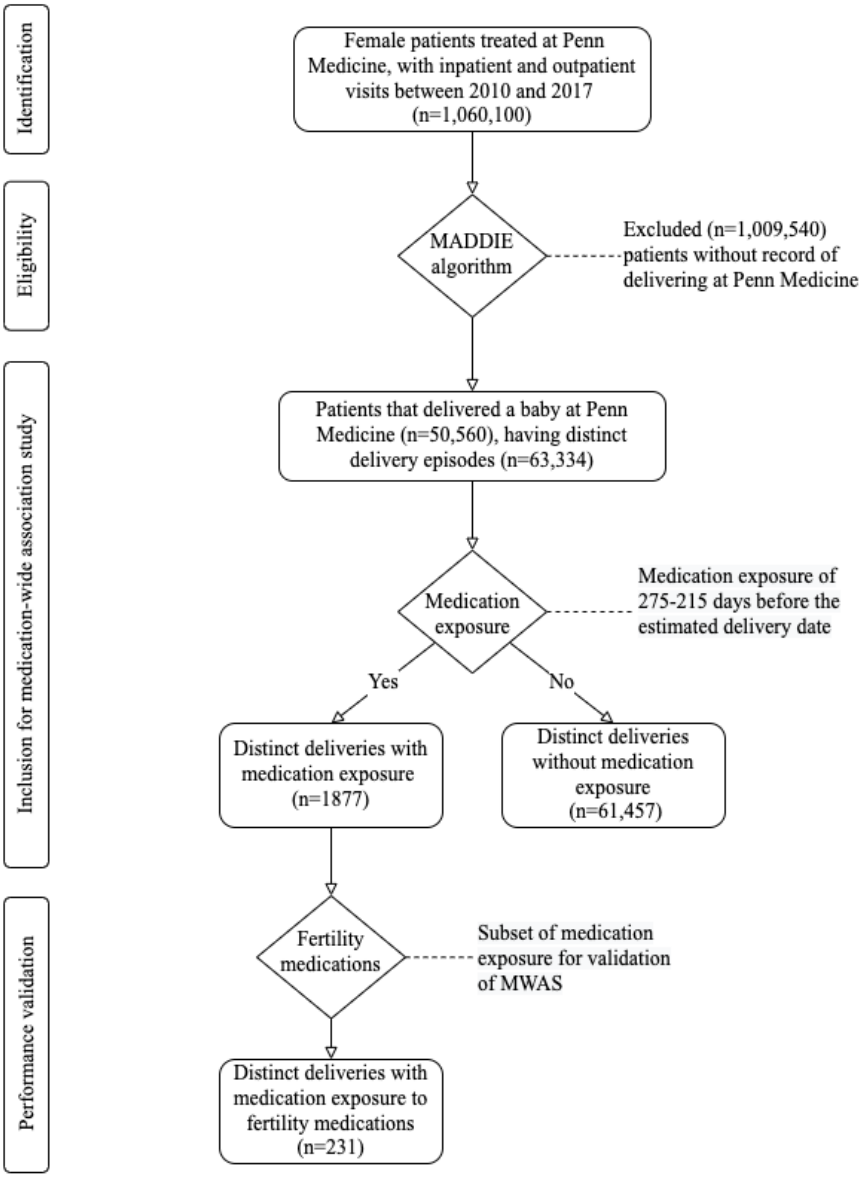
We categorized medications with significant nominal P values into three categories: (1) fertility medications used in ART, (2) medications used for comorbidities associated with MB, and (3) medications not associated with MB in the current literature.

Results

Cohort Characteristics

We obtained EHR data from 1,060,100 female patients treated at Penn Medicine, with inpatient and outpatient visits between 2010 and 2017. A previously developed algorithm called MADDIE identified 50,560 patients who delivered a baby at Penn Medicine having 63,334 distinct deliveries [23]. Figure 1 illustrates the study selection process of the cohort. As shown in Figure 1, our cohort contained 63,334 pregnancies delivered between 2010 and 2017 at Penn Medicine, which was determined by the previously developed MADDIE algorithm [23]. We found that 1562 pregnancies included multiples (eg, twins, triplets, and other higher order multiples), amounting to 2.47% (1562/63,334) of our cohort. We found that of 63,334 pregnancies, 1877 (2.96%) had a recorded prescription medication exposure during the defined exposure time. Furthermore, we found that 5.5% (86/1562) MB pregnancies had a recorded prescription medication exposure during pregnancy.

Figure 1. Retrospective cohort selection process. MADDIE: Method to Acquire Delivery Date Information from Electronic Health Records; MWAS: medication-wide association study.



Drug Classification (Exposure Classification)

We manually annotated 123 medications that were prescribed during the preconception period and the first trimester period of 1877 pregnancies of the 63,334 (2.96%) total distinct deliveries in our cohort (Table 1). These 123 medications belonged to 25 broad drug classes. In our cohort, 15 medications that are typically used as part of fertility treatment were prescribed during pregnancy (Multimedia Appendix 4) [21,22]. Pregnancies with fertility medication exposure (231/63,334, 0.4%) are described in Table 1; the mean age difference (34.6, SD 4.0 years) and the higher incidence of MB (37/231, 16%), ART (16/231, 6.9%), and infertility (4/231, 1.7%) diagnoses are notable, as expected with patients using fertility medication.

Aside from fertility medications, the list contained several types of pain (15/123, 12.2%), antibiotic (11/123, 8.9%), and antihistamine medications (8/123, 6.5%). Most of the extracted medications were not formally assigned (48/123, 39%), followed by category C (31/123, 25.2%) and category B (24/123, 19.5%) medications. As expected, fewer medications were categorized as category A (2/123, 1.6%) and category D (5/123, 4.1%). We found 9.8% (12/123) of medications were categorized as category X, contraindicated in pregnancy, medications—all of which are medications indicated for fertility treatment, contraception, or other indications in obstetrics and gynecology practice.

Table 1. Retrospective cohort medication exposure data.

	Total distinct deliveries (N=63,334)	No prescription medication exposure (n=61,457)	Prescription medication exposure[a] (n=1877)	Fertility medication[b] exposure (n=231)
Pregnancy outcome, n (%)				
Multiple birth[c]	1562 (2.47)	1476 (2.4)	86 (4.58)	37 (16)
Diagnosis, n (%)				
Assisted reproductive technology[d]	246 (0.39)	218 (0.35)	28 (1.49)	16 (6.9)
Infertility[e]	48 (0.08)	39 (0.06)	9 (0.48)	4 (1.7)
Maternal age, mean (SD)	29.5 (6.1)	29.5 (6.1)	30.5 (5.7)	34.6 (4.0)

[a]Prescription medication exposure is during the preconception period and the first trimester period only in this cohort.

[b]Multimedia Appendix 4 provides a list of medications with indication for infertility treatment; note that this is a subset of patients with prescription medication exposure.

[c]Multiple birth determined by International Classification of Diseases (ICD) codes shown in Multimedia Appendix 1.

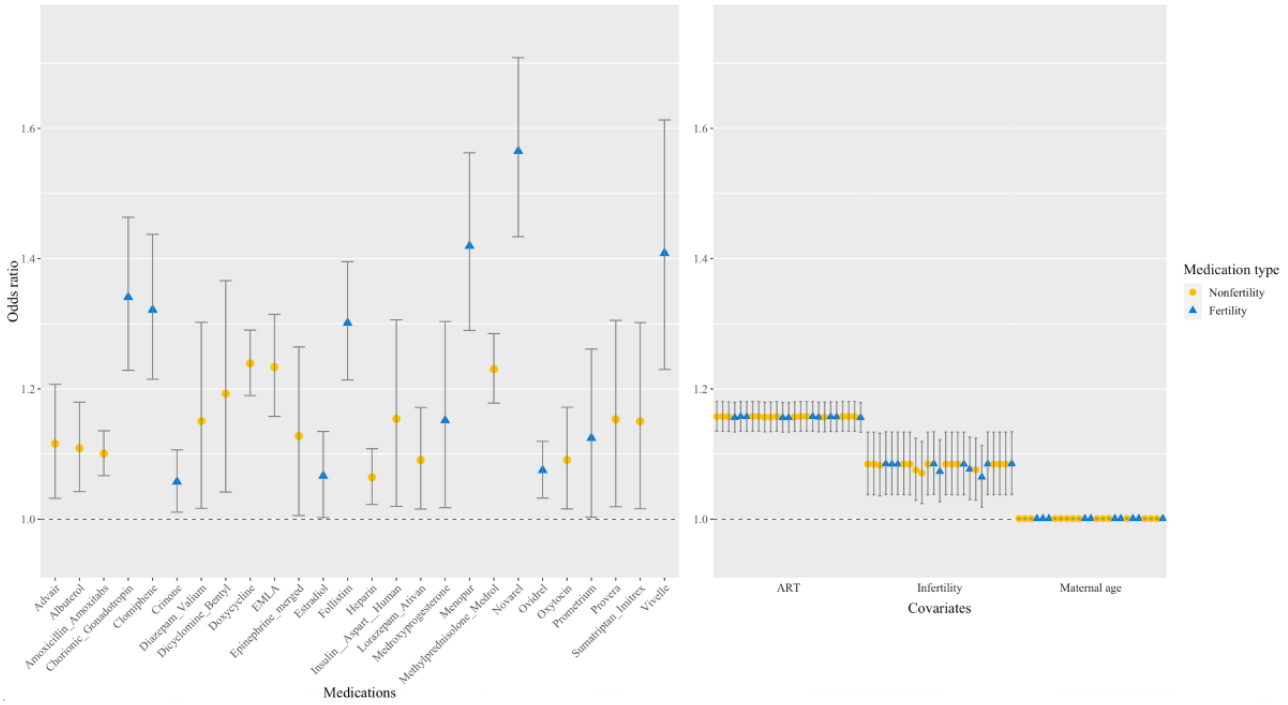
[d]Pregnancy resulting from assisted reproductive was determined by the ICD codes shown in Multimedia Appendix 2.

[e]Infertility diagnosis determined by ICD codes shown in Multimedia Appendix 2.

MWAS: MB

In Figure 2, the significant medications (P<.001 to P=.04) from our fully adjusted model (ie, model 3) are shown with ORs (95% CIs) in a forest plot. The results for all 3 models are presented in Multimedia Appendix 5. Several fertility treatment medications have higher ORs in comparison, namely, Vivelle, Novarel, Menopur, Follistim, clomiphene, and chorionic gonadotropin. The medication class with the highest number of drugs associated with MB (P<.001 to P=.04) was fertility treatment (11/123, 8.9%) prescriptions. The forest plot in Figure 2 illustrates the OR (95% CI) of the significant medications by the covariates in model 3, where an association with ART-resulting pregnancy and infertility diagnoses is shown. The resulting ORs with 95% CIs are listed in Multimedia Appendix 6 for significant (P<.001 to P=.04) and nonsignificant (P=.5 to P=.98) medications.

Figure 2. Medications and covariates significantly associated with multiple birth, using odds ratio (95% CIs). Medication names found significant in our logistic regression model 3 (P<.05) are categorized by drug classification. Odds ratio and CIs are plotted for the covariates in model 3, by each medication: assisted reproductive technology (ART)–resulting pregnancy, infertility diagnosis, and maternal age. Fertility medications are indicated in blue.



Validation Set Performance

Validation set performance was evaluated for our fully adjusted model (model 3). Of the 123 medications extracted, we found 26 (21.1%) medications nominally associated with MB ($P<.001$ to $P=.04$) and 11 (8.9%) medications associated with MB using the Bonferroni adjustment ($P<.001$ to $P=.04$). Multimedia Appendix 5 provides the confusion tables from the performance analysis of all 3 models. Using the Bonferroni correction method, 57% (8/14) fertility medications were captured, whereas 79% (11/14) were captured using the raw or nominal *P* value (Multimedia Appendix 7); therefore, sensitivity performance was greater using noncorrected *P* values (Table 2). This indicates the utility of using nominal *P* values in exploratory MWAS.

Table 2. Performance validation of assisted reproductive technology medications in medication-wide association study.

	Performance metric[a]				
	Sensitivity	Specificity	Accuracy	Precision	F1 score
Model 1: no adjustment					
P value	.80	.84	.84	.41	.54
P value with Bonferroni adjustment	.47	.96	.90	.64	.54
Model 2: adjustment for maternal age[b]					
P value	.73	.85	.84	.41	.53
P value with Bonferroni adjustment	.47	.96	.90	.64	.50
Model 3: adjustment for maternal age and assisted reproductive technology diagnosis[c] and infertility diagnosis[d]					
P value	.73	.86	.85	.42	.53
P value with Bonferroni adjustment	.40	.96	.89	.60	.48

[a]Performance metrics were calculated using formulas shown in Multimedia Appendix 3.

[b]Maternal age determined by age at delivery encounter.

[c]Pregnancy resulting from assisted reproductive technology determined by the International Classification of Diseases codes shown in Multimedia Appendix 2.

[d]Infertility diagnosis determined by the International Classification of Diseases codes shown in Multimedia Appendix 2.

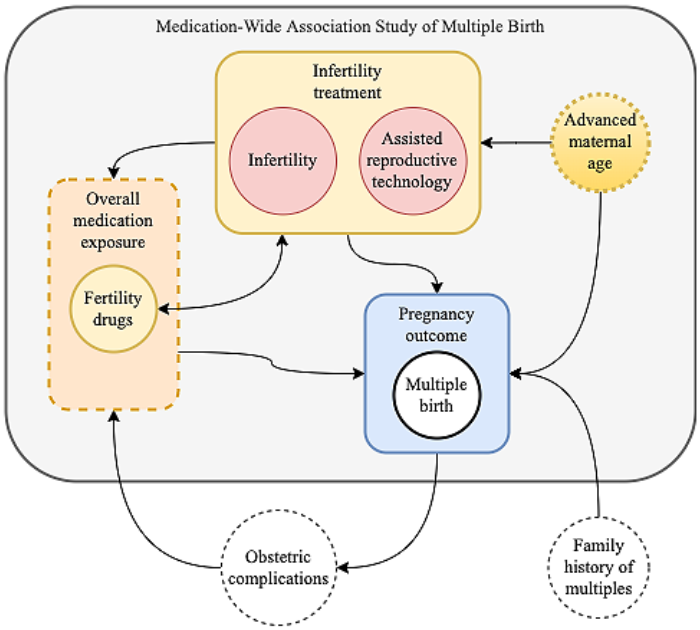
Known, Confounding, and Unknown Associations

Prescription medications associated with comorbidities of fertility and ART treatment were found, as well as medications that may be used for obstetric complications related to multifetal pregnancy care. Of the 26 significant medications using nominal *P* value, 11 (42%) were potential fertility treatment medications; 12 (46%) were associated with infertility and ART use or complications associated with multifetal pregnancy; and 3 (12%) were not previously associated with MB, ART, or fertility-related problems (ie, novel findings or unexpected agents; Table 3). As shown in Figure 3, the validation set included medications used for infertility treatment (medications listed in Multimedia Appendix 4). Nevertheless, the MWAS for MB included confounding medication exposure during multifetal pregnancy, prescribed for (1) treatment of comorbidities of infertility and ART use and (2) treatment of obstetric complications of multifetal pregnancy. Medications associated with ART treatment were associated with MB even after adjusting for ART procedure codes and infertility diagnosis codes. Although 3 asthma medications were found to be significant, previous studies showed mixed results when examining the relationship between asthma, asthma medication use, and fertility [32-34]. The association between irritable bowel disease (IBD) and fertility is complex; patients with quiescent IBD have fertility rates comparable with those of the general population [28], whereas patients with an active disease or those who had undergone a pelvic surgery may have reduced fertility [30]. Overall, we found 3 medications not previously reported to be associated with an increased risk of MB following prescription during the preconception and periconceptional period: sumatriptan and imitrex ($P=.03$), oxytocin ($P=.02$), and lorazepam and ativan ($P=.02$).

Table 3. Medications associated with multiple birth after adjustment for assisted reproductive technology, infertility, and maternal age (model 3).

Indicated comorbidity	Generic medication name or names	Medications associated with multiple birth, n (%)
Associated with infertility and assisted reproductive technology		
Assisted reproductive technology treatment	EMLA, methylprednisolone, diazepam, amoxicillin, doxycycline, and medroxyprogesterone acetate	6 (23)
Asthma	Albuterol, fluticasone propionate and salmeterol, and epinephrine	3 (12)
Irritable bowel disease	Dicyclomine	1 (4)
Associated with multifetal pregnancy		
Cardiovascular-related diagnoses (gestational hypertension and thrombosis)	Heparin	1 (4)
Gestational diabetes mellitus	Insulin aspart, human	1 (4)
Not previously associated with multiple birth, assisted reproductive technology, or fertility-related problems	Sumatriptan, oxytocin, and lorazepam	3 (12)

Figure 3. Conceptual schema for medication-wide association study (MWAS) analyses on multiple birth. Confounding relationships for medication-outcome associations are illustrated. Within the MWAS, we adjust for maternal age, infertility diagnosis, and assisted reproductive technology–resulting pregnancy diagnosis. The study does not adjust for all known associations of multiple birth such as obstetric complications or family history of multiples. The validation of the MWAS models observed performance in capturing fertility medication exposure.



Discussion

Overview

We applied 3 logistic regression models to retrospective EHR data of a cohort of patients who delivered at Penn Medicine between 2010 and 2017 (n=63,334) to explore potential associations between the medications prescribed during the preconception and first trimester period (binary variable) and the occurrence of MB (binary outcome). We discuss the results of our MWAS from our fully adjusted model that was adjusted for age and ART and infertility diagnosis (model 3) on MB for all associations revealed using nominal.

Reason for Conducting an MWAS on MB

The application of an MWAS approach to MB allows the analysis of medications used outside the scope of obstetric treatment, capturing comorbidities that may increase the risk of the outcome. Not all of this is known, as MB is more commonly used as an adjustment for analysis of other pregnancy outcomes of interest. Off-label use is common in pregnancy and infertility treatment [28,35]. Our MWAS approach with annotation of known off-label uses can further improve the identification of comorbidities associated with MB (eg, infertility and subsequent ART use). Research into the side effects of medications is more focused on adverse outcomes than MB, notwithstanding the risks of multifetal pregnancy.

The graph in Multimedia Appendix 8 illustrates the overlap between patients with the respective fertility diagnoses and fertility medication prescriptions in our cohort. This also demonstrates that many patients with fertility or ART medications were not assigned the corresponding diagnostic code, indicating that fertility studies using EHR data should include medication history to fully capture affected patients. A recently enhanced algorithm to detect ectopic pregnancy in the EHR used diagnosis and procedure codes as well as medication exposure [36]. The complete picture of a patient's medical encounters during pregnancy is likely not captured in the EHR of a health care system (eg, engagement with more than one health care system, over-the-counter medications, etc). Adjustment for fertility diagnoses and pregnancies resulting from ART treatments may not truly represent patients undergoing ART and fertility treatment if the diagnosis codes are used without the inclusion of medication histories. The same is true when only fertility medications are observed, especially medications with multiple indications for care in obstetrics and gynecology. Fertility treatment, meaning without eggs or embryos handled, may involve medical treatment (eg, clomiphene and gonadotropins) that increases the chances of multifetal pregnancy. Pregnancies resulting from fertility treatment do not necessarily indicate ART use; therefore, EHR may not reflect ART use diagnosis. The EHR should include an infertility diagnosis in these cases, but we found that in many instances, both infertility diagnoses and ART use codes were absent from those receiving these medications (Multimedia Appendix 8). Using diagnosis codes and medication exposure should allow for better capture of MB in comparison with using only one or the other.

EHR for MWAS Versus SART Database (or Other ART Cohort Database)

MB outcomes can also be observed in ART cohort databases, such as the SART Clinical Outcomes Reporting System. However, fertility medication treatment without the intention of egg retrieval will not necessarily be captured within such databases, as they are beyond the scope of ART. Moreover, not all multifetal pregnancies result from infertility treatment and ART. Finally, such databases are reported from ART clinics and are not necessarily representative of all the medications prescribed during pregnancy. An ART cohort database may have a wealth of data elements specific to ART treatment; however, these data are reported using inconsistent methods, often from a variety of reporting services [37]. In contrast, EHRs may also have missing prescription information due to offsite care; however, the scope of captured health information is likely more comprehensive overall than that of an ART clinic because it includes medications for comorbidities and other aspects of patients' care that may be overlooked by ART specialists.

Medication Exposure During the Preconception and First Trimester Period

We found that 5.51% (86/1562) MB pregnancies had prescription medication exposure during pregnancy. Therefore, pregnancies resulting in MB were more likely to have recorded prescription medication during the preconception and first trimester period. This is consistent with (1) the fact that ART often uses medications early on to induce pregnancy [20] and (2) multifetal pregnancies are at higher risk of pregnancy complications [21] and therefore may be more likely to receive prescription medication treatment. Moreover, a higher proportion of MB (37/231, 16%) was found for those exposed to fertility medications in comparison with the occurrence of MB (1562/63,334, 2.47%) for the overall cohort.

Our Evaluation Using Known Fertility and ART Medications That Increase the Risk of MB Is Not Perfect

Overview

To assess the ability of our MWAS to capture medications that increase the risk of MB, we used medications that are known to increase an individual's chance of conceiving and have been implicated in increasing the risk of MB in the literature (Multimedia Appendix 4). We know that this list of medications is incomplete (and hence part of the reason for this study), but we wanted to understand how many known medications we were able to capture using our MWAS approach. The indication of medication prescribed is not necessarily straightforward; without observing clinical notes and ICD codes from an encounter, there are often several therapeutic uses for which the medication could have been prescribed (eg, progesterone). More context and research are required to understand the discovered associations further. Although known fertility medications were missed by our approach (3/14, 21%), we observed a large number of drugs not known to be associated with MB with insignificant nominal *P* values (93/123, 75.6%), which is comforting. We observed drugs used in fertility treatment (11/26, 42%) and drugs known to be associated with multifetal pregnancy (12/26, 46%), along with 3 (11%) novel associations. Associated comorbidities of infertility overlap with obstetric complications during multifetal pregnancy, including diabetes mellitus, cardiovascular disease, thyroid dysregulation, and liver dysregulation.

Medications Associated With Infertility and ART

Medications used in fertility treatment themselves may be captured solely because of reverse causation, although they do not have a truly strong association with multifetal pregnancy. Several medications may be prescribed during IVF treatment cycles for preventive care or other indications, including antibiotics (doxycycline and amoxicillin), a corticosteroid (methylprednisolone), pain management (EMLA), progestin-induced menstruation (medroxyprogesterone acetate), and conscious sedation (diazepam) [38,39].

ART and ovulation induction procedures are used for fertility treatment. A comprehensive review of infertility comorbidities in women suggests that infertility is a complex health care issue, and women with infertility are at a higher risk of psychiatric disorders and endometrial cancer [40]. Infertility and fertility treatment are associated with other pathologies, such as polycystic ovarian syndrome, endometriosis, thyroid disorders, breast cancer, cardiovascular disease, metabolic syndrome, diabetes mellitus, and liver dysfunction [41].

Medications associated with comorbidities of infertility were identified, including treatments for asthma and IBD. Research

shows that women with asthma have higher pregnancy losses [32] and a prolonged time to pregnancy [33]; in contrast, some studies have shown no association [34]. Bronchodilators (albuterol, epinephrine, and fluticasone propionate or salmeterol) may be pharmacological treatments for asthma, which has been linked to a prolonged time to pregnancy and is associated with a higher need for fertility treatment among women aged ≥35 years [42]. In addition, a retrospective study on asthma during pregnancy in Sweden found that women hospitalized for asthma had a higher risk of twinning [43]. Dicyclomine is used to treat IBD; however, the association between IBD and fertility is complex, and patients with quiescent IBD have fertility rates comparable with those of the general population [44], whereas patients with an active disease or those who had undergone a pelvic surgery may have reduced fertility [45].

Medications Associated With Obstetric Complications During Multifetal Pregnancy

Medications identified by the MWAS may be prescribed for obstetric complications associated with multifetal pregnancy. These pregnancies are at an increased risk of obstetric complications, such as preterm birth, placental problems, gestational diabetes mellitus, anemia, and preeclampsia. Owing to the time exposure range, medications typically used to treat complications typically past the first trimester of pregnancy were not captured by the MWAS. One antidiabetic medication (ie, insulin aspart, human) was identified; however, other forms of insulin and the insulin sensitizer metformin were not identified as significant. A single antithrombotic medication, heparin, was identified, but other anticoagulants and cardiovascular-related medications were not identified in our models.

Novel Findings of Medications Associated With MB

Migraines have a high incidence in obstetrics; one migraine pharmacological treatment (sumatriptan) was found to be associated with MB. An association between migraine history and development of ovarian hyperstimulation syndrome may indicate the risk of multifetal pregnancy [46], as ovarian hyperstimulation syndrome–complicated pregnancy is linked to a higher incidence of MB [47]. However, further research is required to understand the biological mechanisms, if any, underlying this association. Oxytocin could be associated with MB because of prior pregnancy delivery episodes (as oxytocin is used during labor), indicating a short time between pregnancies.

Limitations and Future Work

Although the World Health Organization's anatomical therapeutic chemical classification system is applicable, the proportion of RxNorm drugs mapped to anatomical therapeutic chemicals would result in fewer medications being included in the analysis. However, the therapeutic use of the medications has not been explicitly determined. Medications classified as fertility related are based on the SART references; however, without discrete indications, they potentially underpower performance in the validation process. In addition, a major limitation of using standard pharmacology and drug-related terminologies is that approximately 11% of medications used in women's health are off-label [48]. This includes several popular medications commonly used in the obstetrics and gynecology domain [49-51]. Use of off-label medications requires manual review of medications, which is laborious. Manual review and classification of the prescription medications were conducted by an informaticist (LD) and not a pharmacologist. Several extracted prescriptions were available over the counter (38/123, 30.9%); therefore, exposure to such drugs is likely underrepresented in our cohort. Potentially highly related variables were not considered in the analysis, introducing a possible omitted variable bias (eg, drug dose, drug form, route of application, and temporal component of exposure). Medication exposure of 275-215 days before subsequent delivery may likely include medication exposure before conception (ie, prior pregnancy delivery episodes), especially regarding the length of gestation due to preterm birth. We did not ensure that medication exposure occurred before conception; therefore, the medications associated with multifetal pregnancies in this study are not causal in nature. Unexpected agents and probable confounding medications require further adjustment in the MWAS technique to provide more reliable, meaningful results.

Conclusions

Our research demonstrates opportunities in using an MWAS approach with EHR data to explore agents previously unknown to be associated with MB outcomes. The results indicated that a number of medications used in ART and infertility treatment were associated with an increased incidence of MB likely due to multifetal pregnancy, as expected. Using these medications as our gold standard, we found that our algorithm had an accuracy of 85% and 89%, using nominal *P* values and Bonferroni-adjusted *P* values, respectively. Sensitivity and F1 score were improved using nominal *P* values in comparison with Bonferroni-adjusted *P* values, indicating the applicability of nominal *P* values in exploratory MWAS studies. A total of 6 novel agents were linked to MB, with the remaining 20 medications potentially linked to the comorbidities of infertility, ART use, and obstetric complications during multifetal pregnancy. The MWAS approach can facilitate hypothesis-driven data exploration, informing the adjustments needed in the models in further research. Our approach also highlights the importance of exploring medication histories, as many patients receiving ART and fertility treatments do not have corresponding diagnosis codes indicating treatment. If medication information was not used, these patients were mistakenly labeled as having not received ART and infertility treatment. This underscores the importance of multidata modalities in retrospective EHR studies, especially for those exploring the effects and outcomes related to pregnancy.

Acknowledgments

The authors would like to thank the Perelman School of Medicine at the University of Pennsylvania for providing generous funds to support this project.

Conflicts of Interest

None declared.

Multimedia Appendix 1
International Classification of Diseases (ICD)-9 and ICD-10 codes to capture multiple birth.
[DOCX File , 20 KB - medinform_v10i6e32229_app1.docx]

Multimedia Appendix 2
International Classification of Diseases (ICD)-9 and ICD-10 codes used to identify comorbidity diagnosis from the electronic health record for model adjustment.
[DOCX File , 15 KB - medinform_v10i6e32229_app2.docx]

Multimedia Appendix 3
Formulas for validation.
[PDF File (Adobe PDF File), 562 KB - medinform_v10i6e32229_app3.pdf]

Multimedia Appendix 4
Medications that can be indicated for infertility treatment.
[DOCX File , 20 KB - medinform_v10i6e32229_app4.docx]

Multimedia Appendix 5
Medication-wide association study results for models 1, 2, and 3.
[XLSX File (Microsoft Excel File), 60 KB - medinform_v10i6e32229_app5.xlsx]

Multimedia Appendix 6
Odds ratio with 95% CIs for observed prescription of medications.
[DOCX File , 26 KB - medinform_v10i6e32229_app6.docx]

Multimedia Appendix 7
Performance analysis confusion tables.
[DOCX File , 18 KB - medinform_v10i6e32229_app7.docx]

Multimedia Appendix 8
Graph of patients with assisted reproductive technology or fertility diagnosis and patients with fertility medication prescriptions.
[PDF File (Adobe PDF File), 76 KB - medinform_v10i6e32229_app8.pdf]

References

1. Ru Y, Pressman EK, Cooper EM, Guillet R, Katzman PJ, Kent TR, et al. Iron deficiency and anemia are prevalent in women with multiple gestations. Am J Clin Nutr 2016 Oct;104(4):1052-1060. [doi: 10.3945/ajcn.115.126284] [Medline: 27581469]
2. Choi SH, Park YS, Shim KS, Choi YS, Chang JY, Hahn WH, et al. Recent trends in the incidence of multiple births and its consequences on perinatal problems in Korea. J Korean Med Sci 2010 Aug;25(8):1191-1196 [FREE Full text] [doi: 10.3346/jkms.2010.25.8.1191] [Medline: 20676332]
3. Senat MV, Ancel PY, Bouvier-Colle MH, Bréart G. How does multiple pregnancy affect maternal mortality and morbidity? Clin Obstet Gynecol 1998 Mar;41(1):78-83. [doi: 10.1097/00003081-199803000-00013] [Medline: 9504226]
4. Santana DS, Silveira C, Costa ML, Souza RT, Surita FG, Souza JP, WHO Multi-Country Survey on MaternalNewborn Health Research Network. Perinatal outcomes in twin pregnancies complicated by maternal morbidity: evidence from the WHO Multicountry Survey on Maternal and Newborn Health. BMC Pregnancy Childbirth 2018 Nov 20;18(1):449 [FREE Full text] [doi: 10.1186/s12884-018-2082-9] [Medline: 30453908]
5. Parazzini F, Villa A, Moroni S, Tozzi L, Restelli S. The epidemiology of multiple pregnancies. Acta Genet Med Gemellol (Roma) 1994;43(1-2):17-23. [doi: 10.1017/s0001566000002919] [Medline: 7847017]

6. Schachter M, Raziel A, Friedler S, Strassburger D, Bern O, Ron-El R. Monozygotic twinning after assisted reproductive techniques: a phenomenon independent of micromanipulation. Hum Reprod 2001 Jun;16(6):1264-1269. [doi: 10.1093/humrep/16.6.1264] [Medline: 11387303]
7. Yanaihara A, Yorimitsu T, Motoyama H, Watanabe H, Kawamura T. Monozygotic multiple gestation following in vitro fertilization: analysis of seven cases from Japan. J Exp Clin Assist Reprod 2007 Sep 22;4:4 [FREE Full text] [doi: 10.1186/1743-1050-4-4] [Medline: 17888172]
8. Derom C, Vlietinck R, Derom R, Van den Berghe H, Thiery M. Increased monozygotic twinning rate after ovulation induction. Lancet 1987 May 30;1(8544):1236-1238. [doi: 10.1016/s0140-6736(87)92688-2] [Medline: 2884372]
9. Kaufman M. The embryology of conjoined twins. Childs Nerv Syst 2004 Aug;20(8-9):508-525. [doi: 10.1007/s00381-004-0985-4] [Medline: 15278382]
10. Ryan PB, Madigan D, Stang PE, Schuemie MJ, Hripcsak G. Medication-wide association studies. CPT Pharmacometrics Syst Pharmacol 2013 Sep 18;2:e76 [FREE Full text] [doi: 10.1038/psp.2013.52] [Medline: 24448022]
11. Patel CJ, Ji J, Sundquist J, Ioannidis JP, Sundquist K. Systematic assessment of pharmaceutical prescriptions in association with cancer risk: a method to conduct a population-wide medication-wide longitudinal study. Sci Rep 2016 Aug 10;6:31308 [FREE Full text] [doi: 10.1038/srep31308] [Medline: 27507038]
12. Andreassen BK, Støer NC, Martinsen JI, Ursin G, Weiderpass E, Thoresen GH, et al. Identification of potential carcinogenic and chemopreventive effects of prescription drugs: a protocol for a Norwegian registry-based study. BMJ Open 2019 Apr 08;9(4):e028504 [FREE Full text] [doi: 10.1136/bmjopen-2018-028504] [Medline: 30962244]
13. Marić I, Winn VD, Borisenko E, Weber KA, Wong RJ, Aziz N, et al. Data-driven queries between medications and spontaneous preterm birth among 2.5 million pregnancies. Birth Defects Res 2019 Oct 01;111(16):1145-1153. [doi: 10.1002/bdr2.1580] [Medline: 31433567]
14. Sen A, Vardaxis I, Lindqvist BH, Brumpton BM, Strand LB, Bakken IJ, et al. Systematic assessment of prescribed medications and short-term risk of myocardial infarction - a pharmacopeia-wide association study from Norway and Sweden. Sci Rep 2019 Jun 04;9(1):8257 [FREE Full text] [doi: 10.1038/s41598-019-44641-1] [Medline: 31164670]
15. Coloma PM, Schuemie MJ, Trifirò G, Furlong L, van Mulligen E, Bauer-Mehren A, EU-ADR consortium. Drug-induced acute myocardial infarction: identifying 'prime suspects' from electronic healthcare records-based surveillance system. PLoS One 2013;8(8):e72148 [FREE Full text] [doi: 10.1371/journal.pone.0072148] [Medline: 24015213]
16. Blitz MJ, Yukhayev A, Pachtman SL, Reisner J, Moses D, Sison CP, et al. Twin pregnancy and risk of postpartum hemorrhage. J Matern Fetal Neonatal Med 2020 Nov;33(22):3740-3745. [doi: 10.1080/14767058.2019.1583736] [Medline: 30836810]
17. Lu Y, Ding Z, Li W, Mei L, Shen L, Shan H. Prediction of twin pregnancy preeclampsia based on clinical risk factors, early pregnancy serum markers, and uterine artery pulsatility index. Pak J Med Sci 2021;37(7):1727-1733 [FREE Full text] [doi: 10.12669/pjms.37.7.5041] [Medline: 34912386]
18. Benn P, Rebarber A. Non-invasive prenatal testing in the management of twin pregnancies. Prenat Diagn 2021 Sep;41(10):1233-1240 [FREE Full text] [doi: 10.1002/pd.5989] [Medline: 34170028]
19. Takano M, Nakata M, Nagasaki S, Sakuma J, Morita M. Prediction of twin-to-twin transfusion syndrome using characteristic waveforms of ductus venosus in recipient twins. Twin Res Hum Genet 2020 Oct;23(5):292-297. [doi: 10.1017/thg.2020.73] [Medline: 33004103]
20. Breathnach FM, McAuliffe FM, Geary M, Daly S, Higgins JR, Dornan J, Perinatal Ireland Research Consortium. Definition of intertwin birth weight discordance. Obstet Gynecol 2011 Jul;118(1):94-103. [doi: 10.1097/AOG.0b013e31821fd208] [Medline: 21691168]
21. Lannon BM, Choi B, Hacker MR, Dodge LE, Malizia BA, Barrett CB, et al. Predicting personalized multiple birth risks after in vitro fertilization-double embryo transfer. Fertil Steril 2012 Jul;98(1):69-76. [doi: 10.1016/j.fertnstert.2012.04.011] [Medline: 22673597]
22. Kuhrt K, Hezelgrave-Elliott N, Stock SJ, Tribe R, Seed PT, Shennan AH. Quantitative fetal fibronectin for prediction of preterm birth in asymptomatic twin pregnancy. Acta Obstet Gynecol Scand 2020 Sep;99(9):1191-1197 [FREE Full text] [doi: 10.1111/aogs.13861] [Medline: 32249408]
23. Canelón SP, Burris HH, Levine LD, Boland MR. Development and evaluation of MADDIE: method to acquire delivery date information from electronic health records. Int J Med Inform 2021 Jan;145:104339 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104339] [Medline: 33232918]
24. Landy HJ, Keith LG. The vanishing twin: a review. Hum Reprod Update 1998;4(2):177-183. [doi: 10.1093/humupd/4.2.177] [Medline: 9683354]
25. Almog B, Levin I, Wagman I, Kapustiansky R, Lessing JB, Amit A, et al. Adverse obstetric outcome for the vanishing twin syndrome. Reprod Biomed Online 2010 Feb;20(2):256-260. [doi: 10.1016/j.rbmo.2009.11.015] [Medline: 20113963]
26. Adashi EY, Barri PN, Berkowitz R, Braude P, Bryan E, Carr J, et al. Infertility therapy-associated multiple pregnancies (births): an ongoing epidemic. Reprod Biomed Online 2003 Nov;7(5):515-542. [doi: 10.1016/s1472-6483(10)62069-x] [Medline: 14686351]

27. Davidson L, Boland MR. Comparative analysis and evaluation of state-of-the-art medication mapping tools to transform a local medication terminology to RxNorm. AMIA Jt Summits Transl Sci Proc 2020;2020:126-135 [FREE Full text] [Medline: 32477631]
28. Legro RS. Introduction: on-label and off-label drug use in reproductive medicine. Fertil Steril 2015 Mar;103(3):581-582. [doi: 10.1016/j.fertnstert.2015.01.028] [Medline: 25726701]
29. Drugs.com. URL: https://www.drugs.com/ [accessed 2022-05-26]
30. Briggs GG, Freeman RK, Roger K. Drugs in Pregnancy and Lactation: A Reference Guide to Fetal and Neonatal Risk, 10th Edition. Philadelphia, PA, USA: Lippincott Williams & Wilkins; 2014.
31. SART: ART Medications. Society for Assisted Reproductive Technology. URL: https://www.sart.org/patients/a-patients-guide-to-assisted-reproductive-technology/general-information/art-medications/ [accessed 2022-05-16]
32. Turkeltaub PC, Lockey RF, Holmes K, Friedmann E. Author Correction: asthma and/or hay fever as predictors of fertility/impaired fecundity in U.S. women: national survey of family growth. Sci Rep 2020 Apr 23;10(1):7193 [FREE Full text] [doi: 10.1038/s41598-020-64338-0] [Medline: 32322018]
33. Gade EJ, Thomsen SF, Lindenberg S, Kyvik KO, Lieberoth S, Backer V. Asthma affects time to pregnancy and fertility: a register-based twin study. Eur Respir J 2014 Apr;43(4):1077-1085 [FREE Full text] [doi: 10.1183/09031936.00148713] [Medline: 24232708]
34. Tata LJ, Hubbard RB, McKeever TM, Smith CJ, Doyle P, Smeeth L, et al. Fertility rates in women with asthma, eczema, and hay fever: a general population-based cohort study. Am J Epidemiol 2007 May 01;165(9):1023-1030. [doi: 10.1093/aje/kwk092] [Medline: 17255115]
35. Usadi RS, Merriam KS. On-label and off-label drug use in the treatment of female infertility. Fertil Steril 2015 Mar;103(3):583-594. [doi: 10.1016/j.fertnstert.2015.01.011] [Medline: 25660647]
36. Getahun D, Shi JM, Chandra M, Fassett MJ, Alexeeff S, Im TM, et al. Identifying ectopic pregnancy in a large integrated health care delivery system: algorithm validation. JMIR Med Inform 2020 Nov 30;8(11):e18559 [FREE Full text] [doi: 10.2196/18559] [Medline: 33141678]
37. Mendola P, Gilboa SM. Reporting of birth defects in SART CORS: time to rely on data linkage. Fertil Steril 2016 Sep 01;106(3):554-555 [FREE Full text] [doi: 10.1016/j.fertnstert.2016.06.020] [Medline: 27343954]
38. Kaser DJ, Ginsburg ES, Carrell DT, Racowsky C. Chapter 31 - Assisted reproduction. In: Yen & Jaffe's Reproductive Endocrinology Physiology, Pathophysiology, and Clinical Management, 8th edition. USA: Elsevier; 2017.
39. Trout SW, Vallerand AH, Kemmann E. Conscious sedation for in vitro fertilization. Fertil Steril 1998 May;69(5):799-808. [doi: 10.1016/s0015-0282(98)00031-4] [Medline: 9591482]
40. Hanson B, Johnstone E, Dorais J, Silver B, Peterson CM, Hotaling J. Female infertility, infertility-associated diagnoses, and comorbidities: a review. J Assist Reprod Genet 2017 Feb;34(2):167-177 [FREE Full text] [doi: 10.1007/s10815-016-0836-8] [Medline: 27817040]
41. Lorzadeh N, Kazemirad N, Kazemirad Y. Human immunodeficiency: extragonadal comorbidities of infertility in women. Immun Inflamm Dis 2020 Sep 03;8(3):447-457 [FREE Full text] [doi: 10.1002/iid3.327] [Medline: 32621331]
42. Vejen Hansen A, Ali Z, Malchau SS, Blafoss J, Pinborg A, Ulrik CS. Fertility treatment among women with asthma: a case-control study of 3689 women with live births. Eur Respir J 2019 Feb;53(2):1800597 [FREE Full text] [doi: 10.1183/13993003.00597-2018] [Medline: 30464019]
43. Källén B, Rydhstroem H, Åberg A. Asthma during pregnancy – a population based study. Eur J Epidemiol 2000 Feb;16(2):171. [doi: 10.1023/A:1007678404911]
44. Tavernier N, Fumery M, Peyrin-Biroulet L, Colombel J, Gower-Rousseau C. Systematic review: fertility in non-surgically treated inflammatory bowel disease. Aliment Pharmacol Ther 2013 Oct;38(8):847-853 [FREE Full text] [doi: 10.1111/apt.12478] [Medline: 24004045]
45. Kokoszko-Bilska A, Sobkiewicz S, Fichna J. Inflammatory bowel diseases and reproductive health. Pharmacol Rep 2016 Aug;68(4):859-864. [doi: 10.1016/j.pharep.2016.03.009] [Medline: 27117378]
46. Rollene NL, Khan Z, Schroeder DR, Cutrer FM, Coddington CC. Migraines and ovarian hyperstimulation syndrome: a dopamine connection. Fertil Steril 2011 Jan;95(1):417-419. [doi: 10.1016/j.fertnstert.2010.08.041] [Medline: 20889153]
47. Raziel A, Schachter M, Friedler S, Ron-El R. Outcome of IVF pregnancies following severe OHSS. Reprod Biomed Online 2009 Jan;19(1):61-65. [doi: 10.1016/s1472-6483(10)60047-8]
48. Radley DC, Finkelstein SN, Stafford RS. Off-label prescribing among office-based physicians. Arch Intern Med 2006 May 08;166(9):1021-1026. [doi: 10.1001/archinte.166.9.1021] [Medline: 16682577]
49. Rayburn WF, Turnbull GL. Off-label drug prescribing on a state university obstetric service. J Reprod Med 1995 Mar;40(3):186-188. [Medline: 7776301]
50. Voigt F, Goecke TW, Najjari L, Pecks U, Maass N, Rath W. Off-label use of misoprostol for labor induction in Germany: a national survey. Eur J Obstet Gynecol Reprod Biol 2015 Apr;187:85-89. [doi: 10.1016/j.ejogrb.2014.11.026] [Medline: 25553610]
51. Marret H, Fauconnier A, Dubernard G, Misme H, Lagarce L, Lesavre M, et al. Overview and guidelines of off-label use of methotrexate in ectopic pregnancy: report by CNGOF. Eur J Obstet Gynecol Reprod Biol 2016 Oct;205:105-109. [doi: 10.1016/j.ejogrb.2016.07.489] [Medline: 27572300]

Abbreviations

ART: assisted reproductive technology
EHR: electronic health record
IBD: irritable bowel disease
ICD: International Classification of Diseases
IVF: in vitro fertilization
MADDIE: Method to Acquire Delivery Date Information from Electronic Health Records
MB: multiple birth
MWAS: medication-wide association study
OR: odds ratio
SART: Society for Assisted Reproductive Technology

Edited by C Lovis; submitted 19.07.21; peer-reviewed by Y Chu, R Poluru, S Matsuda, I Ioakeim-Skoufa; comments to author 02.01.22; revised version received 19.02.22; accepted 21.04.22; published 07.06.22.

Please cite as:
Davidson L, Canelón SP, Boland MR
Medication-Wide Association Study Using Electronic Health Record Data of Prescription Medication Exposure and Multifetal Pregnancies: Retrospective Study
JMIR Med Inform 2022;10(6):e32229
URL: https://medinform.jmir.org/2022/6/e32229
doi:10.2196/32229
PMID:35671076

Original Paper

Error and Timeliness Analysis for Using Machine Learning to Predict Asthma Hospital Visits: Retrospective Cohort Study

Xiaoyi Zhang[1], MS; Gang Luo[1], DPhil

Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States

Corresponding Author:
Gang Luo, DPhil
Department of Biomedical Informatics and Medical Education
University of Washington
UW Medicine South Lake Union
850 Republican Street, Building C, Box 358047
Seattle, WA, 98195
United States
Phone: 1 206 221 4596
Fax: 1 206 221 2671
Email: gangluo@cs.wisc.edu

Abstract

Background: Asthma hospital visits, including emergency department visits and inpatient stays, are a significant burden on health care. To leverage preventive care more effectively in managing asthma, we previously employed machine learning and data from the University of Washington Medicine (UWM) to build the world's most accurate model to forecast which asthma patients will have asthma hospital visits during the following 12 months.

Objective: Currently, two questions remain regarding our model's performance. First, for a patient who will have asthma hospital visits in the future, how far in advance can our model make an initial identification of risk? Second, if our model erroneously predicts a patient to have asthma hospital visits at the UWM during the following 12 months, how likely will the patient have ≥1 asthma hospital visit somewhere else or ≥1 surrogate indicator of a poor outcome? This work aims to answer these two questions.

Methods: Our patient cohort included every adult asthma patient who received care at the UWM between 2011 and 2018. Using the UWM data, our model made predictions on the asthma patients in 2018. For every such patient with ≥1 asthma hospital visit at the UWM in 2019, we computed the number of days in advance that our model gave an initial warning. For every such patient erroneously predicted to have ≥1 asthma hospital visit at the UWM in 2019, we used PreManage and the UWM data to check whether the patient had ≥1 asthma hospital visit outside of the UWM in 2019 or any surrogate indicators of poor outcomes. Such surrogate indicators included a prescription for systemic corticosteroids during the following 12 months, any type of visit for asthma exacerbation during the following 12 months, and asthma hospital visits between 13 and 24 months later.

Results: Among the 218 asthma patients in 2018 with asthma hospital visits at the UWM in 2019, 61.9% (135/218) were given initial warnings of such visits ≥3 months ahead by our model and 84.4% (184/218) were given initial warnings ≥1 day ahead. Among the 1310 asthma patients in 2018 who were erroneously predicted to have asthma hospital visits at the UWM in 2019, 29.01% (380/1310) had asthma hospital visits outside of the UWM in 2019 or surrogate indicators of poor outcomes.

Conclusions: Our model gave timely risk warnings for most asthma patients with poor outcomes. We found that 29.01% (380/1310) of asthma patients for whom our model gave false-positive predictions had asthma hospital visits somewhere else during the following 12 months or surrogate indicators of poor outcomes, and thus were reasonable candidates for preventive interventions. There is still significant room for improving our model to give more accurate and more timely risk warnings.

International Registered Report Identifier (IRRID): RR2-10.2196/5039

(JMIR Med Inform 2022;10(6):e38220) doi:10.2196/38220

KEYWORDS

asthma; machine learning; clinical decision support; forecasting; patient care management; healthcare outcome; emergency department; health outcome; prediction model

Introduction

Background

Over 262 million people in the world have asthma [1]. In the United States, around 7.8% of people have asthma, which leads to 1.6 million emergency department (ED) visits, 179,000 inpatient stays [2], and an aggregate medical cost of US $50.3 billion annually [3]. A main goal in asthma management is to curtail asthma hospital visits, ie, ED visits and inpatient stays for asthma. Part of the state of the art for achieving this goal is to implement a predictive model to find patients who are at significant risk of having asthma hospital visits in the future. If deemed high risk, a patient can be considered for enrollment in a care management program to receive preventive interventions. Then a care manager regularly follows up with the patient to monitor the patient's asthma control status, alter the patient's asthma medications as the need arises, and help book relevant services. This approach is employed by many health care systems, such as Intermountain Healthcare, the University of Washington Medicine (UWM), and Kaiser Permanente Northern California [4], along with many health plans, such as the health plans in 9 of 12 urban communities [5]. When used properly, this approach can curtail asthma hospital visits by up to 40% [5-9].

A care management program typically accommodates no more than 3% of patients due to capacity constraints [10]. To optimize the efficacy of such programs, we recently employed extreme gradient boosting (XGBoost) [11], a machine learning algorithm, and the UWM data to build the world's most accurate model to forecast which asthma patients will have asthma hospital visits during the following 12 months [12]. Our model obtained an area under the receiver operating characteristic curve of 0.902, a specificity of 90.91% (13,115/14,426 patients), a sensitivity of 70.2% (153/218 patients), a positive predictive value of 10.45% (153/1464 patients), a negative predictive value of 99.51% (13,115/13,180 patients), and an accuracy of 90.6% (13,268/14,644 patients) [12]. Compared with every prior model for this prediction task [4,13-26], our model improved the area under the receiver operating characteristic curve by ≥10%.

Objectives

Currently, two questions remain regarding our model's performance. First, for a patient who will have asthma hospital visits in the future, how far in advance can our model make an initial identification of risk? Since any preventive intervention requires sufficient time to take effect [27,28], a model should identify the risk as early as possible to provide preventive interventions in time to avoid a poor outcome. Second, if our model erroneously predicts a patient to have ≥1 asthma hospital visit at the UWM during the following 12 months, how likely will the patient have ≥1 asthma hospital visit at a facility outside of the UWM or ≥1 surrogate indicator of a poor outcome? As our model was trained on the UWM data, it can only predict future asthma hospital visits at the UWM. The goal of this work was to answer these two questions. Part of the analysis that we conducted to answer the second question has previously been published as an abstract at the 2022 American Academy of Allergy, Asthma & Immunology Annual Meeting [29].

Methods

Study Elements Reused From Previous Work

The following parts were reused from our prior paper on model building using the UWM data [12]: patient cohort, features, prediction target, cutoff point for conducting binary classification, training set, test set, and predictive model.

Ethics Approval

The institutional review board of the UWM approved this retrospective cohort study (STUDY00000118).

Patient Cohort

As the biggest academic health care system in Washington State, the UWM maintains an enterprise data warehouse that stores clinical and administrative data from 12 clinics and 3 hospitals for adults. The patient cohort was composed of every adult asthma patient ≥18 years old who received care at any of the 15 UWM facilities between 2011 and 2018. A patient was deemed to have asthma in a given year if the patient's visit billing data in that year included ≥1 asthma diagnosis code according to the International Classification of Diseases (ICD) tenth revision (ie, code J45.x) or ninth revision (ie, code 493.1x, 493.0x, 493.9x, or 493.8x) [13,30]. This asthma case-finding method has been shown to strike the best balance between sensitivity and positive predictive value among several rule-based asthma case-finding methods, does not require the patient to have >1 year of historical data, and is suited for use in population health management [30]. Patients who died during that year were excluded.

Data Sets

Two data sets were used. The first data set was retrieved from the UWM's enterprise data warehouse. This data set held structured administrative and clinical data for visits by the patient cohort to the 15 UWM facilities from 2011 to 2020. The second data set came from a commercial product, PreManage (Collective Medical Technologies Inc) [31]. This data set contained structured visit and diagnosis data for ED visits and inpatient stays during 2019 by our patient cohort at every hospital in Washington State, as well as at many other American hospitals outside of Washington State.

Overview of Our Predictive Model

Prediction Target, Training Set, and Test Set

For an asthma patient at a given time point, the prediction target was whether the patient would have ≥1 asthma hospital visit during the following 12 months. The prediction was made based on the patient's data up to that time point. An asthma hospital visit was defined as an ED visit or an inpatient stay with a principal diagnosis of asthma (ICD tenth revision code J45.x or ICD ninth revision code 493.1x, 493.0x, 493.9x, or 493.8x). During model training and testing, for each patient with asthma in a given year, we used the data of that patient by the end of the year to predict the outcome of the patient in the following 12 months [12]. Since the prediction target was in the following 12 months, the UWM data between 2011 and 2019 provided 8 years of effective data for model training and testing. The

effective data from 2011 to 2017 were used as the training set for training our predictive model, and the effective data from 2018 were used as the test set for testing our model. To answer our study's two questions, we focused on the asthma patients in the test set (ie, the asthma patients in 2018), and examined the predictions made by our model for these patients. For the asthma patients in 2018 who were erroneously predicted to have asthma hospital visits at the UWM in 2019, the UWM data from 2020 were used to compute one of the surrogate indicators of poor outcomes.

Machine Learning Algorithm and Features

Our predictive model was constructed using 71 features and the XGBoost classification algorithm [11]. These 71 features are presented in the online multimedia appendix of our previous paper on model building using the UWM data [12]. The features were constructed using the attributes in our UWM data set, which cover diverse aspects such as diagnoses, patient demographics, vital signs, visits, laboratory tests, procedures, and medications. Two exemplary features are the number of days from the patient's most recent ED visit and the number of asthma diagnoses that the patient received in the previous 12 months. These 71 features were included in every data instance that was inputted to our predictive model.

Cutoff Point for Conducting Binary Classification

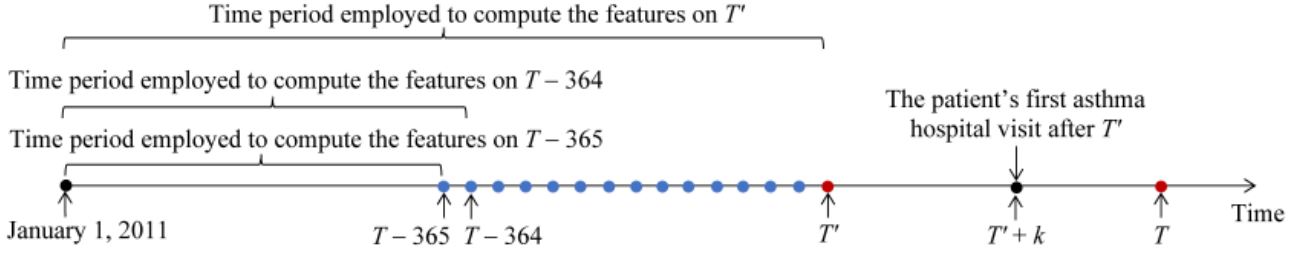
We set the cutoff point for conducting binary classification at the highest 10% of the risk scores computed by our model. Each patient with a risk score above this cutoff point was projected to have ≥1 asthma hospital visit during the following 12 months.

Assessing the Timeliness of the Initial Warnings of Risk Given by Our Model

Given a predictive model and an asthma patient in 2018 whose first asthma hospital visit in 2019 happened on date T, we measured k, the number of days in advance that our model gave an initial warning of risk. To compute k, we started from T - 365 and kept moving forward along the timeline to find the earliest date T' (T - 365 ≤ T' ≤ T - 1) such that by taking the feature values computed on the patient's historical data up to T' as inputs, the model would predict the patient to have ≥1 asthma hospital visit during the 12 months after T'. In this case, the model warned the patient's first asthma hospital visit after T' k ($1 \le k \le T - T'$) days in advance, with $T' + k$ being the starting date of the patient's first asthma hospital visit after T' (see Figure 1). Otherwise, if the model still predicted no future asthma hospital visit when we reached T - 1, the model warned the patient's asthma hospital visit on T k = 0 day in advance. The larger the value of k, the more timely the initial warning of risk that the model gave for the patient. k reflected how early before a poor outcome occurred the care manager would be prompted for the first time to consider giving the patient preventive interventions. The value of k was not affected by any prediction made by the model when the feature values computed based on the patient's historical data up to a given date after T' were taken as inputs.

For our predictive model, we computed k for every asthma patient in 2018 who had ≥1 asthma hospital visit at the UWM in 2019. We present the mean and the distribution of k.

Figure 1. Method of calculating k. T: the date on which the patient's first asthma hospital visit in 2019 happened. T': the earliest date between T - 365 and T - 1 such that by taking the feature values computed on the patient's historical data up to T' as inputs, the model would predict the patient to have ≥1 asthma hospital visit during the 12 months after T'. k: the number of days of advanced warning that the model gave for the patient for the first time.



Analyzing False-Positive Predictions Made by Our Model

For each asthma patient in 2018 whom our model erroneously predicted to have ≥1 asthma hospital visit at the UWM in 2019, we used PreManage data to check whether the patient had ≥1 asthma hospital visit outside of the UWM in 2019. We also used the UWM data to check whether the patient had any surrogate indicator of a poor outcome. Surrogate indicators of poor outcomes included a prescription for systemic corticosteroids during the following 12 months (ie, during 2019), any type of visit with a primary or principal diagnosis of asthma exacerbation during the following 12 months (ie, during 2019), and an asthma hospital visit between 13 and 24 months later (ie, during 2020). Systemic corticosteroids are used to treat asthma exacerbation. In addition, if the patient had ≥1 prescription for systemic corticosteroids in 2019, we computed the number of systemic corticosteroids ordered for the patient in 2019 counting multiplicity. This number partially reflected how poorly the patient's asthma was controlled. We present the distribution of this number.

Results

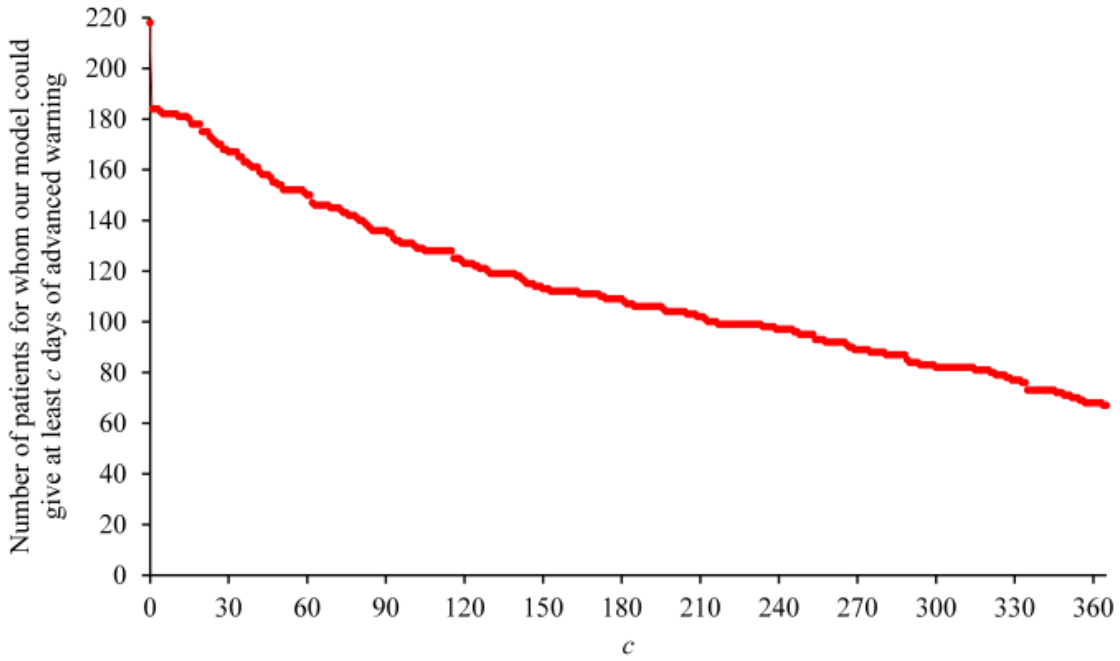
Clinical Characteristics and Demographics of Our Patient Cohort

Multimedia Appendix 1 shows the clinical characteristics and demographics for the UWM asthma patients, presented separately for the period between 2011 and 2017 and for 2018. Every data instance is linked to a distinct index year and patient pair and is used to project the outcome for the patient in the following 12 months. Our previous paper [12] included a detailed comparison of the clinical characteristics and demographics of the 2 sets of patients.

The Timeliness of Initial Warnings of Risk Given by Our Model

Of the 14,644 asthma patients in 2018, 218 (1.49%) had asthma hospital visits at the UWM in 2019. Figure 2 plots the distribution of the number of days in advance that our model gave an initial warning of an asthma hospital visit for every such patient. Our model gave a mean of 190 (SD 150) days of advanced warning. Our model gave an initial warning of risk ≥12 months in advance for 67 of these 218 (30.7%) patients, ≥6 months in advance for 107 of these 218 (49.1%) patients, ≥3 months in advance for 135 of these 218 (61.9%) patients, ≥1 month in advance for 167 of these 218 (76.6%) patients, ≥2 weeks in advance for 181 of these 218 (83%) patients, and ≥1 day in advance for 184 of these 218 (84.4%) patients.

Figure 2. The number of patients for whom our model could give at least *c* days of advanced warning versus *c* ($0 \leq c \leq 365$) among the 218 patients with asthma in 2018 who had asthma hospital visits at the University of Washington Medicine in 2019.



Breakdown of False-Positive Predictions Made by Our Model

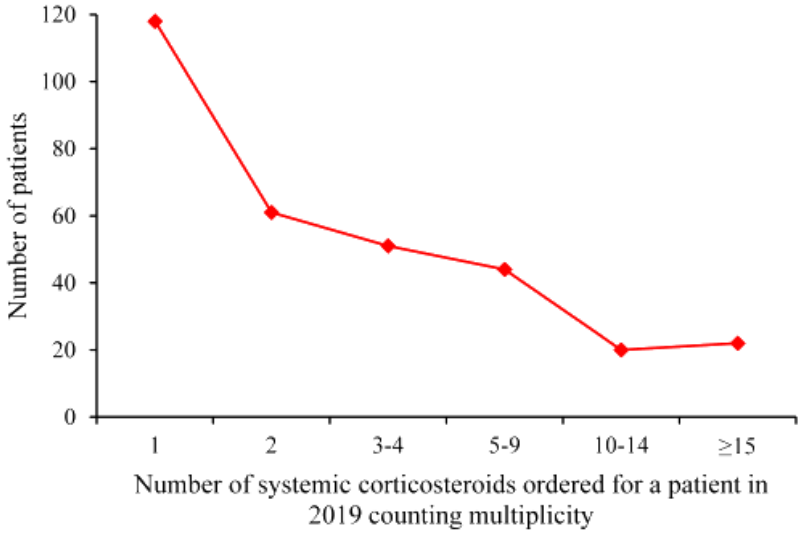
Our model erroneously predicted that 1310 asthma patients in 2018 would have asthma hospital visits at the UWM in 2019 [12]. Table 1 shows the number of these patients who had ≥1 asthma hospital visit outside of the UWM in 2019 or ≥1 surrogate indicator of a poor outcome.

In total, 316 asthma patients in 2018 were erroneously predicted by our model to have ≥1 asthma hospital visit at the UWM in 2019 and also had ≥1 prescription for systemic corticosteroids in 2019. Figure 3 plots the distribution of the number of systemic corticosteroids ordered for every such patient in 2019 counting multiplicity. The maximum value of this number was 118.

Table 1. The number of patients (N=1310) who had ≥1 asthma hospital visit outside of the University of Washington Medicine (UWM) in 2019 or ≥1 surrogate indicator of a poor outcome among the 1310 asthma patients in 2018 whom our model erroneously predicted to have asthma hospital visits at the UWM in 2019.

Outcome	Patients, n (%)
(1) At least 1 prescription for systemic corticosteroids during the following 12 months	316 (24.12)
(2) Any type of visit with a primary or principal diagnosis of asthma exacerbation during the following 12 months	126 (9.62)
(3) Asthma hospital visit between 13 and 24 months later (ie, during 2020)	18 (1.37)
(4) At least 1 asthma hospital visit outside of the UWM during the following 12 months	39 (2.98)
Any of (1), (2), and (3)	358 (27.33)
Any of (1), (2), (3), and (4)	380 (29.01)

Figure 3. The distribution of the number of systemic corticosteroids ordered for every patient in 2019 counting multiplicity among the 316 asthma patients in 2018 who were erroneously predicted by our model to have ≥1 asthma hospital visit at the University of Washington Medicine in 2019 and also had ≥1 prescription for systemic corticosteroids in 2019.



Discussion

Principal Results

Among the 218 asthma patients in 2018 who had asthma hospital visits at the UWM in 2019, the number of patients for whom our model could give at least c days of advanced warning decreased roughly linearly with c ($0 \leq c \leq 365$) at a fast pace. Our model gave timely risk warnings (eg, ≥3 months in advance) for a large proportion of these 218 asthma patients. Nevertheless, for another large proportion of these 218 asthma patients, our model could not give a timely risk warning. The model either gave a risk warning that was at most a few days in advance or did not predict a patient's risk even on the day before an asthma hospital visit.

Among the 1310 asthma patients in 2018 whom our model erroneously predicted to have asthma hospital visits at the UWM in 2019, 380 (29.01%) had asthma hospital visits outside of the UWM in 2019 or surrogate indicators of poor outcomes, and hence were reasonable candidates for preventive interventions. Among the 316 of these patients who had ≥1 prescription for systemic corticosteroids in 2019, a large proportion had rather poor asthma control, as reflected by a nontrivial number of systemic corticosteroids that were ordered for these patients in 2019.

Are the Initial Warnings of Risk Given by Our Model Timely Enough?

A predictive model should identify the risk of having future asthma hospital visits as early as possible in order to give the patient preventive interventions in time to avoid a poor outcome. The time needed for a preventive intervention to take effect varies with the intervention. To the best of our knowledge, there is no consensus on the amount of time needed for a particular preventive intervention or a particular combination of preventive interventions to take effect for averting future asthma hospital visits. Consequently, in this study, we could not compute the exact percentage of patients with future asthma hospital visits for whom our model could give timely risk warnings. Nevertheless, we can shed some light on the rough range of this percentage. In a prior study [27,28], several clinicians gave the opinion that up to 3 months could be needed for any intervention to take effect for averting inpatient stays for an ambulatory care-sensitive, chronic condition such as asthma. For 135 of the 218 (61.9%) asthma patients in 2018 who had asthma hospital visits at the UWM in 2019, our model was able to give an initial warning of risk ≥3 months in advance. Accordingly, we expect that the percentage of patients with future asthma hospital visits for whom our model could give a timely risk warning was at least 61.9%, which is large. On the other hand, for 34 of the 218 (15.6%) asthma patients in 2018 who had asthma hospital visits at the UWM in 2019, our model could not foresee the patient's risk even on the day before the visit. Thus, the percentage of patients with future asthma hospital visits for whom our model could not give a timely risk warning was at least 15.6%, which is also large. Combining these two findings, we estimate that the percentage of patients with future asthma hospital visits for whom our model could give a timely risk warning was somewhere between 61.9% and 84.4%. Thus, there is still significant room for improving our model to give more timely risk warnings.

Potential Impact of False-Positive Predictions Made by Our Model

We previously developed an automated method to supply rule-style explanations for the predictions that an arbitrary machine learning model makes on tabular data and to suggest tailored interventions [32,33]. Whenever our model gave a risk warning for a patient, we could use this method to help clinicians decide whether the patient should be enrolled in a care management program, should receive other less-expensive preventive interventions, or did not need any preventive intervention. For 134 of the 153 (87.6%) asthma patients in 2018 whom our model accurately predicted to have asthma hospital visits at the UWM in 2019, our method supplied rule-style explanations for the predictions made by the model [32]. Each such explanation included ≥1 modifiable risk factor

and linked to ≥1 intervention [32]; nevertheless, the situation could be different for other prediction targets or health care systems.

We found that among the 1310 asthma patients in 2018 whom our model erroneously predicted to have asthma hospital visits at the UWM in 2019, 380 (29.01%) had asthma hospital visits outside of the UWM in 2019 or surrogate indicators of poor outcomes. These patients could have benefited from the information provided by our automated explanation method. For the other 930 of the 1310 (70.99%) asthma patients in 2018 whom our model erroneously predicted to have asthma hospital visits at the UWM in 2019, our model's predictions could be truly inaccurate, leaving significant room for improving our model's accuracy. To know how many of these predictions would mislead clinicians into making incorrect intervention decisions, we would need to perform a user study with clinicians. This is left as an area of interest for future work.

Related Work

To the best of our knowledge, no prior study has used either surrogate indicators of poor outcomes or future asthma hospital visits at other hospitals to analyze the false-positive predictions made by a predictive model for asthma hospital visits. Also, no prior study has assessed the timeliness of the initial warnings of risk given by such a model. For predicting *Clostridium difficile* infection during an inpatient stay, Wiens et al [34] measured the number of days of advanced warning that a model gave on the patient. For predicting the total amount of donations that a fundraiser could obtain on a medical crowdfunding platform, Wang et al [35] measured the prediction timeliness based on the number of days of input data that a model needed in order to produce predictions within a certain percentage error rate and with a given level of confidence. For predicting the onset of sepsis, Guan et al [36] and Lauritsen et al [37] showed how model accuracy varied by the amount of time from when the model made a prediction to when sepsis occurred. Sepsis is an acute condition, whereas asthma is a chronic condition.

Limitations

This study has 5 limitations. First, this study was performed in a single health care system. In the future, we plan to use data from other health care systems to perform similar error and timeliness analyses on predicting asthma hospital visits [38,39].

Second, this study shows that many false-positive predictions made by our model could be truly inaccurate. While this study did not examine the factors that could have caused our model to make incorrect predictions, future work to investigate these factors could help improve model performance.

Third, although the PreManage data set covers every hospital in Washington State and many other American hospitals outside of Washington State, the data set does not cover every hospital in the United States. Consequently, our computational results on asthma hospital visits outside of the UWM in 2019 might have missed a small number of asthma patients in 2018 who had asthma hospital visits in 2019 that were outside of the UWM and whose data were unavailable in PreManage.

Fourth, our 3 surrogate indicators of poor outcomes were computed based on the UWM data. Consequently, our computational results for these surrogate indicators missed the asthma patients in 2018 who had surrogate indicators of poor outcomes outside of the UWM.

Fifth, this study computed the number of days in advance that our model gave an initial warning of an asthma hospital visit for a patient. This number reflected how early before a poor outcome a care manager could be prompted for the first time to consider giving the patient preventive interventions. However, it is currently unknown how likely the care manager would take action after receiving such a warning. This is worth studying in future work.

Conclusions

This study analyzed the errors and timeliness of the risk warnings given by our model for predicting asthma hospital visits. Our results show that our model gave timely risk warnings for most asthma patients with poor outcomes. We found that 380 of the 1310 (29.01%) asthma patients for whom our model gave false-positive predictions had asthma hospital visits outside of our health care system during the following 12 months or surrogate indicators of poor outcomes, and hence were reasonable candidates for preventive interventions. There is thus still significant room for improving our model to give more accurate and more timely risk warnings, such as by using predictive and comprehensible temporal features semiautomatically extracted from longitudinal medical data [35,40,41].

Acknowledgments

We thank Brian Kelly for the helpful discussions. GL was partially supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health (award number R01HL142503). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

XZ took part in the study design and the literature review, performed the computer coding and the experiments, and wrote the first draft of the paper. GL conceptualized and designed the study, performed the literature review, and rewrote the entire paper. Both authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The summary statistics of the clinical characteristics and the demographics of the University of Washington Medicine patients with asthma.
[PDF File (Adobe PDF File), 69 KB - medinform_v10i6e38220_app1.pdf]

References

1. Chronic respiratory diseases: asthma. World Health Organization. 2021. URL: https://www.who.int/news-room/q-a-detail/chronic-respiratory-diseases-asthma [accessed 2022-03-22]
2. Most recent national asthma data. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/asthma/most_recent_national_asthma_data.htm [accessed 2022-03-22]
3. Nurmagambetov T, Kuwahara R, Garbe P. The economic burden of asthma in the United States, 2008-2013. Ann Am Thorac Soc 2018 Mar;15(3):348-356. [doi: 10.1513/AnnalsATS.201703-259OC] [Medline: 29323930]
4. Lieu TA, Quesenberry CP, Sorel ME, Mendoza GR, Leong AB. Computer-based models to identify high-risk children with asthma. Am J Respir Crit Care Med 1998 Apr;157(4 Pt 1):1173-1180. [doi: 10.1164/ajrccm.157.4.9708124] [Medline: 9563736]
5. Mays GP, Claxton G, White J. Managed care rebound? Recent changes in health plans' cost containment strategies. Health Aff (Millwood) 2004;Suppl Web Exclusives:W4-427-436. [doi: 10.1377/hlthaff.w4.427] [Medline: 15451964]
6. Caloyeras JP, Liu H, Exum E, Broderick M, Mattke S. Managing manifest diseases, but not health risks, saved PepsiCo money over seven years. Health Aff (Millwood) 2014 Jan;33(1):124-131. [doi: 10.1377/hlthaff.2013.0625] [Medline: 24395944]
7. Greineder DK, Loane KC, Parks P. A randomized controlled trial of a pediatric asthma outreach program. J Allergy Clin Immunol 1999 Mar;103(3 Pt 1):436-440. [doi: 10.1016/s0091-6749(99)70468-9] [Medline: 10069877]
8. Kelly CS, Morrow AL, Shults J, Nakas N, Strope GL, Adelman RD. Outcomes evaluation of a comprehensive intervention program for asthmatic children enrolled in Medicaid. Pediatrics 2000 May;105(5):1029-1035. [doi: 10.1542/peds.105.5.1029] [Medline: 10790458]
9. Axelrod RC, Zimbro KS, Chetney RR, Sabol J, Ainsworth VJ. A disease management program utilizing life coaches for children with asthma. J Clin Outcomes Manag 2001;8(6):38-42.
10. Axelrod RC, Vogel D. Predictive modeling in health plans. Dis Manag Health Outcomes 2003;11(12):779-787. [doi: 10.2165/00115677-200311120-00003]
11. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
12. Tong Y, Messinger AI, Wilcox AB, Mooney SD, Davidson GH, Suri P, et al. Forecasting future asthma hospital encounters of patients with asthma in an academic health care system: predictive model development and secondary analysis study. J Med Internet Res 2021 Apr 16;23(4):e22796 [FREE Full text] [doi: 10.2196/22796] [Medline: 33861206]
13. Schatz M, Cook EF, Joshua A, Petitti D. Risk factors for asthma hospitalizations in a managed care organization: development of a clinical prediction rule. Am J Manag Care 2003 Aug;9(8):538-547 [FREE Full text] [Medline: 12921231]
14. Grana J, Preston S, McDermott PD, Hanchak NA. The use of administrative data to risk-stratify asthmatic patients. Am J Med Qual 1997;12(2):113-119. [doi: 10.1177/0885713X9701200205] [Medline: 9161058]
15. Loymans RJ, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Assendelft WJ, Schermer TR, et al. Identifying patients at risk for severe exacerbations of asthma: development and external validation of a multivariable prediction model. Thorax 2016 Sep;71(9):838-846. [doi: 10.1136/thoraxjnl-2015-208138] [Medline: 27044486]
16. Eisner MD, Yegin A, Trzaskoma B. Severity of asthma score predicts clinical outcomes in patients with moderate to severe persistent asthma. Chest 2012 Jan;141(1):58-65. [doi: 10.1378/chest.11-0020] [Medline: 21885725]
17. Sato R, Tomita K, Sano H, Ichihashi H, Yamagata S, Sano A, et al. The strategy for predicting future exacerbation of asthma using a combination of the Asthma Control Test and lung function test. J Asthma 2009 Sep;46(7):677-682. [doi: 10.1080/02770900902972160] [Medline: 19728204]
18. Osborne ML, Pedula KL, O'Hollaren M, Ettinger KM, Stibolt T, Buist AS, et al. Assessing future need for acute care in adult asthmatics: the Profile of Asthma Risk Study: a prospective health maintenance organization-based study. Chest 2007 Oct;132(4):1151-1161. [doi: 10.1378/chest.05-3084] [Medline: 17573515]
19. Miller MK, Lee JH, Blanc PD, Pasta DJ, Gujrathi S, Barron H, TENOR Study Group. TENOR risk score predicts healthcare in adults with severe or difficult-to-treat asthma. Eur Respir J 2006 Dec;28(6):1145-1155 [FREE Full text] [doi: 10.1183/09031936.06.00145105] [Medline: 16870656]
20. Peters D, Chen C, Markson LE, Allen-Ramey FC, Vollmer WM. Using an asthma control questionnaire and administrative data to predict health-care utilization. Chest 2006 Apr;129(4):918-924. [doi: 10.1378/chest.129.4.918] [Medline: 16608939]
21. Yurk RA, Diette GB, Skinner EA, Dominici F, Clark RD, Steinwachs DM, et al. Predicting patient-reported asthma outcomes for adults in managed care. Am J Manag Care 2004 May;10(5):321-328 [FREE Full text] [Medline: 15152702]

22. Loymans RJB, Debray TPA, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Schermer TRJ, et al. Exacerbations in adults with asthma: a systematic review and external validation of prediction models. J Allergy Clin Immunol Pract 2018;6(6):1942-1952.e15. [doi: 10.1016/j.jaip.2018.02.004] [Medline: 29454163]
23. Lieu TA, Capra AM, Quesenberry CP, Mendoza GR, Mazar M. Computer-based models to identify high-risk adults with asthma: is the glass half empty of half full? J Asthma 1999 Jun;36(4):359-370. [doi: 10.3109/02770909909068229] [Medline: 10386500]
24. Schatz M, Nakahiro R, Jones CH, Roth RM, Joshua A, Petitti D. Asthma population management: development and validation of a practical 3-level risk stratification scheme. Am J Manag Care 2004 Jan;10(1):25-32 [FREE Full text] [Medline: 14738184]
25. Forno E, Fuhlbrigge A, Soto-Quirós ME, Avila L, Raby BA, Brehm J, et al. Risk factors and predictive clinical scores for asthma exacerbations in childhood. Chest 2010 Nov;138(5):1156-1165 [FREE Full text] [doi: 10.1378/chest.09-2426] [Medline: 20472862]
26. Xiang Y, Ji H, Zhou Y, Li F, Du J, Rasmy L, et al. Asthma exacerbation prediction and risk factor analysis based on a time-sensitive, attentive neural network: retrospective cohort study. J Med Internet Res 2020 Jul 31;22(7):e16981 [FREE Full text] [doi: 10.2196/16981] [Medline: 32735224]
27. Longman JM, Passey ME, Ewald DP, Rix E, Morgan GG. Admissions for chronic ambulatory care sensitive conditions - a useful measure of potentially preventable admission? BMC Health Serv Res 2015 Oct 16;15:472 [FREE Full text] [doi: 10.1186/s12913-015-1137-0] [Medline: 26475293]
28. Johnston JJ, Longman JM, Ewald DP, Rolfe MI, Diez Alvarez S, Gilliland AHB, et al. Validity of a tool designed to assess the preventability of potentially preventable hospitalizations for chronic conditions. Fam Pract 2020 Jul 23;37(3):390-394 [FREE Full text] [doi: 10.1093/fampra/cmz086] [Medline: 31848589]
29. Zhang X, Luo G. Error analysis of machine learning predictions on asthma hospital encounters. J Allergy Clin Immunol 2022 Feb;149(2):Supplement, AB47. [doi: 10.1016/j.jaci.2021.12.184]
30. Howell D, Rogers L, Kasarskis A, Twyman K. Comparison and validation of algorithms for asthma diagnosis in an electronic medical record system. Ann Allergy Asthma Immunol 2022 Mar 30;128(6):667-681. [doi: 10.1016/j.anai.2022.03.025] [Medline: 35367347]
31. Collective Medical and Consonus Healthcare announce partnership to improve postacute transitions of care. Collective Medical Technologies Inc. 2018. URL: https://collectivemedical.com/resources/press-release/collective-medical-and-consonus-healthcare-announce-partnership-to-improve-post-acute-transitions-of-care [accessed 2022-03-22]
32. Tong Y, Messinger AI, Luo G. Testing the generalizability of an automated method for explaining machine learning predictions on asthma patients' asthma hospital visits to an academic healthcare system. IEEE Access 2020;8:195971-195979 [FREE Full text] [doi: 10.1109/access.2020.3032683] [Medline: 33240737]
33. Zhang X, Luo G. Ranking rule-based automatic explanations for machine learning predictions on asthma hospital encounters in patients with asthma: retrospective cohort study. JMIR Med Inform 2021 Aug 11;9(8):e28287 [FREE Full text] [doi: 10.2196/28287] [Medline: 34383673]
34. Wiens J, Guttag JV, Horvitz E. Patient risk stratification with time-varying parameters: a multitask learning approach. J Mach Learn Res 2016;17(79):1-23 [FREE Full text]
35. Wang T, Jin F, Hu Y, Cheng Y. Early predictions for medical crowdfunding: a deep learning approach using diverse inputs. Arxiv Preprint posted online on November 9, 2019. [FREE Full text] [doi: 10.48550/arXiv.1911.05702]
36. Guan Y, Wang X, Chen X, Yi D, Chen L, Jiang X. Assessment of the timeliness and robustness for predicting adult sepsis. iScience 2021 Feb 19;24(2):102106 [FREE Full text] [doi: 10.1016/j.isci.2021.102106] [Medline: 33659874]
37. Lauritsen SM, Kalør ME, Kongsgaard EL, Lauritsen KM, Jørgensen MJ, Lange J, et al. Early detection of sepsis utilizing deep learning on electronic health record event sequences. Artif Intell Med 2020 Apr;104:101820 [FREE Full text] [doi: 10.1016/j.artmed.2020.101820] [Medline: 32498999]
38. Luo G, Nau CL, Crawford WW, Schatz M, Zeiger RS, Rozema E, et al. Developing a predictive model for asthma-related hospital encounters in patients with asthma in a large, integrated health care system: secondary analysis. JMIR Med Inform 2020 Nov 09;8(11):e22689 [FREE Full text] [doi: 10.2196/22689] [Medline: 33164906]
39. Luo G, He S, Stone BL, Nkoy FL, Johnson MD. Developing a model to predict hospital encounters for asthma in asthmatic patients: secondary analysis. JMIR Med Inform 2020 Jan 21;8(1):e16080 [FREE Full text] [doi: 10.2196/16080] [Medline: 31961332]
40. Luo G. A roadmap for semi-automatically extracting predictive and clinically meaningful temporal features from medical data for predictive modeling. Glob Transit 2019;1:61-82 [FREE Full text] [doi: 10.1016/j.glt.2018.11.001] [Medline: 31032483]
41. Luo G, Stone BL, Koebnick C, He S, Au DH, Sheng X, et al. Using temporal features to provide data-driven clinical early warnings for chronic obstructive pulmonary disease and asthma care management: protocol for a secondary analysis. JMIR Res Protoc 2019 Jun 06;8(6):e13783 [FREE Full text] [doi: 10.2196/13783] [Medline: 31199308]

Abbreviations

ED: emergency department
ICD: International Classification of Diseases
UWM: University of Washington Medicine
XGBoost: extreme gradient boosting

Edited by C Lovis; submitted 24.03.22; peer-reviewed by M Hofford; comments to author 15.04.22; revised version received 16.04.22; accepted 13.05.22; published 08.06.22.

Please cite as:
Zhang X, Luo G
Error and Timeliness Analysis for Using Machine Learning to Predict Asthma Hospital Visits: Retrospective Cohort Study
JMIR Med Inform 2022;10(6):e38220
URL: https://medinform.jmir.org/2022/6/e38220
doi:10.2196/38220
PMID:35675129

Original Paper

The Prediction of Preterm Birth Using Time-Series Technology-Based Machine Learning: Retrospective Cohort Study

Yichao Zhang[1*], MSc; Sha Lu[2,3*], MD; Yina Wu[1], MEng; Wensheng Hu[2,3], MD; Zhenming Yuan[1], PhD

[1]Hangzhou Normal University, Hangzhou, China

[2]Department of Obstetrics and Gynecology, Hangzhou Women's Hospital, Hangzhou, China

[3]Department of Obstetrics and Gynecology, The Affiliated Hangzhou Women's Hospital of Hangzhou Normal University, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Zhenming Yuan, PhD
Hangzhou Normal University
2318 Yuhangtang Road
Hangzhou, 311121
China
Phone: 86 13588714850
Email: zmyuan@hznu.edu.cn

Abstract

Background: Globally, the preterm birth rate has tended to increase over time. Ultrasonography cervical-length assessment is considered to be the most effective screening method for preterm birth, but routine, universal cervical-length screening remains controversial because of its cost.

Objective: We used obstetric data to analyze and assess the risk of preterm birth. A machine learning model based on time-series technology was used to analyze regular, repeated obstetric examination records during pregnancy to improve the performance of the preterm birth screening model.

Methods: This study attempts to use continuous electronic medical record (EMR) data from pregnant women to construct a preterm birth prediction classifier based on long short-term memory (LSTM) networks. Clinical data were collected from 5187 pregnant Chinese women who gave birth with natural vaginal delivery. The data included more than 25,000 obstetric EMRs from the early trimester to 28 weeks of gestation. The area under the curve (AUC), accuracy, sensitivity, and specificity were used to assess the performance of the prediction model.

Results: Compared with a traditional cross-sectional study, the LSTM model in this time-series study had better overall prediction ability and a lower misdiagnosis rate at the same detection rate. Accuracy was 0.739, sensitivity was 0.407, specificity was 0.982, and the AUC was 0.651. Important-feature identification indicated that blood pressure, blood glucose, lipids, uric acid, and other metabolic factors were important factors related to preterm birth.

Conclusions: The results of this study will be helpful to the formulation of guidelines for the prevention and treatment of preterm birth, and will help clinicians make correct decisions during obstetric examinations. The time-series model has advantages for preterm birth prediction.

(JMIR Med Inform 2022;10(6):e33835) doi:10.2196/33835

KEYWORDS

preterm birth prediction; temporal data mining; electronic medical records; pregnant healthcare

Introduction

Background

Preterm birth, defined as birth occurring before 37 weeks of completed gestation, is the primary cause of neonatal death and disability and affects the long-term health of newborns [1,2]. According to the World Health Organization global action report on preterm birth, there are approximately 15 million premature infants born in the world every year, with an incidence rate of 5% to 18%; 1 million of these premature infants die [3]. China is the most populous country in the world, and the implementation of the two-child policy has increased the average age of first pregnancy and the incidence of preterm birth [4-6]. Compared to full-term birth, prematurity imposes adverse effects

on the health and safety of both the pregnant woman and the infant. Prematurity increases the incidence of congenital malformation, being small for gestational age, and nervous system diseases associated with immature organs [7-9]. Therefore, early prediction of preterm birth and preventive measures have a significant potential to reduce mortality and improve the survival rate of preterm infants [10,11].

Despite the serious clinical consequences, there are currently no effective early screening methods for preterm birth. It is generally considered that ultrasonography cervical-length assessment is the most effective screening method [11,12], but routine, universal cervical-length screening remains controversial because of its cost [13,14]. Cervical screening is not popular in China and is performed only for pregnant women with cervical insufficiency [15]. Fetal fibronectin is an extracellular matrix glycoprotein that has also been extensively studied as a predictor of preterm birth, and although it has high specificity, it has a low detection rate [16]. Other biomarkers, including inflammatory factors, serum proteomics, and genetic factors, are associated with preterm birth [17], but each of these only has good performance in a subset of cases, and few studies have demonstrated that they are sufficiently useful for clinical use.

There is not a single or combined screening method for preterm birth that has high sensitivity and can reliably identify women at risk for preterm birth [11]. The etiological mechanism of preterm birth is elusive, and the interaction between risk factors is complex. Machine learning algorithms based on time-series technology can solve nonlinear relationships between multi-dimensional variables and analyze and mine their time-series characteristics. These machine learning models have been shown to be effective in the prediction of obstetric diseases [18,19]. Therefore, this paper proposes a time-series preterm birth prediction model based on a long short-term memory (LSTM) network.

Related Work

In the literature, various methods have been proposed to predict the risk of preterm birth with machine learning. These methods can be broadly categorized into 2 types, according to their data source: special examination data or routine clinical data. Special examination data include findings from the cervicovaginal fluid [20], electrohysterography [21], and whole-blood gene expression [22]. These data need special methods to obtain and are not suitable for large-scale initial screening. Therefore, research results based on these data have only been shown to have better prediction performance in small-sample data sets. Other research has sought to build prediction models based on routine clinical examination data and demographic data. Koivu et al [23] used a US Centers for Disease Control and Prevention (CDC) data set of almost sixteen million observations to build a prediction model; the best-performing machine learning model achieved an area under the curve (AUC) of 0.64 for preterm birth when using external the New York City test data. Lee et al [24] used the same CDC and New York City data sets to build an artificial neural network prediction model; it also had an AUC of 0.64. Weber et al [25] assessed the prediction of early (<32 weeks) spontaneous preterm birth among non-Hispanic women by applying machine learning to multilevel data from a large birth cohort; the AUC of this prediction model was 0.67.

Although the above prediction models have relatively reliable performance, they all use huge, complex data sets for analysis. It can be difficult to obtain complete data sets of this size and complexity because of privacy issues. More importantly, these models ignore the influence of time-related factors. Time-series analysis and prediction methods predict future developments according to tendencies in past changes and highlight the role of time factors in making predictions. In fact, obstetric examinations are continuous and repeated time-series records and are considered to be related to pregnancy risk [26]. Previous studies have reported that time-series models perform well in the field of obstetrics. For example, Tao et al [27] used maternal weight change trajectories during pregnancy to establish a time-series hybrid model to predict the birth weight of newborns. Zhou et al [28] predicted the risk of postpartum hemorrhage using continuous data from prenatal physical examinations. Compared with other biological phenomena, the 280-day gestational cycle has a relatively fixed time; pregnant women also have high compliance to obstetric outpatient examinations [29]. Therefore, a time-series model to mine time-series characteristics from data obtained during pregnancy has high potential.

Few studies have described the interpretability of their models. Khatibi et al [30] used Iran's national databank of maternal and neonatal records to design a map/reduce phase-based, parallel feature selection machine learning algorithm to predict the risk of preterm birth. The map phase used parallel feature selection and classification methods to score features, while the reduce phase aggregated the feature scores in order to determine the contribution of predictors to the model. Similar methods include the calculation of frequency statistics, the Gini index and other indicators that trace the decision-making process of the tree model [31], and calculating Shapley values to define the importance of features [32].

Although none of the above methods are suitable for time-series models, it is encouraging that there have been recent proposals for interpretable frameworks for time-series classification that can be used in different medical scenarios. In the field of medical signals, Ivatur et al [33] proposed a post-hoc explainability framework for deep learning models applied to quasi-periodic biomedical time-series classification that included 3 different techniques for explanation: studying ablation, studying permutation, and using a local, interpretable model-agnostic explanation method. Maweu et al [34] proposed a modular framework named the convolutional neural network (CNN) explainability framework for electrocardiogram signals that explains the quality of the deep learning model in terms of quantifiable metrics and feature visualization. Electronic medical record (EMRs) contain time series and multimodal data which further hinder interpretability. Nguyen-Du et al [35] proposed a new deep electronic health record spotlight framework for transforming EMR data into pathways and 2D pathway images, which can then be used with 2D CNN techniques to support visual interpretation. Viton et al [36] proposed an approach based on heat maps as a visual means of highlighting significant

variables over a temporal sequence, which can be applied to the problem of predicting the risk of in-hospital mortality.

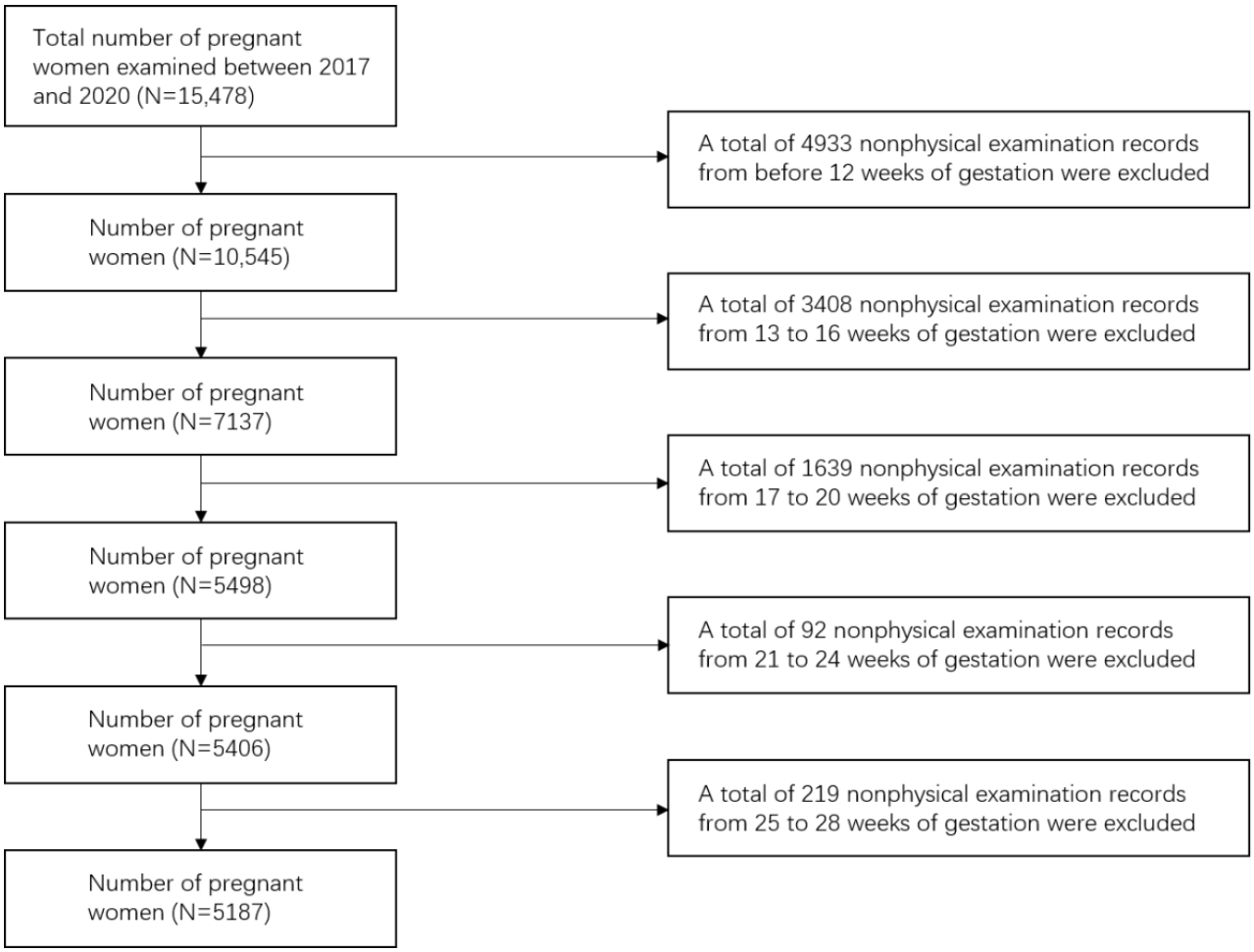
This previous research motivated the current study, which makes the following key contributions: (1) we designed and implemented a complete process for preterm birth screening and providing early warnings based on regular EMR data; (2) we used machine learning based on time-series technology to analyze the obstetric examination data and improve the performance of the prediction model; (3) we provide a preliminary explanation of the quantitative interpretability of the model and explore time-series predictors affecting preterm birth.

Methods

Setting and Study Population

The data were collected from Hangzhou Women's Hospital (Hangzhou Maternity and Child Health Care Hospital), Hangzhou, Zhejiang Province, China, between 2017 and 2020. This study included >25,000 pregnant women who received antenatal care at Hangzhou Women's Hospital and eventually gave birth naturally through the vagina. The exclusion criteria were as follows: presence of multiple pregnancies, assisted reproduction, severe cardio- or cerebrovascular complications or comorbidities, and performance of cervical cerclage during pregnancy. The inclusion criterion was a first pregnancy test taken before 12 gestational weeks. According to the Chinese guidelines for prenatal examination [37], pregnant women should have a monthly outpatient examination before 28 weeks of gestation. Figure 1 shows the filtering and processing flow chart used to select the study population. Some women were excluded owing to failure to obtain data or implausible pregnancy outcomes. Data from a final total of 5187 women were available for analysis.

Figure 1. Flow chart showing participant selection.



Clinical Measurements and Data Collection

Demographic data, physical examination data, ultrasound records, and laboratory data from the antenatal period were retrieved from EMRs. At registration for pregnancy, information on maternal demographic characteristics (eg, age, education, and occupation), anthropometrics (eg, body weight, height, and blood pressure), and clinical history (eg, parity and disease history) were recorded. As shown in Table 1, repeated pregnancy data were obtained for each individual from the first pregnancy test to the final pregnancy test, taken between 25 to 28 weeks. The clinical data included age, weight, uterine height, abdominal circumference, blood pressure, and findings from ultrasonic examination. Laboratory tests (eg, routine blood examination and blood biochemistry examination, including blood lipids and glucose) were performed at 24 weeks of gestation.

Participants were asked to wear light clothing when their height and weight were measured. BMI was calculated as body weight in kilograms divided by body height in meters squared. Sitting blood pressure was examined after at least 10 minutes of rest using a standard mercury sphygmomanometer with the patient's right arm held at heart level. Maternal venous blood samples were drawn in the morning after an overnight fast of ≥8 hours.

Table 1. Description of data sources.

Gestational age	Ultrasonic examination	Laboratory tests
Before 12 weeks	✓[a]	N/A[b]
From 13 to 16 weeks	✓	N/A
From 17 to 20 weeks	✓	N/A
From 21 to 24 weeks	✓	N/A
From 25 to 28 weeks	✓	✓

[a]✓ indicates that the pregnant woman has made relevant clinical examination in this pregnancy stage.

[b]N/A: not applicable.

Model Design

Based on the above-mentioned features, 2 machine learning models were constructed to predict preterm birth. One was an early prediction model based on the data sources in Table 1. For each cross-sectional gestational age category, extreme gradient boosting (XGB) combined with decision trees was employed to establish the prediction model. XGB is an improvement on the gradient lifting algorithm and is widely used in the field of obstetric auxiliary diagnosis [38]. The second model used temporal prediction techniques. Long short-term memory networks (LSTMs) are a type of time-cyclic neural network that are suitable for processing and predicting events with relatively long intervals and delays in the time series [39]. LSTMs can avoid the gradient disappearance of conventional recurrent neural networks and are widely used in the field of disease diagnosis [40].

LSTMs realize information protection and control through 3 control gates, namely the input gate, the forgetting gate, and the output gate. The key in LSTMs is the unit state. The LSTM unit judges whether the output of the previous time step is useful; only useful information is saved and the rest is forgotten at the forgetting gate. Equations (1) through (5) represent the parameter update process, where σ represents the sigmoid function, h_{t-1} represents the output of the LSTM at the previous time step, and h_t represents the current output; *I*, *f*, and *o*, respectively, represent the input gate, forgetting gate, and output gate in the LSTM unit. Equation (4) represents the process of the state transition of the memory unit, where c_t is the state of the memory unit at the current time step. The current state is calculated by the previous time step state, c_{t-1}, and the result of the forgetting gate and the input gate of the current-time LSTM unit.

$$i_t = \sigma (W_{\chi i}\chi_t + W_{hi}h_{t-1} + b_i) \textbf{ (1)}$$

$$f_t = \sigma (W_{\chi f}\chi_t + W_{hf}h_{t-1} + b_f) \textbf{ (2)}$$

$$o_t = \sigma (W_{Xo}\chi_t + W_{ho}h_{t-1} + b_o) \textbf{ (3)}$$

$$C_t = f_t c_{t-1} + i_t \tanh(W_{xc}\chi_t + W_{hc}h_{t-1} + b_c) \textbf{ (4)}$$

$$h_t = o_t \tanh(c_t) \textbf{ (5)}$$

The parameters of these prediction models were determined by grid search. The models were validated with 5-fold cross-validation. The 5-fold cross-validation splits the training dataset into 2 sections, where 80% of the dataset is used for training and the remaining 20% is used for testing. Simultaneously, the incidence rate of preterm birth is about 5%, so in situations where there were imbalanced class data combined with unequal error costs, random oversampling was used to balance the dataset to get true performance values for the classifier. The random oversampling method makes the number of minority classes the same as the number of majority classes by randomly copying minority class samples to get new equilibrium data.

Under the Python 3.6 environment (Python Software Foundation), the data analysis and visualization were completed by using NumPy, Pandas, Matplotlib, Seaborn, and other libraries [41,42]. The machine learning model comes from the scikit-learn library and the deep learning framework adopts PyTorch [43]. Based on the amount of data in this study, the LSTM network was able to run on a personal computer. The adaptive learning rate of the Adam optimizer [44] was used to accelerate the convergence speed of the LSTM model. Table 2 shows the values of the parameters for the 2 models.

Table 2. Summary of parameter values in each model.

Parameters	Values
Extreme gradient boosting model	
Learning rate	0.01
N_estimators	200
Min_samples_leaf	4
Min_samples_split	3
Max_depth	2
Long short-term memory model	
Loss function	CrossEntropy
Num_layers	2
Optimizer	Adam
Hidden_size	130
Input size	65
Learning rate	0.001
Batch-size	256
Epochs	20

Model Evaluation

The characteristics were compared between the preterm birth and full-term birth groups. Statistical tests were 2-sided; *P* values <.05 were considered statistically significant. All analyses were performed using the statistical software SPSS 22.0 (IBM).

The prediction performance was considered an important factor to evaluate the proposed model. In this paper, the receiver operating characteristic (ROC) curve and AUC were used to evaluate the model's ability to predict preterm birth. In addition, the evaluation indicators of the confusion matrix, including accuracy, sensitivity, and specificity, were used to analyze the relationship between the actual values and the predicted values for the risk of preterm birth. Accuracy, sensitivity, and specificity were calculated as follows: accuracy = (TN + TP) / (TN + TP + FN + FP); sensitivity = TP / (TP + FN); and specificity = TN / (TN + FP), where TP indicates true positive, FP indicates false positive, TN indicates true negative, and FN indicates false negative.

Feature importance reflects the contribution each variable makes in classifying preterm birth, which explains the results of the model decision. In this study, feature importance for the XGB model was calculated by the sum of the decrease in error when split by a variable [31]. For the LSTM model, feature ablation was used, which provides feature importance at a given time step for each input feature [45], computing attribution as the difference in output after replacing each feature with a baseline; a lower AUC indicates a more important feature.

Ethics Approval

The study design was approved by the local Ethical and Research Committee (written permission, with approval number 2019-02-2). All medical procedures were performed following the relevant guidelines and regulations. The informed consent requirement for this study was waived by the board because the researchers only accessed the database for analysis purposes and all patient data were deidentified.

Results

General Characteristics of the Study Participants

The data set used in this paper comes from a hospital in eastern China and is very extensive, including maternal ultrasound records, prenatal examination reports, and laboratory data. Of the 5187 pregnant women enrolled in the present study, 4966 gave birth at full term. The remaining 221 women gave birth preterm. The general characteristics of the participants are presented in Table 3. Table 4 summarizes the clinical characteristics of the study subjects at the second trimester (25-28 weeks).

Table 3. General characteristics of the study population (N=5187)

Characteristics	Mean (SD)
Age, years	29.63 (3.52)
Prepregnancy weight, kg	53.65 (8.15)
Height, cm	161.45 (4.84)
Prepregnancy BMI, kg/m^2	20.57 (2.92)
Parity, number	0.26 (0.46)
Gravidity, number	1.71 (0.98)
Prepregnancy SBP[a], mmHg	106.12 (13.02)
Prepregnancy DBP[b], mmHg	67.29 (9.31)
Number of preterm births in reproductive history, parity number	0.003 (0.05)
Menarche, years	13.47 (1.22)
Period, days	6.07 (3.03)
Cycle, days	29.55 (7.06)

[a]Systolic blood pressure.

[b]Diastolic blood pressure.

Table 4. Clinical characteristics and laboratory parameters at the second trimester.

Characteristics	Full-term birth (n=4966) Mean (SD)	Preterm birth (n=221) Mean (SD)	*P* value
General characteristics			
Age, years	29.61 (3.49)	30.14 (3.64)	.02
Prepregnancy weight, kg	53.92 (7.16)	53.74 (8.12)	.31
Prepregnancy SBP[a], mmHg	106.70 (10.45)	106.19 (11.98)	.48
Prepregnancy DBP[b], mmHg	67.65 (7.96)	67.47 (7.41)	.53
Physical data			
Gestational age, weeks	26.02 (1.17)	26.09 (1.19)	.73
Pulse rate, beats per minute	77.63 (7.27)	77.32 (6.82)	.56
Maternal weight at pregnancy, kg	61.16 (7.28)	60.39 (8.29)	.29
SBP, mmHg	111.42 (10.62)	113.19 (11.24)	.04
DBP, mmHg	65.29 (7.78)	66.09 (8.20)	<.001
Uterine height, cm	24.48 (1.82)	24.02 (2.28)	.45
Mother abdominal circumference, cm	88.76 (5.45)	86.98 (8.33)	.45
Ultrasonic data			
Biparietal diameter, cm	6.70 (0.23)	6.84 (0.48)	.05
Head circumference, cm	24.60 (0.76)	25.02 (1.47)	.13
Femur length, cm	4.83 (0.17)	4.93 (0.34)	.06
Fetal abdominal circumference, cm	22.18 (0.86)	22.94 (1.45)	.03
Laboratory data			
Triglyceride, mmol/L	2.15 (0.78)	2.25 (0.79)	.02
Total bile acid, μmol/L	2.22 (1.75)	2.17 (1.52)	.43
Uric acid, μmol/L	244.05 (49.69)	246.05 (49.60)	.12
Platelets, cells × 10^9/L	209.12 (45.24)	212.26 (46.10)	.11
Fasting blood glucose, mmol/L	4.35 (0.38)	4.40 (0.46)	.04
Total cholesterol, mmol/L	6.23 (1.01)	6.19 (1.07)	.28
Activated partial thromboplastin time, seconds	26.25 (2.97)	26.26 (3.31)	.75
Fibrinogen, g/L	3.77 (0.63)	3.85 (0.64)	.03
Hemoglobin, g/L	115.96 (8.44)	116.79 (8.61)	.04

[a]Systolic blood pressure.

[b]Diastolic blood pressure.

Model Performance

Based on the above-mentioned features in Table 3 and Table 4, 2 machine learning models were constructed to predict preterm birth. An XGB model was used for cross-sectional research and an LSTM model was used for time-series research. The optimal parameters were set for each predictive model and corroborated via a test data set that was derived from the training data set by 5-fold cross-validation. The accuracy, sensitivity, specificity, and AUC of the models for predicting preterm birth are shown in Table 5, which compares the performance of these 2 models in identical testing data sets. Notably, the LSTM model, used for time-series research, had the best overall prediction ability. Its accuracy, sensitivity, specificity, and AUC were 0.739, 0.407, 0.982, and 0.651, respectively. Furthermore, the model performance gradually improved with the number of gestational weeks. The overall performance of the model was best in the last cross-sectional gestational age group, with an overall accuracy of 0.689, sensitivity of 0.407, specificity of 0.979, and AUC of 0.601.

Based on the validation result for the training data set, an independent testing data set was used for predicting preterm birth. The matrices and ROC curves for the predictive models in the testing data set are shown in Figure 2. Compared with cross-sectional designs, the LSTM model had a lower misdiagnosis rate at the same detection rate. The high specificity

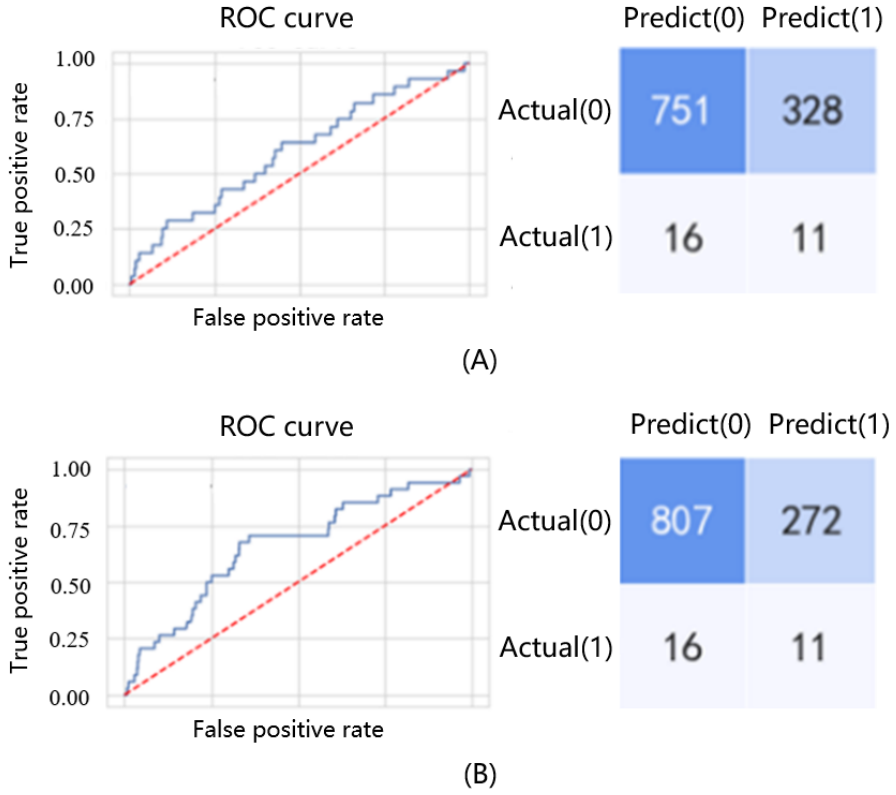
of the model excluded more true negative samples, lowering the cost of screening.

Table 5. Average prediction results of different methods after 5-fold cross-validation.

Prediction results	Observation period (gestational weeks)					Time series
	Before 12 weeks	Weeks 13-16	Weeks 17-20	Weeks 21-24	Weeks 25-28	
AUC[a]	0.532	0.558	0.516	0.568	0.601	0.651
Sensitivity	0.286	0.365	0.362	0.387	0.407	0.407
Specificity	0.974	0.978	0.977	0.977	0.979	0.982
Accuracy	0.525	0.574	0.584	0.622	0.689	0.739

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

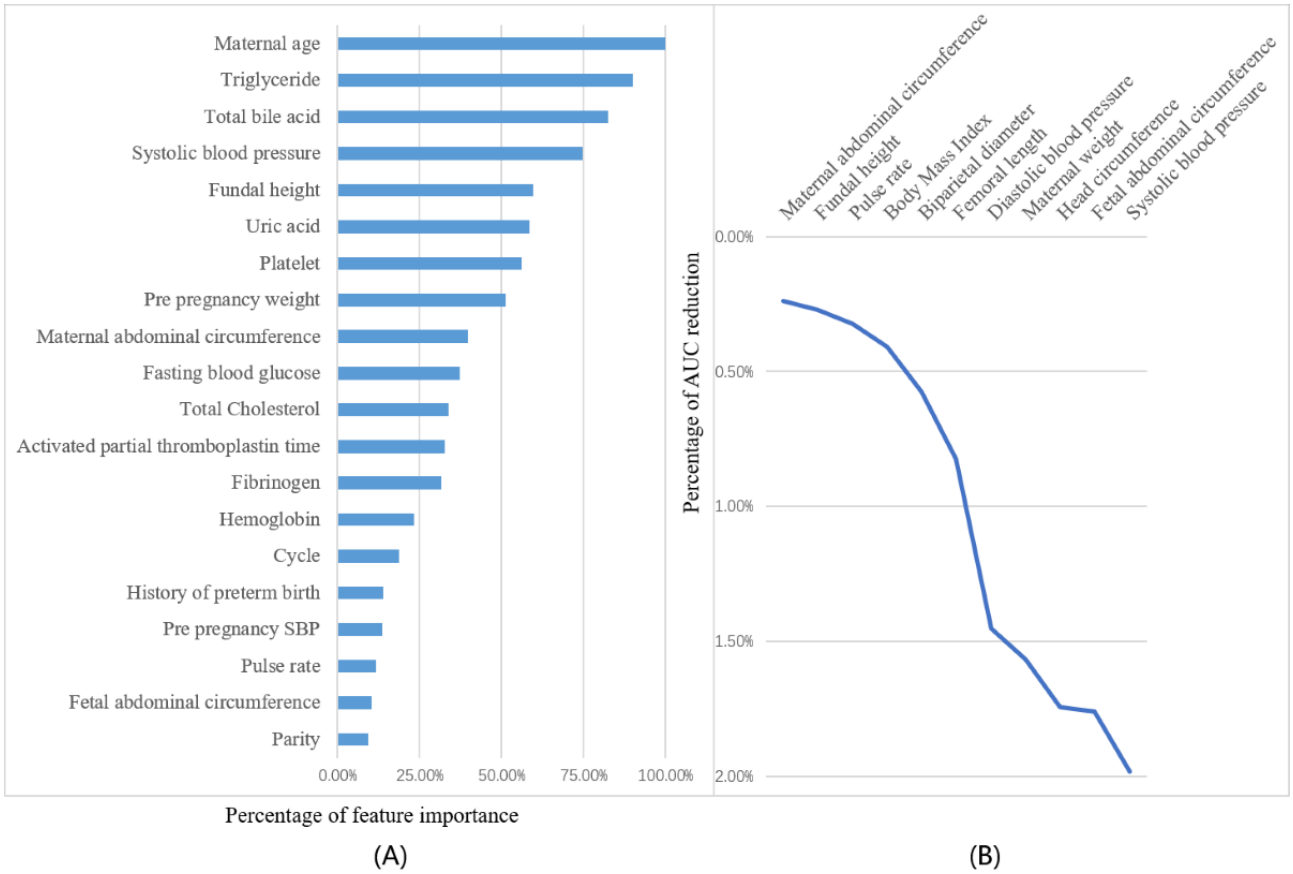
Figure 2. Receiver operating characteristic curves and confusion matrix of prediction models: (A) cross-sectional prediction of the extreme gradient boosting model at weeks 25 to 28; (B) prediction results of the long short-term memory model. ROC: receiver operating characteristic.



Influence of Variables on Predictions

The identification of important features by the XGB and LSTM models is shown in Figure 3. Feature importance was calculated by XGB as the sum of the decrease in error when split by a variable, which reflects the contribution each variable makes in classifying. Maternal age was the most important variable to predict preterm birth, followed by triglyceride level, total bile acid level, systolic pressure during pregnancy, fundal height, uric acid level, platelet level, and prepregnancy weight. The LSTM model for time-series research achieved the best performance, and feature ablation provided feature importance for a given time-series input feature. The importance of features was evaluated according to the degree of AUC decrease. The results indicated that the AUC decrease rate for systolic blood pressure was 2%, which was the most important time-series feature, followed by fetal abdominal circumference, head circumference, and maternal weight.

Figure 3. Importance of the variables: (A) identification of important features by the extreme gradient boosting model at weeks 25 to 28; (B) identification of important features by the long short-term memory model. AUC: area under the curve; SBP: systolic blood pressure.



Discussion

Principal Findings

Premature birth is widely recognized as an increasingly serious problem. In this study, 5 pregnancy test records in the first and second trimesters of pregnancy were selected to construct a time-series model to predict preterm delivery. Compared with traditional machine learning models, the use of a time-series model improved prediction performance for preterm birth and allowed the identification of important variables for predicting preterm birth.

The early prediction of preterm birth has always been challenging. The input index of traditional prediction model research has usually been a special test item or a combination of tests that aim to find new markers that have a high contribution to preterm birth prediction; most past studies have not been clinically verified [11,17,46]. Many studies have tried to effectively predict preterm birth, which would allow early detection and prompt management. Cervical screening, fetal fibronectin measurement, or the combination of these methods can effectively predict preterm birth [12-14,16,47]. However, there are still flaws in the forecasts. For asymptomatic women, the performance of the fetal fibronectin test is too low to be clinically relevant [48]. Many studies have found that cervical status is an independent risk factor for preterm birth. In China's 2014 edition of the Clinical Diagnosis and Treatment Guidelines for Preterm Delivery [49], it is recommended that when cervical length is <25 mm, transvaginal ultrasound should be performed before 24 weeks to predict preterm birth in high-risk patients. In fact, cervical examinations are still controversial for screening of the general population. Some studies advocate for dynamic cervical examination regardless of whether a subject is high- or low-risk [50,51]. On the other hand, a greater number of studies either oppose or do not recommend large-scale cervical screening, for reasons that include but are not limited to the material cost, the time required, the lack of unified standards, and the professional training of laboratory personnel [13,14,52-55], which may lead to costs that do not conform to health economics. The prediction model in this study effectively predicts the early development of preterm labor based on demographic factors and prenatal laboratory data. These data are easy to obtain in routine clinical practice. Therefore, the prediction model of preterm birth proposed in this study can be used as a practical screening method for preterm birth in the first and second trimesters of pregnancy.

In fact, earlier works have already reported very close or even higher accuracy than this study. Compared with the large national databases used in previous studies, the conventional data used in this paper is still relatively weak, especially in its lack of key information, such as obstetric and gynecological history and family history. However, we are excited that this paper significantly improves the performance of prediction models through a machine learning method based on time-series technology.

This study reveals various new factors that affect the prediction of preterm birth. Additionally, parameters that have been

traditionally reported to be related to delivery date, such as age, prepregnancy weight, history of preterm birth, and menstrual cycle, were confirmed to be influential factors in preterm birth prediction [1,56]. Interestingly, blood pressure, blood glucose, lipids, uric acid, and other metabolic factors were also very important factors related to preterm birth. Although it has not been thoroughly investigated, the relationship between metabolic risk factors and preterm birth has been preliminarily recognized in several previous studies [57,58]. In a recent observational study of 5535 deliveries, pregnant women with a cluster of metabolic risk factors during early pregnancy were more likely to give birth preterm [59]. The metabolic reaction during pregnancy normally meets the needs of fetal growth; however, an excessive metabolic stress reaction can lead to the occurrence of various pathologies in pregnancy [60]. Despite the controversy, changes in metabolic levels during pregnancy have been observed in women who give birth preterm.

Limitations

This study has several limitations. First, the laboratory examinations of the pregnant women were completed in their respective communities before 20 weeks of gestation, precluding them from being included in the analysis due to differences in test standards. In addition, the prepregnancy characteristics were affected by recall bias; moreover, most of the included women were primipara. Thus, the contribution of preterm birth history to the model was limited. Second, the performance of the model still needs to be improved, although LSTM has great potential. Nonetheless, considering this prediction model is a baseline model based on conventional data, it can continue to add biochemical and biophysical markers to increase screening performance. In addition, advanced maternal age was a clear confounding factor [61], and stratified analysis by age will be considered in a follow-up study. Third, this paper is only a preliminary explanation of the interpretability of the machine learning model. Future work will consider using a more sophisticated post hoc explainability framework, especially for time-series problems. Finally, the study was possibly affected by selection bias due to its single-center design. The prediction model has not been widely used in clinical practice, and its accuracy and practicality should be verified in prospective studies with larger samples.

Conclusions

Preterm birth is the primary cause of neonatal death and disability, and early prediction of preterm birth has great potential to improve the survival rate of preterm infants. In this work, we analyzed obstetric medical data based on time-series machine learning and evaluated the risk of preterm birth. Our study can screen high-risk groups for preterm birth in the early and middle trimesters of pregnancy. Compared with a traditional cross-sectional study, the time-series LSTM model in this study had better overall prediction ability with a lower misdiagnosis rate and the same detection rate. In future work, we will further improve the data set, especially regarding some key characteristics of premature birth that have been reported by past relevant research, and build a more sophisticated post hoc explainability framework for the time series model.

Acknowledgments

This research was funded by the National Health Commission Scientific Research Fund—Major Science and Technology Program of Medicine and Health of Zhejiang Province (WKJ-ZJ-1911), “Pioneer” and “Leading Goose” research and development programs of Zhejiang (2022C03102), Zhejiang Public Welfare Technology Research (GF20F020063), Primary Research and Development Plan of Zhejiang Province in China (2020C03107), Scientific and Technological Research Projects in Key Fields of the Corps (2021AB034-2), Natural Science Foundation of Zhejiang Province in China (GF20F020009, LQ21H040001), National Natural Science Foundation of China (82173530), and the Science and Technology Program of Medicine and Health of Hangzhou (ZD20200035 and OO2019054). We would also like to thank all the pregnant women and health care professionals who participated in the different stages of the development of the prediction model.

Authors' Contributions

YW and YZ were responsible for the study design. WH extracted the data. YW completed the relevant experiments. YW, WH, SL and YZ provided feedback on analyses and interpretation of results. YZ, YW, and ZY wrote the paper. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Vogel JP, Chawanpaiboon S, Moller A, Watananirun K, Bonet M, Lumbiganon P. The global epidemiology of preterm birth. Best Pract Res Clin Obstet Gynaecol 2018 Oct;52:3-12. [doi: 10.1016/j.bpobgyn.2018.04.003] [Medline: 29779863]
2. Luu TM, Rehman Mian MO, Nuyt AM. Long-Term Impact of Preterm Birth: Neurodevelopmental and Physical Health Outcomes. Clin Perinatol 2017 Jun;44(2):305-314. [doi: 10.1016/j.clp.2017.01.003] [Medline: 28477662]
3. Howson C, Kinney M, McDougall L, Lawn JE, Born Too Soon Preterm Birth Action Group. Born too soon: preterm birth matters. Reprod Health 2013;10 Suppl 1:S1 [FREE Full text] [doi: 10.1186/1742-4755-10-S1-S1] [Medline: 24625113]

4. Xue Q, Shen F, Gao Y, Tong M, Zhao M, Chen Q. An analysis of the medical indications for preterm birth in an obstetrics and gynaecology teaching hospital in Shanghai, China. Midwifery 2016 Apr;35:17-21. [doi: 10.1016/j.midw.2016.01.013] [Medline: 27060395]
5. Chen S, Zhang C, Chen Y. Analysis of factors influencing safety of department of obstetrics based on the second child policy and investigation of countermeasures. J Shanghai Jiaotong Univ (Med Sci) 2016;5:742-746. [doi: 10.3969/j.issn.1674-8115.2016.05.025]
6. Jing S, Chen C, Gan Y, Vogel J, Zhang J. Incidence and trend of preterm birth in China, 1990-2016: a systematic review and meta-analysis. BMJ Open 2020 Dec 12;10(12):e039303 [FREE Full text] [doi: 10.1136/bmjopen-2020-039303] [Medline: 33310797]
7. Tokariev A, Stjerna S, Lano A, Metsäranta M, Palva JM, Vanhatalo S. Preterm Birth Changes Networks of Newborn Cortical Activity. Cereb Cortex 2019 Feb 01;29(2):814-826. [doi: 10.1093/cercor/bhy012] [Medline: 30321291]
8. Bensi C, Costacurta M, Belli S, Paradiso D, Docimo R. Relationship between preterm birth and developmental defects of enamel: A systematic review and meta-analysis. Int J Paediatr Dent 2020 Nov 02;30(6):676-686. [doi: 10.1111/ipd.12646] [Medline: 32243004]
9. Blencowe H, Cousens S, Oestergaard M, Chou D, Moller A, Narwal R, et al. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. Lancet 2012 Jun 09;379(9832):2162-2172. [doi: 10.1016/S0140-6736(12)60820-4] [Medline: 22682464]
10. da Fonseca EB, Damião R, Moreira DA. Preterm birth prevention. Best Pract Res Clin Obstet Gynaecol 2020 Nov;69:40-49. [doi: 10.1016/j.bpobgyn.2020.09.003] [Medline: 33039310]
11. Oskovi Kaplan ZA, Ozgu-Erdinc AS. Prediction of Preterm Birth: Maternal Characteristics, Ultrasound Markers, and Biomarkers: An Updated Overview. J Pregnancy 2018;2018:8367571 [FREE Full text] [doi: 10.1155/2018/8367571] [Medline: 30405914]
12. Reicher L, Fouks Y, Yogev Y. Cervical Assessment for Predicting Preterm Birth-Cervical Length and Beyond. J Clin Med 2021 Feb 07;10(4):627 [FREE Full text] [doi: 10.3390/jcm10040627] [Medline: 33562187]
13. Rosenbloom JI, Raghuraman N, Temming LA, Stout MJ, Tuuli MG, Dicke JM, et al. Predictive Value of Midtrimester Universal Cervical Length Screening Based on Parity. J Ultrasound Med 2020 Jan 08;39(1):147-154. [doi: 10.1002/jum.15091] [Medline: 31283038]
14. Rozenberg P. Universal cervical length screening for singleton pregnancies with no history of preterm delivery, or the inverse of the Pareto principle. BJOG 2017 Jun 04;124(7):1038-1045. [doi: 10.1111/1471-0528.14392] [Medline: 27813278]
15. Lu C, Li Z, Wang Z, Guo H, Zhong C, Li Y. Methods of detecting cervical incompetence and their evidence-based evaluation. Journal of Practical Obstetrics and Gynecology 2018;034(005):347-351.
16. Kuhrt K, Hezelgrave-Elliott N, Stock SJ, Tribe R, Seed PT, Shennan AH. Quantitative fetal fibronectin for prediction of preterm birth in asymptomatic twin pregnancy. Acta Obstet Gynecol Scand 2020 Sep 20;99(9):1191-1197 [FREE Full text] [doi: 10.1111/aogs.13861] [Medline: 32249408]
17. Glover AV, Manuck TA. Screening for spontaneous preterm birth and resultant therapies to reduce neonatal morbidity and mortality: A review. Semin Fetal Neonatal Med 2018 Apr;23(2):126-132 [FREE Full text] [doi: 10.1016/j.siny.2017.11.007] [Medline: 29229486]
18. Jhee JH, Lee S, Park Y, Lee SE, Kim YA, Kang S, et al. Prediction model development of late-onset preeclampsia using machine learning-based methods. PLoS One 2019 Aug 23;14(8):e0221202 [FREE Full text] [doi: 10.1371/journal.pone.0221202] [Medline: 31442238]
19. Maragatham G, Devi S. LSTM Model for Prediction of Heart Failure in Big Data. J Med Syst 2019 Mar 19;43(5):111. [doi: 10.1007/s10916-019-1243-3] [Medline: 30888519]
20. Park S, Oh D, Heo H, Lee G, Kim SM, Ansari A, et al. Prediction of preterm birth based on machine learning using bacterial risk score in cervicovaginal fluid. Am J Reprod Immunol 2021 Sep;86(3):e13435. [doi: 10.1111/aji.13435] [Medline: 33905152]
21. Fergus P, Cheung P, Hussain A, Al-Jumeily D, Dobbins C, Iram S. Prediction of preterm deliveries from EHG signals using machine learning. PLoS One 2013 Oct 28;8(10):e77154 [FREE Full text] [doi: 10.1371/journal.pone.0077154] [Medline: 24204760]
22. Tarca AL, Pataki, Romero R, Sirota M, Guan Y, Kutum R, DREAM Preterm Birth Prediction Challenge Consortium, et al. Crowdsourcing assessment of maternal blood multi-omics for predicting gestational age and preterm birth. Cell Rep Med 2021 Jun 15;2(6):100323 [FREE Full text] [doi: 10.1016/j.xcrm.2021.100323] [Medline: 34195686]
23. Koivu A, Sairanen M. Predicting risk of stillbirth and preterm pregnancies with machine learning. Health Inf Sci Syst 2020 Dec 25;8(1):14 [FREE Full text] [doi: 10.1007/s13755-020-00105-9] [Medline: 32226625]
24. Lee K, Ahn KH. Application of Artificial Intelligence in Early Diagnosis of Spontaneous Preterm Labor and Birth. Diagnostics (Basel) 2020 Sep 22;10(9):733 [FREE Full text] [doi: 10.3390/diagnostics10090733] [Medline: 32971981]
25. Weber A, Darmstadt GL, Gruber S, Foeller ME, Carmichael SL, Stevenson DK, et al. Application of machine-learning to predict early spontaneous preterm birth among nulliparous non-Hispanic black and white women. Ann Epidemiol 2018 Nov;28(11):783-789.e1. [doi: 10.1016/j.annepidem.2018.08.008] [Medline: 30236415]

26. Goldstein RF, Abell SK, Ranasinha S, Misso M, Boyle JA, Black MH, et al. Association of Gestational Weight Gain With Maternal and Infant Outcomes: A Systematic Review and Meta-analysis. JAMA 2017 Jun 06;317(21):2207-2225 [FREE Full text] [doi: 10.1001/jama.2017.3635] [Medline: 28586887]
27. Tao J, Yuan Z, Sun L, Yu K, Zhang Z. Fetal birthweight prediction with measured data by a temporal machine learning method. BMC Med Inform Decis Mak 2021 Jan 25;21(1):26 [FREE Full text] [doi: 10.1186/s12911-021-01388-y] [Medline: 33494752]
28. Zhou T, Yu K, Yuan Z, Lu S, Hu W. Predictive Analysis of Postpartum Hemorrhage Based on LSTM and XGBoost Hybrid Model. Computer Systems and Applications 2020;29(3):148-154.
29. Miremberg H, Ben-Ari T, Betzer T, Raphaeli H, Gasnier R, Barda G, et al. The impact of a daily smartphone-based feedback system among women with gestational diabetes on compliance, glycemic control, satisfaction, and pregnancy outcome: a randomized controlled trial. Am J Obstet Gynecol 2018 Apr;218(4):453.e1-453.e7. [doi: 10.1016/j.ajog.2018.01.044] [Medline: 29425836]
30. Khatibi T, Kheyrikoochaksarayee N, Sepehri MM. Analysis of big data for prediction of provider-initiated preterm birth and spontaneous premature deliveries and ranking the predictive features. Arch Gynecol Obstet 2019 Dec 24;300(6):1565-1582. [doi: 10.1007/s00404-019-05325-3] [Medline: 31650230]
31. Zhang Z, Ho KM, Hong Y. Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. Crit Care 2019 Apr 08;23(1):112 [FREE Full text] [doi: 10.1186/s13054-019-2411-z] [Medline: 30961662]
32. Fu Y, Gou W, Hu W, Mao Y, Tian Y, Liang X, et al. Integration of an interpretable machine learning algorithm to identify early life risk factors of childhood obesity among preterm infants: a prospective birth cohort. BMC Med 2020 Jul 10;18(1):184 [FREE Full text] [doi: 10.1186/s12916-020-01642-6] [Medline: 32646442]
33. Ivaturi P, Gadaleta M, Pandey AC, Pazzani M, Steinhubl SR, Quer G. A Comprehensive Explanation Framework for Biomedical Time Series Classification. IEEE J Biomed Health Inform 2021 Jul;25(7):2398-2408. [doi: 10.1109/JBHI.2021.3060997] [Medline: 33617456]
34. Maweu BM, Dakshit S, Shamsuddin R, Prabhakaran B. CEFEs: A CNN Explainable Framework for ECG Signals. Artif Intell Med 2021 May;115:102059. [doi: 10.1016/j.artmed.2021.102059] [Medline: 34001319]
35. Nguyen-Duc T, Mulligan N, Mannu GS, Bettencourt-Silva JH. Deep EHR Spotlight: a Framework and Mechanism to Highlight Events in Electronic Health Records for Explainable Predictions. AMIA Jt Summits Transl Sci Proc 2021;2021:475-484 [FREE Full text] [Medline: 34457163]
36. Viton F, Elbattah M, Guerin JL. Heatmaps for Visual Explainability of CNN-Based Predictions for Multivariate Time Series with Application to Healthcare. 2020 Presented at: 2020 IEEE International Conference on Healthcare Informatics (ICHI); Nov 30-Dec 3, 2020; Oldenburg, Germany. [doi: 10.1109/ichi48887.2020.9374393]
37. Obstetrics group, Branch of Obstetrics and Gynecology, Chinese Medical Association. Pre pregnancy and pregnancy care guidelines (2018). Chinese Journal of Obstetrics and Gynecology 2018;53(1):7-13.
38. Lu Y, Fu X, Chen F, Wong KK. Prediction of fetal weight at varying gestational age in the absence of ultrasound examination using ensemble learning. Artif Intell Med 2020 Jan;102:101748. [doi: 10.1016/j.artmed.2019.101748] [Medline: 31980089]
39. Petrozziello A, Jordanov I, Papageorghiou AT. Deep Learning for Continuous Electronic Fetal Monitoring in Labor. 2018 Presented at: 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Jul 18-21, 2018; Honolulu, HI p. 5866-5869. [doi: 10.1109/embc.2018.8513625]
40. Sun L, Wang Y, He J, Li H, Peng D, Wang Y. A stacked LSTM for atrial fibrillation prediction based on multivariate ECGs. Health Inf Sci Syst 2020 Dec 21;8(1):19 [FREE Full text] [doi: 10.1007/s13755-020-00103-x] [Medline: 32346472]
41. McKinney W. Python for data analysis: data wrangling with Pandas, NumPy, and IPython. Sebastopol, CA: O'Reilly Media; 2017.
42. Komer B, Bergstra J, Eliasmith C. Hyperopt-Sklearn: Automatic Hyperparameter Configuration for Scikit-Learn. In: Proc Of The 13th Python in Science Conf. 2014 Presented at: SciPy2014; Jul 6-12, 2014; Austin, TX p. 32-37 URL: https://conference.scipy.org/proceedings/scipy2014/pdfs/komer.pdf [doi: 10.25080/majora-14bd3278-006]
43. Paszke A, Gross S, Massa F. PyTorch: An Imperative Style, High-Performance Deep Learning Library. 2019 Presented at: 2019 Conference on Neural Information Processing Systems; Dec 8-14, 2019; Vancouver, Canada URL: https://www.researchgate.net/publication/337756689_PyTorch_An_Imperative_Style_High-Performance_Deep_Learning_Library
44. Kingma DP, Ba JL. Adam: A method for stochastic optimization. 2015 Presented at: The 3rd International Conference on Learning Representations; May 7-9, 2015; San Diego, CA URL: https://arxiv.org/pdf/1412.6980.pdf
45. Ismail AA, Gunady M, Bravo HC. Benchmarking Deep Learning Interpretability in Time Series Predictions. ArXiv Preprint posted online on Oct 26, 2020. [doi: 10.48550/arXiv.2010.13924]
46. Suff N, Story L, Shennan A. The prediction of preterm delivery: What is new? Semin Fetal Neonatal Med 2019 Feb;24(1):27-32. [doi: 10.1016/j.siny.2018.09.006] [Medline: 30337215]
47. Son M, Miller ES. Predicting preterm birth: Cervical length and fetal fibronectin. Semin Perinatol 2017 Dec;41(8):445-451 [FREE Full text] [doi: 10.1053/j.semperi.2017.08.002] [Medline: 28935263]

48. Faron G, Balepa L, Parra J, Fils J, Gucciardo L. The fetal fibronectin test: 25 years after its development, what is the evidence regarding its clinical utility? A systematic review and meta-analysis. J Matern Fetal Neonatal Med 2020 Feb;33(3):493-523. [doi: 10.1080/14767058.2018.1491031] [Medline: 29914277]
49. Obstetrics Subgroup, Chinese Society of Obstetrics and Gynecology, Chinese Medical Association. [Diagnosis and therapy guideline of preterm birth (2014)]. Zhonghua Fu Chan Ke Za Zhi 2014 Jul;49(7):481-485. [Medline: 25327726]
50. Einerson BD, Grobman WA, Miller ES. Cost-effectiveness of risk-based screening for cervical length to prevent preterm birth. Am J Obstet Gynecol 2016 Jul;215(1):100.e1-100.e7. [doi: 10.1016/j.ajog.2016.01.192] [Medline: 26880732]
51. Werner EF, Han CS, Pettker CM, Buhimschi CS, Copel JA, Funai EF, et al. Universal cervical-length screening to prevent preterm birth: a cost-effectiveness analysis. Ultrasound Obstet Gynecol 2011 Jul 24;38(1):32-37 [FREE Full text] [doi: 10.1002/uog.8911] [Medline: 21157771]
52. Rozenberg P. [Is universal screening for cervical length among singleton pregnancies with no history of preterm birth justified?]. J Gynecol Obstet Biol Reprod (Paris) 2016 Dec;45(10):1337-1345. [doi: 10.1016/j.jgyn.2016.09.023] [Medline: 28166925]
53. Berghella V. Universal cervical length screening for prediction and prevention of preterm birth. Obstet Gynecol Surv 2012 Oct;67(10):653-658. [doi: 10.1097/OGX.0b013e318270d5b2] [Medline: 23112072]
54. Hermans FJ, Koullali B, van Os MA, van der Ven JE, Kazemier BM, Woiski MD, Triple P group. Repeated cervical length measurements for the verification of short cervical length. Int J Gynaecol Obstet 2017 Dec 28;139(3):318-323. [doi: 10.1002/ijgo.12321] [Medline: 28884811]
55. Masters HR, Warshak C, Sinclair S, Rountree S, DeFranco E. Time required to complete transvaginal cervical length in women receiving universal cervical length screening for preterm birth prevention. J Matern Fetal Neonatal Med 2020 Aug 30:1-5. [doi: 10.1080/14767058.2020.1811666] [Medline: 32862742]
56. Torchin H, Ancel PY. [Epidemiology and risk factors of preterm birth]. J Gynecol Obstet Biol Reprod (Paris) 2016 Dec;45(10):1213-1230. [doi: 10.1016/j.jgyn.2016.09.013] [Medline: 27789055]
57. Grieger JA, Bianco-Miotto T, Grzeskowiak LE, Leemaqz SY, Poston L, McCowan LM, et al. Metabolic syndrome in pregnancy and risk for adverse pregnancy outcomes: A prospective cohort of nulliparous women. PLoS Med 2018 Dec 4;15(12):e1002710 [FREE Full text] [doi: 10.1371/journal.pmed.1002710] [Medline: 30513077]
58. Le TM, Nguyen LH, Phan NL, Le DD, Nguyen HV, Truong VQ, et al. Maternal serum uric acid concentration and pregnancy outcomes in women with pre-eclampsia/eclampsia. Int J Gynaecol Obstet 2019 Jan 08;144(1):21-26 [FREE Full text] [doi: 10.1002/ijgo.12697] [Medline: 30353543]
59. Lei Q, Niu J, Lv L, Duan D, Wen J, Lin X, et al. Clustering of metabolic risk factors and adverse pregnancy outcomes: a prospective cohort study. Diabetes Metab Res Rev 2016 Nov 10;32(8):835-842. [doi: 10.1002/dmrr.2803] [Medline: 27037671]
60. Mouzon SH, Lassance L. Endocrine and metabolic adaptations to pregnancy; impact of obesity. Horm Mol Biol Clin Investig 2015 Oct;24(1):65-72. [doi: 10.1515/hmbci-2015-0042] [Medline: 26613331]
61. Frick AP. Advanced maternal age and adverse pregnancy outcomes. Best Pract Res Clin Obstet Gynaecol 2021 Jan;70:92-100. [doi: 10.1016/j.bpobgyn.2020.07.005] [Medline: 32741623]

Abbreviations

AUC: area under the curve
CDC: US Centers for Disease Control and Prevention
CNN: convolutional neural network
EMR: electronic medical records
FN: false negative
FP: false positive
LSTM: long short-term memory
RNN: recurrent neural network
ROC: receiver operating characteristic
TN: true negative
TP: true positive
XGB: extreme gradient boosting

Edited by C Lovis; submitted 26.09.21; peer-reviewed by M Elbattah, I Mircheva; comments to author 31.01.22; revised version received 21.04.22; accepted 25.04.22; published 13.06.22.

Please cite as:
Zhang Y, Lu S, Wu Y, Hu W, Yuan Z
The Prediction of Preterm Birth Using Time-Series Technology-Based Machine Learning: Retrospective Cohort Study
JMIR Med Inform 2022;10(6):e33835
URL: https://medinform.jmir.org/2022/6/e33835
doi:10.2196/33835
PMID:35700004

Original Paper

Machine Learning Support for Decision-Making in Kidney Transplantation: Step-by-step Development of a Technological Solution

François-Xavier Paquette[1], BSc; Amir Ghassemi[1], MSc; Olga Bukhtiyarova[1], MD, PhD; Moustapha Cisse[1], MSc; Natanael Gagnon[1], DEC; Alexia Della Vecchia[1,2], BSc; Hobivola A Rabearivelo[1], MSc; Youssef Loudiyi[1], MSc

[1]BI Expertise, Quebec, QC, Canada

[2]Research Institute McGill University Heath Centre, Montreal, QC, Canada

Corresponding Author:
Youssef Loudiyi, MSc
BI Expertise
Complexe Samuel Holland, bureau 315
830, avenue Ernest Gagnon
Quebec, QC, G1S3R3
Canada
Phone: 1 418 473 9729
Email: youssef.loudiyi@biexpertise.com

Abstract

Background: Kidney transplantation is the preferred treatment option for patients with end-stage renal disease. To maximize patient and graft survival, the allocation of donor organs to potential recipients requires careful consideration.

Objective: This study aimed to develop an innovative technological solution to enable better prediction of kidney transplant survival for each potential donor-recipient pair.

Methods: We used deidentified data on past organ donors, recipients, and transplant outcomes in the United States from the Scientific Registry of Transplant Recipients. To predict transplant outcomes for potential donor-recipient pairs, we used several survival analysis models, including regression analysis (Cox proportional hazards), random survival forests, and several artificial neural networks (DeepSurv, DeepHit, and recurrent neural network [RNN]). We evaluated the performance of each model in terms of its ability to predict the probability of graft survival after kidney transplantation from deceased donors. Three metrics were used: the C-index, integrated Brier score, and integrated calibration index, along with calibration plots.

Results: On the basis of the C-index metrics, the neural network–based models (DeepSurv, DeepHit, and RNN) had better discriminative ability than the Cox model and random survival forest model (0.650, 0.661, and 0.659 vs 0.646 and 0.644, respectively). The proposed RNN model offered a compromise between the good discriminative ability and calibration and was implemented in a technological solution of technology readiness level 4.

Conclusions: Our technological solution based on the RNN model can effectively predict kidney transplant survival and provide support for medical professionals and candidate recipients in determining the most optimal donor-recipient pair.

(JMIR Med Inform 2022;10(6):e34554) doi:10.2196/34554

KEYWORDS

machine learning; artificial intelligence; medical decision support; kidney transplantation

Introduction

Current State of Organ Allocation

Deceased organ donation is the most common type of kidney donation [1] and can be defined as donation after neurological death (neurological determination of death [NDD]) and donation after circulatory death (DCD) [2]. Despite being authorized in Canada since 2006, DCD donations represented only 17% of deceased organ donations in Canada in 2012 [3]. The number of patients waiting for organ transplantation greatly exceeds the number of organs donated [4]. Ensuring an optimal donor identification and referral process and improving efficiency in identifying compatible donors would help avoid missed donation opportunities [3] and increase the rate of DCD [4]. Assisting

informed decision-making regarding the acceptance of donor kidney by helping patients to better understand the treatment options and potential transplant outcomes would promote better treatment efficiency [5].

In current clinical practice, several kidney allocation algorithms are used to match donor organs with potential recipients. In the United States, the Organ Procurement and Transplantation Network uses a list of potential recipients that are ranked according to objective medical criteria (eg, blood type, tissue type, and size of the organ as well as medical urgency, time spent on the waiting list, and distance between the donor and recipient) [6]. Several simple numerical tools have also been implemented to guide kidney allocation. An example is the Estimated Post Transplant Survival score [7]. This score is assigned to all adult candidates on the kidney transplant waiting list and is based on 4 factors: candidate's time on dialysis, current diagnosis of diabetes, prior solid organ transplants, and candidate's age. The kidney donor risk index [8] combines various donor factors to summarize the risk of graft failure after kidney transplantation into a single number. It uses features such as donor's age, height, weight, ethnicity (or race), history of hypertension, history of diabetes, cause of death, serum creatinine level, hepatitis C status, and DCD criteria. The kidney donor risk index is then remapped to a percentile scale where the lower percentiles (0%-20%) represent a lower risk of graft failure. Candidates with Estimated Post Transplant Survival scores ≤20% will receive offers for kidneys from donors with Kidney Donor Profile Index scores ≤20% before other candidates at the local, regional, and national levels of distribution [9]. Similar candidate and donor variables have also been considered in Canadian kidney allocation systems [10]. According to the recommendations of the Canadian Council for Donation and Transplantation, priority should be given to young recipients (especially when the organ donor is also young), donor-recipient pairs with zero mismatch for HLA ABDR, highly sensitized patients, and those requiring combined transplants.

Machine Learning Support for Organ Donation

When deciding the suitability of a kidney graft for a recipient, it is important to estimate how long the donated organ will remain functional. To address this question, numerous studies have used machine learning (ML) models to predict kidney transplant outcomes, each differing in variable and outcome definitions.

Some models were built using data from either living donor [11] or deceased donor transplants only [12,13], whereas others considered both donor types [14].

In 2010, Reinaldo et al [15] evaluated several simple and interpretable ML models, in which the decision tree model showed 94% accuracy in predicting graft survival 1 year after transplant.

A recent study by Luck et al [16] proposed a neural network model built on data from the Scientific Registry of Transplant Recipients (SRTR) database, where the outcome of interest was graft failure. A total of 436 different variables were used to build the neural network model. The survival predictions were evaluated using a C-index (the percentage of transplant pairs correctly ordered by the model according to the observed survival durations), which was slightly higher than that obtained using the Cox model (0.655 compared with 0.65).

These studies built and evaluated various ML models; however, their termination at the stage of proof of concept makes it difficult to use the results for assistance in clinical decision-making.

Several tools have reached advanced technological readiness levels. Patzer et al [14] built a mobile app to predict 1- and 3-year patient survival using multivariate logistic regression analysis. Kilambi et al [17] quantified the benefits of accepting a kidney transplant based in part on the expected patient survival using Cox regression models. Loupy et al [18] designed a tool to predict long-term kidney allograft failure to guide posttransplant care, also using a Cox model. To the best of our knowledge, all published results are based on linear models that may not capture the nonlinear relationships between the input variables.

The *objective* of this project is to develop an innovative solution of technology readiness level 4 (TRL-4; component and validation in a laboratory environment) that would use ML to support medical decisions about accepting kidney transplants for particular donor-recipient pairs, with specific attention to DCD donations.

This study describes all stages of development of the ML technological solution: data acquisition and preparation, training and evaluation of ML models, and deployment of the solution.

Methods

Data Access and Data Security

BI Expertise obtained permission from SRTR (United States) to access its extensive historical data on organ transplants that were previously used in research [1,19].

Special measures were taken to maintain both the confidentiality and security of personal data. The BI Expertise team leveraged Microsoft Azure public cloud to ensure that all the data were secured and only the team could access it remotely. Data exfiltration risk was avoided by disabling all direct remote accesses. The environment was only visible to end users using a virtual machine inside Azure. This virtual machine was entirely isolated from the computers that were accessing it (no cut and paste).

The predictive modeling environment was based on the Azure ML data science platform and all the data resided in Azure Synapse Analytics. Both platforms were fully integrated to optimize the data preparation process and feature engineering activities. Once the predictive model was built and validated, it was deployed to a specific virtual machine that also hosted the user interface, which was accessible through a browser using a computer, tablet, or mobile device.

Ethics Approval

The proposed architecture was approved by the SRTR research ethics board (REB 2020-020H), and upon deployment, BI Expertise agreed to submit it to unannounced audits.

Data Set

This study was based on several data tables from SRTR, namely, *DONOR_DECEASED*, *REC_HISTO*, *CAND_KIPA*, *TXF_KI*, and *TX_KI*. The tables contained individual deidentified sociodemographic and medical characteristics of kidney donors and recipients as well as outcomes of kidney transplantation *such as graft failure, recipient death, or loss to* follow-up.

We included first-time kidney recipients who underwent transplantation between January 1, 2000, and December 31, 2019. This choice of subset was motivated by important progress made in the field of kidney transplantation at the beginning of the year 2000, and the chosen data set included transplants after these changes were made. In addition, by selecting recipients from the same transplant era, we ensured that all recipients would have undergone similar methods of matching donor-recipient pairs [20].

Data Cleaning and Selection of Variables

The selection of variables to be used for survival analysis was based on expert knowledge, data completeness, and previously published studies [12,21,22]. The input variables included sociodemographic characteristics of donors and recipients, history of comorbidities, blood type, details on donors' death and levels of creatinine, time on the waiting list for recipients, and number of HLA mismatches. These data are typically known before the decision-making about the transplant and therefore can be reliably used as input for the ML mode. The exclusion criteria were the following: (1) variables not known before the transplantation (ie, immunosuppression therapy), (2) variables specific for the US medical system (ie, payment source for transplant recipients), and (3) variables with >20% of missing observations. Multimedia Appendix 1 provides a complete list of the variables and their definitions.

Outcome Definition

The primary outcome was death-censored kidney graft survival, defined as the time elapsed between transplantation and diagnosis of graft failure. Data were censored at the time of the most recent follow-up for recipients who still had functioning grafts, at the time of their last record for those who were lost to follow-up, and at the time of death for those who died before experiencing graft failure. Probability of graft survival was predicted at set time points ranging from 0 to 15 years after transplantation, with intervals of 3 months between each time point.

Feature Engineering

Some variables contained duplicate information, such as racial and ethnic groups. In this case, they were regrouped into a single variable. This resulted in the creation of new variables, which are described in detail in Multimedia Appendix 1.

LassoCV, ElasticNetCV, and recursive feature elimination feature selection methods from the scikit-learn package were used to select the most important variables.

Survival Analysis Models

Several linear and nonlinear survival models were considered.

Cox Proportional Hazards

The Cox proportional hazards model [23] evaluates the effects of covariates on survival time and is commonly used in multivariate survival analysis because of its ease of implementation and interpretation. The Python package *scikit-survival* was used in this study to perform computations related to the Cox model.

DeepSurv

DeepSurv is a variant of the Cox model [24] that handles nonlinear data. The hazard ratio is produced by a neural network, which enables the model to learn from the interactions between covariates. The Python package *pycox* was used to perform training and testing of the DeepSurv model.

DeepHit

DeepHit [25] is an artificial neural network whose output vector is the joint probability distribution of all possible events (graft failure in this study) at each time point, which enables the model to learn the time-varying effects of each covariate on graft survival. The Python package *pycox* was used to perform training and testing of the DeepHit model.

Random Survival Forest

Random survival forest (RSF) [26] is an extension of the random forest model [27] that takes into account right-censoring of survival data. An RSF is an ensemble of survival trees, and each tree is grown on a subsample of the training data. The Python package *scikit-survival* was used to build and test the RSF model.

Recurrent Neural Network

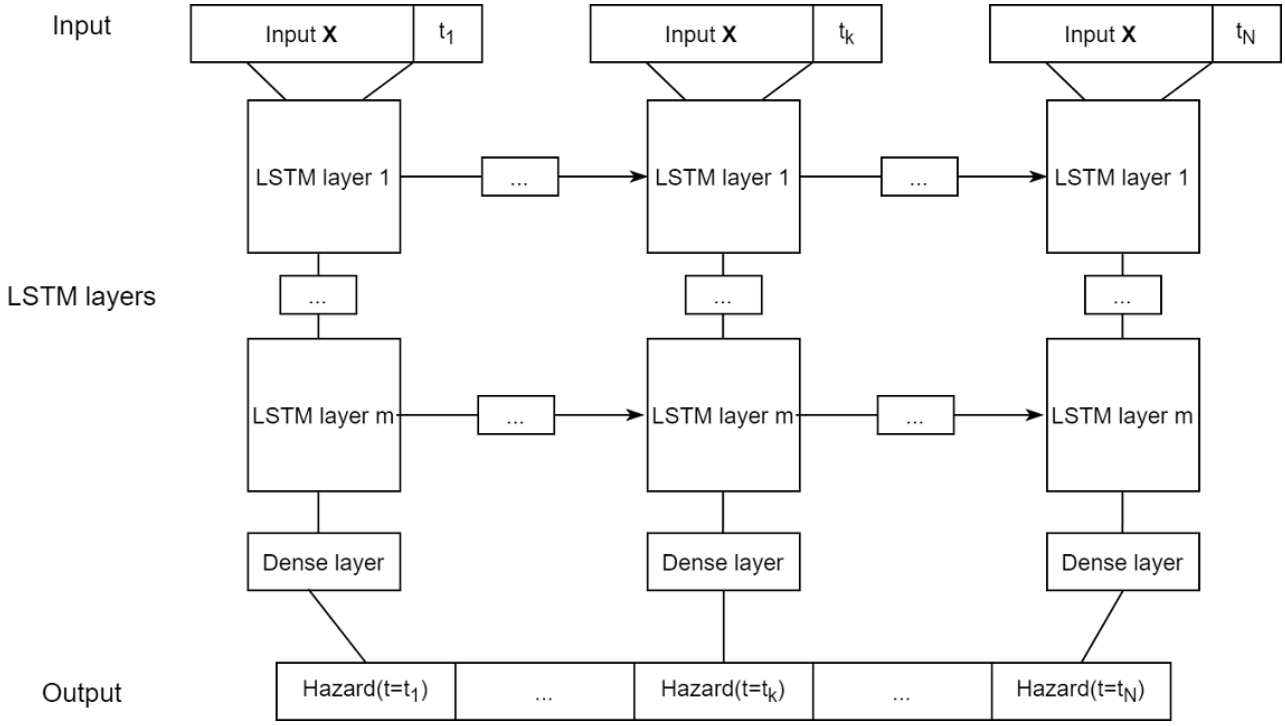
Overview

The structure of our recurrent neural network (RNN) was inspired by previous studies that described deep recurrent survival analysis [28] and RNN-SURV [29]. The RNN presented in this study was implemented in Python using *TensorFlow 2.2* (Google Inc).

Structure of the RNN Model

For each of the N time intervals, the covariate vector X is passed, along with the time interval value *t*, through a series of m long short-term memory layers (Figure 1). The time interval value is added to explicitly capture the time-varying effects of the covariates. The N outputs are then passed through a dense layer with sigmoid activation to obtain the hazard rate at each time step. The hazard rates can be used to compute the estimated probability of survival at any time step *t* as follows:

Figure 1. Structure of the recurrent neural network model. LSTM: long short-term memory.



Loss Function

We compared 2 variants of loss functions, namely, the negative log-likelihood of the cumulative distribution function on all samples added to the negative log-likelihood of the probability density distribution on uncensored samples [28] and the ranking loss proposed in DeepHit [25].

Postprocessing of the RNN Output

To increase the calibration (refer to the *Model PerformanceEvaluation Metrics* section), a method to use the outputs of the RNN (individual hazard rates) as relative risk factors was devised, similar to the individual risk scores obtained from a Cox model. The main difference is that the risk factors vary over time. Therefore, for each patient, we interpreted the hazard rates at each time step as a risk score. From these risk scores, we aimed to obtain calibrated hazard rates to produce better calibrated survival predictions.

One approach to predict the hazard rates from the Cox model risk scores is as follows:

Where the baseline hazard can be estimated from the training data with:

Where *d(t)* is the number of events at *t* and *R(t)* is the risk set at *t*, composed of all individuals still susceptible to the event of interest at time *t* [30].

A similar method was implemented for our RNN model, with the modification that the risk scores at each time step are associated with one of n risk bins, with each risk bin having its own baseline hazard. The cutoff points for the risk bins are determined by computing the n-quantiles of the estimated risk scores of the training samples at each time step.

Calibrated_Hazardi,t = *Ri,t* * *BHk,t*: estimated calibrated hazard rate for transplant i at time step t

$R_{i,t}$: risk score for individual *i* at time step *t*

$BH_{k,t}$: baseline hazard for risk bin *k* at time step *t*

Where the baseline hazards are estimated from the training data with:

which represents the number of observed events at time *t* for samples of bin *k*, divided by the sum of risk scores at time *t* for samples of bin *k* that are still susceptible to the event of interest at time *t*.

The individual calibrated hazard rates can then be used to compute survival probabilities.

Training and Evaluation Data Sets

The results presented in this study were obtained using 5-fold cross-validation. It consists of randomly splitting the data set into 5 partitions of equal size and repeating the training and evaluation process 5 times, each time using one partition (20%) as the evaluation set and the remaining (80%) as the training set.

Training and evaluating for hyperparameter tuning, choice of loss function, and choice of training approach were performed using 5-fold cross-validation (each with different permutations of the 5-fold partitions). These steps were performed on the same set used to compare ML models.

Model Performance Evaluation Metrics

Concordance Index

The concordance index [31] is a measure of the discrimination power of a model. It measures the concordance between the ranking of the predicted risk metrics (eg, risk score, failure time, or probability of failure) and the observed failure times for all pairs of transplants. A pair of samples *i*,*j* is concordant if the predicted risk score of *i* is greater than that of *j* and sample *i* has a shorter survival period than *j*. The C-index is the number of concordant pairs of transplants divided by the total number of comparable pairs. The result can take any value between 0 and 1, with 0.5 representing no discrimination (random predictions) and 1 representing a perfect model.

Harell C-index =  where  is the risk score for transplant *i*.

As the C-index uses a single time-independent risk metric to rank the transplants, it fails to account for the time-dependent effects of covariates on the risk of a patient. In the case of proportional hazard models such as Cox, this has no incidence (ie, risk scores do not change over time). However, for models that output individual survival distributions, the estimated risk of patients may vary with time. For example, a patient with a higher failure probability than others at an earlier time point might have a lower failure probability than others later on. Therefore, the time-dependent concordance index was used to evaluate the models [32]. For this index, a pair of transplants *i*,*j* is considered concordant if *i* experienced failure at a time t_i sooner than t_j and the probability of *i* surviving beyond t_i is lower than that of *j* surviving beyond t_i.

Antolini time-dependent C-index = 

Integrated Brier Score

The Brier score [33] for a time point *t* is the average squared distance between the predicted probability of surviving beyond time *t* and the observed status at *t*. In the presence of right censored data, the distances must be weighed using an inverse probability of the censoring weight method [34].

Brier score *(t)* = 

Where *G(t)* = P[censoring time > *t*] (estimated with the Kaplan-Meier estimator on censoring data).

The integrated Brier score (IBS) is simply the average Brier score across all prediction time points.

IBS = 

Calibration

Calibration of a model refers to the goodness-of-fit of its survival predictions [35]. For example, a model predicts that a patient has a 70% probability of surviving to time *t**. Evaluating the model's calibration aims to answer the question whether the patient can trust this prediction. If 100 patients with identical characteristics as this one were under observation, it would be possible to look at their actual survival times and verify if approximately 70 of them survived to *t**. If there was a significant difference between the predicted and observed survival rates, it would mean that the model was not well calibrated [35].

In reality, the data sets are composed of patients with different characteristics. One common method for evaluating a model's calibration at a chosen time point *t** is to stratify all the patients into groups based on the predicted probability of failure by time *t**. For example, one method is to stratify the patients into 10 groups, where the cutoff points are the deciles of the distribution of the predicted probabilities. For each group, the observed failure rate by time *t** is computed using a Kaplan-Meier estimator fitted to the patients of the group. This observed failure rate is then compared with the average predicted probability of failure by time *t** for all patients in the group. The resulting pairs of predicted and observed values can be visually examined side-by-side or on a plot. This process can be repeated for all time points [36].

However, Harrell [37] argued that the binning of the predicted probabilities leads to a loss of precision. To address this issue, Austin et al [36] proposed using regression splines to model the observed failure rate as a function of the complementary log-log transformation of the predicted failure rate, using the relationship:  . For a visual evaluation of the calibration at a time *t**, an estimate of the observed failure probability before *t** for every predicted failure probability $F_i(t^*)$ can be obtained using the regression splines, and the resulting pairs can be plotted. With a perfectly calibrated model, this would yield a diagonal curve.

One of the suggested metrics for numerically assessing the calibration is the integrated calibration index (ICI) [38], which is simply the mean absolute difference between the predicted and estimated observed values.

ICI (*t**) = 

Development of the Technological Solution

The developed end-user application provides the relevant graft survival probabilities in 3 steps. First, users must enter the required information related to the donor and the transplant candidate (Multimedia Appendix 1). Second, the predictive model is run to obtain survival probabilities under 3 simulated scenarios: the recipient receives the deceased donor kidney (as per the input of step 1), the recipient receives a kidney from a predefined average DCD donor, and the recipient receives a kidney from a predefined average NDD donor. Third, the graft survival predictions are shown (Multimedia Appendix 1). The average DCD and NDD results at the current time point are included to enable the comparison between multiple donor-recipient matches and to support medical decision-making about accepting the proposed donor kidney or waiting for the next available one.

Software Used for the Project

JIRA (project management; Atlassian), Bitbucket (code management; Atlassian), Confluence (documentation management; Atlassian), Azure (Microsoft) Cloud Platform (cloud), Azure Machine Learning (computations), Google Suite,

Teams (team communication), Azure Secured Virtual Machine (data security), VS Code (Microsoft), Python (ML model design and coding), and Expo.io (framework for client web applications, expo.dev) were the software used for the project.

Code and Model Availability

The code and the trained model can be available upon request if permission from Health Canada and SRTR is obtained in each particular case, which is needed for ethical considerations.

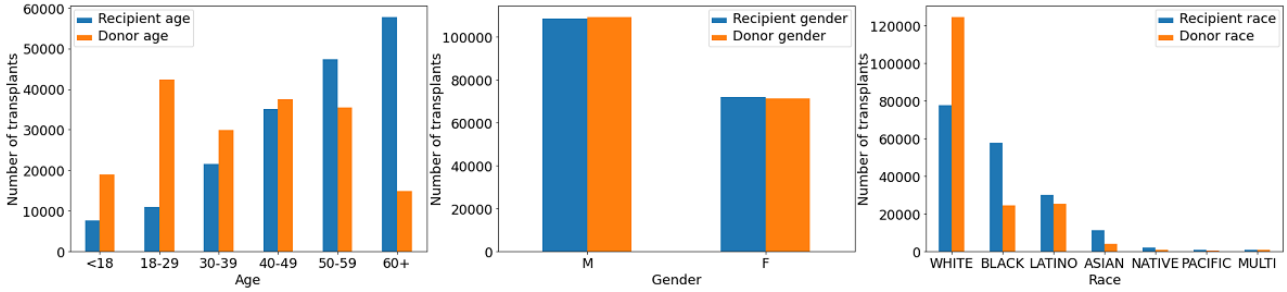
Results

Characteristics of the Data Sets

The initial data sets contained information on 210,688 first-time kidney transplant recipients from deceased donors and included 402 variables. The final data set obtained after data cleaning and selection of variables contained data on 180,141 transplants (154,292 from NDD donations and 25,849 from DCD donations) and included 35 variables. Feature selection methods such as LassoCV, ElasticNetCV, and recursive feature elimination did not recommend changing the set of variables chosen based on manually set exclusion criteria. After one-hot encoding of the categorical variables, the total number of input covariates was 170 (Multimedia Appendix 1).

Demographics of the patients are shown in Figure 2. This study considered donor-recipient pairs of all ages, including pediatric patients (aged <18 years). The data set contained an unequal number of donors and recipients belonging to different sociodemographic groups. The number of kidney transplant recipients increased with age, which may reflect the fact that the older population is more likely to have end-stage kidney disease. In contrast, the fewest number of eligible donors per age group was the ≥60 cohort. This may also be attributed to the fact that not all kidneys retrieved from the older adult donors are viable. Older adult donors are likely to have more comorbidities, making them illegible to donate. The study population included a large number of male recipients and donors. It was also imbalanced regarding racial groups, with a predominant number of White donors over donors of other races, as well as an unequal number of recipients of different races.

Figure 2. Sociodemographic characteristics of kidney donors and recipients.



Choice of Hyperparameters and Training

The 3 neural network–based models were trained using the Adam optimizer with a learning rate of 0.001 and batch size of 128. The optimal number of hidden layers and the number of nodes in the layers were determined separately for each model by testing a range of possible values, starting with small networks and gradually increasing their size. In the 3 cases, increasing the number of hidden layers in the past 3 models resulted in overfitting and decreased discriminative performance. Batch normalization and dropout with a rate of 0.10 were used. In addition, L2 regularization with a factor of 0.001 was used for the RNN model.

DeepSurv consists of 2 dense layers, with 32 and 16 neurons in layers 1 and 2, respectively. DeepHit consists of 3 dense layers with 64, 32, and 16 neurons, respectively. The long short-term memory layers of RNN contain the same number of neurons.

The RSF consists of 100 trees, with a maximum depth of 25 nodes. At each node, 13 randomly selected covariates were considered to split (the square root of the number of covariates). The minimum number of samples required to split a node was 400, and the minimum number of samples in the leaf nodes was 200. Adding more trees did not increase the discriminative ability of the model, and reducing the minimum number of samples to split resulted in overfitting.

Comparison of RNN Loss Functions

Different loss functions (or objective functions) were tested when building the RNN model. It was found that using the ranking loss proposed in DeepHit [25] yielded a model with better discrimination ability. With the deep recurrent survival analysis [28] loss function, the average C-index was 0.64 on the graft survival task, whereas with the DeepHit ranking loss, the C-index averaged approximately 0.66. Therefore, the latter loss function was used to train the proposed RNN model.

Definition of the loss function:

where

α=.1 (a calibration parameter) and

c_i=0 indicates that patient *i* experienced the event of interest during observation period.

Using this loss function to train the neural network yields a model with good discrimination ability but produces poorly calibrated survival predictions. This is because the loss function was mainly designed to encourage the correct ordering of pairs. This issue motivated the postprocessing of the RNN outputs, which is presented in the *Survival Analysis Models* section.

Model Performance

In preliminary experiments, 3 approaches were tested to obtain survival predictions for DCD kidney transplants with survival analysis models. DeepHit was used as a benchmark for this purpose. The first method was to train the model using only the DCD transplant data, which yielded an average C-index of 0.604. The second method was to train the model using data from both NDD and DCD transplants, which yielded an average C-index of 0.631 on the DCD evaluation set. The third method was to use transfer learning, which consisted of training the model on the larger NDD transplant data set (to gain general knowledge on kidney transplants), then training the model a second time on the DCD transplant data set to gain knowledge specific to DCD grafts. This approach yielded an average C-index of 0.625 on the DCD-only evaluation set. Thus, the model trained only on DCD transplants yielded the poorest results, which may be explained by the lower volume of data available for this specific transplant cohort. The best performance was obtained with the model trained on a data set that included both NDD and DCD transplants. Therefore, further development of ML models was based on the combined data set.

For the final evaluation of the models, a 5-fold cross-validation was used. It consists of randomly splitting the data set into 5 partitions of equal size and repeating the training and evaluation process 5 times, each time using one partition (20%) as the evaluation set and the remaining (80%) as the training set. Table 1 presents the evaluation results for the 5 models that were explored. The C-index obtained by using the Cox proportional hazards model was 0.646. The decision tree–based RSF had a time-dependent C-index of 0.644, whereas the neural network–based models (DeepSurv, DeepHit, and our proposed RNN) obtained time-dependent C-indexes of 0.650, 0.661, and 0.659, respectively. Table 1 also presents IBS and ICI for the 1-year, 5-year, and 15-year time points. The ICI for each time point was the lowest for the Cox proportional hazards model, whereas the C-index and IBS showed the best values for DeepHit and RNN, respectively.

Figure 3 shows the smoothed calibration curve for the cumulative probability of graft failure at 1 year, 5 years, and 15 years. These plots help to visualize the discrepancy between the graft failure probability predicted by the model and the observed graft failure rate.

For the probability of graft failure in the first year, all 5 tested models had similar calibration, as shown by the ICIs in Table 1 and the calibration curves shown in Figure 3. They all tended to slightly underestimate the survival rate. There were more significant differences in the calibration of the models for the probabilities of graft failure in the first 5 and 15 years. The calibration curves for Cox and DeepSurv are almost perfectly aligned with the identity line and have very low ICIs, indicating that the 2 models produce the most reliable individual survival predictions.

In the case of DeepHit, it is interesting to see that although it had the best discriminative ability, the model failed to produce sufficiently accurate survival predictions, especially at later time points. For the 5- and 15-year time points, DeepHit had the worst ICI (0.0285 and 0.1356) of all models, and its calibration curve had the most significant deviation from the identity line.

The survival predictions produced by the RNN were better calibrated than those produced by DeepHit and RSF. However, as seen on the calibration plots, they are not as well calibrated as those obtained using the Cox and DeepSurv models.

Table 1. Evaluation results for the tested machine learning models.

Model	C-index	IBS[a]	ICI[b] for 1 year	ICI for 5 years	ICI for 15 years
Cox proportional hazards	0.646	0.15439	*0.00942*[c]	*0.00949*	*0.00748*
DeepSurv	0.650	0.15361	0.00957	0.00999	0.01189
DeepHit	*0.661*	0.15259	0.01171	0.02858	0.13561
RSF[d]	0.644	0.15288	0.01058	0.01739	0.04559
RNN[e]	0.659	*0.15220*	0.00989	0.01076	0.02634

[a]IBS: integrated Brier score.

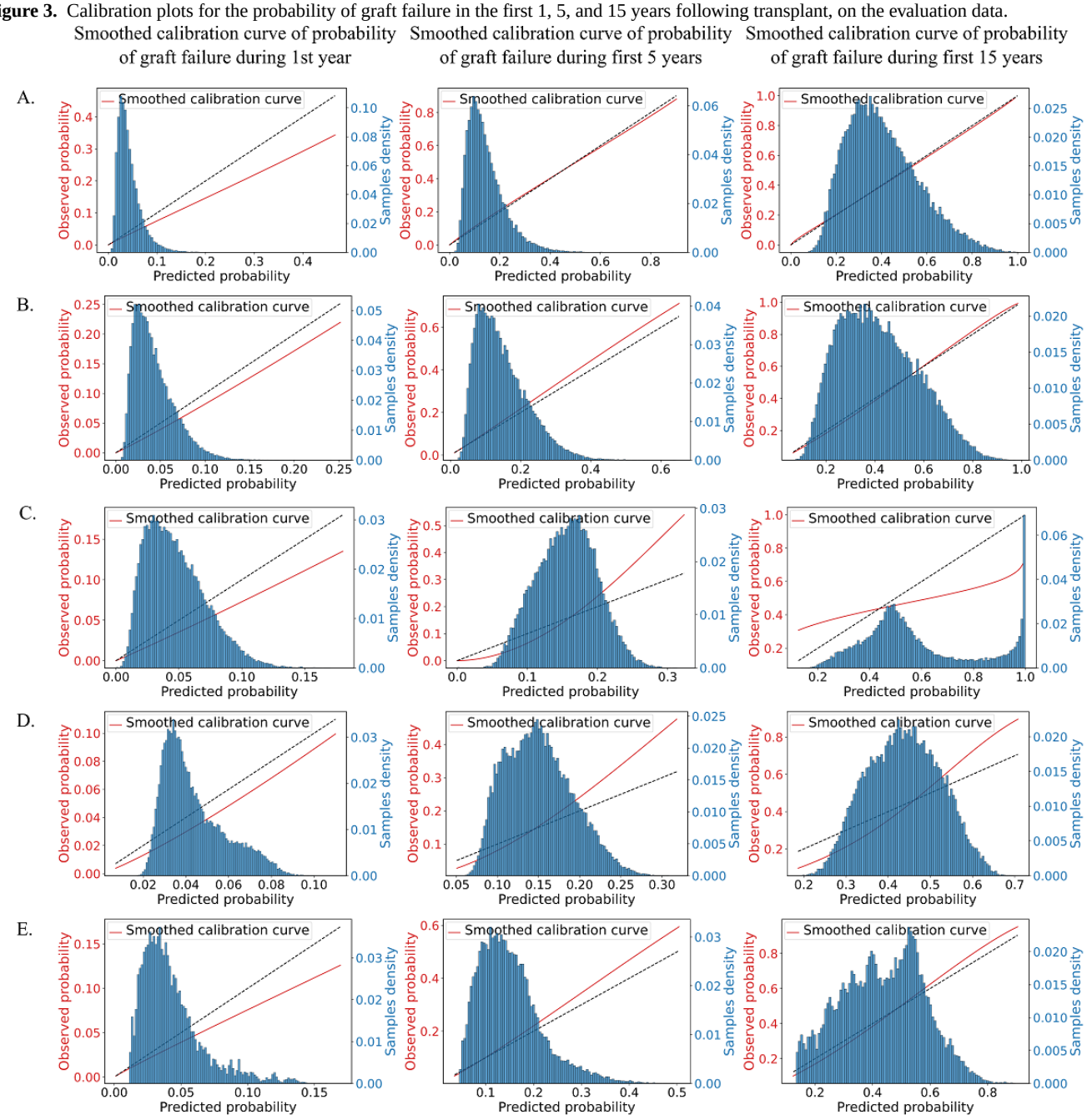
[b]ICI: integrated calibration index.

[c]The italicized values represent the best result obtained for each evaluation metric.

[d]RSF: random survival forest.

[e]RNN: recurrent neural network.

Figure 3. Calibration plots for the probability of graft failure in the first 1, 5, and 15 years following transplant, on the evaluation data.



Discussion

Overview

This study focused on the development of an ML-based decision support solution to help kidney transplant practitioners and their patients make informed decisions when a deceased donor kidney becomes available. All stages of the development process are described: data acquisition and preparation, evaluation of existing survival analysis models, development and evaluation of a new survival analysis model, and deployment of the technological solution of TRL-4.

Principal Findings

When building survival analysis models in the context of kidney transplantation, there are several factors that characterize the models and ultimately influence the final quality of the prediction tool.

One factor is the size of the data sets used to build these models. It varies widely between studies, ranging from 80 [39] to 131,709 transplants [16]. It has been demonstrated that large sample sizes improve the predictive performance of ML models [40]. Another important factor is the period for which the risk of mortality or graft failure is predicted. This may depend on data availability and duration of the observation period. Mark et al [22] built an ensemble model to predict patient survival throughout the first 5 years following kidney transplantation. Luck et al [16] evaluated the graft survival probability at each anniversary date of the graft for 15 years following transplantation. Our study was based on the most recent available data and included up to 19 years of observations of 180,141 transplant procedures. The models presented here

evaluate graft survival probabilities at each quarterly anniversary of the graft for 15 years. To the best of our knowledge, this is the largest data set with the longest observation period used to build ML models for predictions in the kidney transplantation area.

The performance of a predictive model is also strongly dependent on incorporating prognostically significant variables into the models. The number of variables used for survival analysis in the literature ranges from 6 to several hundred [16,21,41,42]. Selection of a very small number of variables may lead to the exclusion of important factors that may influence the outcome of the transplantation, whereas including a very large number of variables may increase the sparsity of the data, which in turn may cause overfitting. In this study, variables were selected based on medical expertise, previous studies [18,22], and characteristics such as data completeness and data duplication for the first step (35 variables).

The choice of a survival analysis model is also critical. Multiple options have been described in the literature, such as the Cox regression model [18], decision trees [43], support vector machines [44], Bayesian belief networks [12], RSF [22], and artificial neural networks [16,21].

In this study, 5 different models were explored: a regression-based Cox proportional hazards model; RSF; and 3 neural network models, namely, DeepSurv, DeepHit, and a proposed RNN. To the best of our knowledge, the latter was used on kidney transplantation data for the first time in this study. These models were evaluated on the task of predicting kidney graft survival throughout the first 15 years following transplantation. Three metrics were used to evaluate each model: the C-index, IBS, and ICI, along with calibration plots.

Evaluation of ML Models

The results for the C-index metric shown in Table 1 indicate that the neural network–based models (DeepSurv, DeepHit, and RNN) had better discriminative ability than the Cox model and RSF. In fact, the DeepHit model and our proposed RNN model performed best with a C-index of 0.661 and 0.659, respectively. This indicates their ability to discern groups of donor-recipient pairs that were at a higher risk of experiencing graft failure after transplant from groups that had a lower risk. The improvement compared with the widely used Cox model (C-index of 0.646) may be because of the higher capacity for feature extraction by the neural networks.

The main drawback of the Cox proportional hazards model and DeepSurv is the assumption that the computed hazard ratio is time invariant. In contrast, DeepHit and RNN make no assumptions about the distribution of time-to-event data and can learn the time-varying effects of covariates, making them more flexible. This is important when evaluating survival over a wide time frame, as in our study, over 15 years. For example, a covariate could have a negative effect on survival in the first few years after transplantation but no impact in the later years.

Previously published articles on the prediction of survival of kidney grafts from deceased donors often described different evaluation metrics, such as accuracy [15,44], mean relative absolute error, root mean square error, mean absolute error [15], and C-index [14,16,18], which makes it difficult to perform a comparison between the studies.

The performance of the proposed DeepHit and RNN models evaluated with the C-index is comparable with the previously published iChooseKidney technological solution (0.6640 at 3 years after transplantation) [14] and slightly exceeds the performance of the deep learning survival model described by Luck et al [16] (0.6550). However, the comparison of models based on the C-index alone is limited to the evaluation of their discriminative ability and does not consider the average accuracy of the survival predictions. Making use of ICI and smoothed calibration curves [31,32] helped shed light on the model's predictive quality.

From the results presented in Table 1 and Figure 3, we can see that there is often an imbalance between a model's discriminative ability and its calibration. As discriminative ability is required to differentiate between high-risk and low-risk kidney transplants, one might prefer a model with a higher C-index if a comparison of donor-candidate pairs is to be performed, for example, in the case of organ allocation. In contrast, as good calibration is required to provide reliable graft survival predictions, a model with better calibration may be preferable in cases where personalized expected graft survival distributions are to be presented, for example, to a transplant candidate.

Characteristics of the Developed Technological Solution

We developed a client web application to predict organ survival probability for each potential kidney donor-recipient pair for a period between 1 and 15 years after the transplantation. We opted to use the proposed RNN model to deploy our prototype application. This model offers a compromise between the good discriminative ability and the calibration necessary for the purpose of our application. Indeed, one of the main uses of the decision support application is to simultaneously present graft survival probabilities to a kidney transplant candidate and to offer a point of comparison by presenting graft survival predictions that the patient could expect with other potential donors.

It would be possible to use an alternative approach for computing the predictions at the time *now + average time before a new kidney is available*. To achieve this purpose, it would be necessary to compute the survival for every possible additional wait time and the probability of that wait time occurring, along with the patient survival to that wait time. This could be an objective for future studies.

The presented choice of approach to evaluate the *average donor* predictions at the same time *now* as the predictions for the offered donor kidney is a matter of simplicity and an effective way for patients without statistical background to look at 2 options (accept or refuse the transplant) and understand the possible outcomes.

The client application is at the prototype stage (TRL-4), aiming to demonstrate the capabilities of the ML predictive model. The following information about the candidate recipient is entered in the first step of the application: age, height, weight, ethnicity,

sex, diagnosis, number of years on dialysis, presence of diabetes, and presence of angina. The details about potential donor that are entered in the next step are donor's age, height, weight, ethnicity, donation type, creatinine level, history of diabetes, hypertension diagnosis, hepatitis C diagnosis, and smoking habit. These covariates are used as input for the trained RNN model. In the next step, the user selects the number of years for the prediction target. The output page displays the probability of survival of the transplant for the given donor-recipient pair and specified period as well as for the candidate recipient and average NDD and DCD donors for comparison. It is also possible to expand the result boxes to obtain a detailed view of the results for any specific transplant prediction.

Future Perspectives

The current application is recipient-oriented and specific to kidney transplantation. Future research could expand this application to other transplanted organs and nonrecipient users. For example, if connected to a candidate database, the application can produce an ordered list of optimal donor-recipient matches when an organ becomes available. The Expo.io development environment for the client was chosen for its capability to support web, Android, and iOS environments, leaving many options open for the distribution and accessibility of the service. The client also connects to the model by using an application programming interface. Thus, although the initial prototype was entirely run in a local environment, the solution could easily be transferred to a cloud-based environment.

In the future, the application could also be extended to include additional predictive models to further inform patients. For example, when a kidney is offered to a patient, it would be instructive to predict the expected waiting time before a *better* kidney becomes available should the patient decide to remain on the waiting list. The solution could also be upgraded to enable the recommendation of the best candidate recipient for each newly available kidney from the existing candidate waiting list based on the predicted graft survival.

Limitations

Our study has certain limitations, which are important to mention. A built-in selection bias exists in the SRTR data set. It is evident that deceased donor kidneys accepted for transplantation have superior characteristics than those that were never used for transplantation and therefore do not appear in the data. The data were imbalanced according to different age, sex, and racial groups. These selection biases may negatively affect the accuracy of predictions made for candidate recipients or donors who fall into underrepresented populations.

Another limitation is the level of detail available in the data set. The registry-level data from the SRTR certainly does not encapsulate all the characteristics of the clinical and functional status of donor-recipient pairs. Consequently, there must be factors that influence graft survival that were not present in the data. We also did not consider HLA typing, an important variable when matching donors and recipients, because of the complexity of modeling HLA mismatches. We must also consider the population of the United States, on which the models were built. Multiple factors, such as age, race, and state of residency, may reflect the socioeconomic status of patients, which itself may affect access to health care. To use the models built in this study in other countries, for example, in Canada, one must consider that some factors may differently affect graft survival.

Conclusions

We analyzed and tested 5 ML models to predict kidney graft survival for a period of up to 15 years after transplantation. This study focused on patients who received deceased donor kidney transplants in the United States between 2000 and 2019 and included both NDD and DCD transplants. The resulting RNN predictive model was integrated into a decision support application designed to help kidney transplant practitioners and their patients make informed decisions regarding transplant options.

Acknowledgments

The project was supported by Health Canada as a part of Innovative Solutions Canada Challenge "Machine learning to improve organ donation rates and make better matches" Phase 1. ADV received funding from the Mathematics of Information Technology and Complex Systems Accelerate internship.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Variables included in the machine learning models training. The original categories, the CAN_DGN variable (Candidate kidney diagnosis) and REC_FUNCTN_STAT (Candidate functional status), from the SRTR data set were grouped according to the previous work of Mark et al [22].

[DOCX File , 21 KB - medinform_v10i6e34554_app1.docx]

References

1. OPTN/SRTR 2017 annual data report: preface. Am J Transplant 2019 Feb;19 Suppl 2:1-10 [FREE Full text] [doi: 10.1111/ajt.15272] [Medline: 30811889]

2. Shemie SD, Baker AJ, Knoll G, Wall W, Rocker G, Howes D, et al. National recommendations for donation after cardiocirculatory death in Canada: donation after cardiocirculatory death in Canada. CMAJ 2006 Oct 10;175(8):S1 [FREE Full text] [doi: 10.1503/cmaj.060895] [Medline: 17124739]
3. Deceased organ donor potential in Canada. CIHI. URL: https://www.cihi.ca/sites/default/files/organdonorpotential_2014_en_0.pdf [accessed 2022-03-03]
4. Organ donation and transplantation in Canada. Government of Canada. URL: https://publications.gc.ca/site/eng/9.856553/publication.html [accessed 2022-03-04]
5. Hart A, Bruin M, Chu S, Matas A, Partin MR, Israni AK. Decision support needs of kidney transplant candidates regarding the deceased donor waiting list: a qualitative study and conceptual framework. Clin Transplant 2019 May;33(5):e13530 [FREE Full text] [doi: 10.1111/ctr.13530] [Medline: 30865323]
6. Organ Procurement and Transplantation Network (OPTN) policies. Organ Procurement and Transplantation Network. URL: https://optn.transplant.hrsa.gov/media/eavh5bf3/optn_policies.pdf [accessed 2022-05-14]
7. A guide to calculating and interpreting the Estimated Post-Transplant Survival (EPTS) score used in the Kidney Allocation System (KAS). Organ Procurement and Transplantation Network. URL: https://optn.transplant.hrsa.gov/media/1511/guide_to_calculating_interpreting_epts.pdf [accessed 2022-03-04]
8. Lee AP, Abramowicz D. Is the Kidney Donor Risk Index a step forward in the assessment of deceased donor kidney quality? Nephrol Dial Transplant 2015 Aug 04;30(8):1285-1290. [doi: 10.1093/ndt/gfu304] [Medline: 25282158]
9. Husain SA, King KL, Dube GK, Tsapepas D, Cohen DJ, Ratner LE, et al. Regional disparities in transplantation with deceased donor kidneys with kidney donor profile index less than 20% among candidates with top 20% estimated post transplant survival. Prog Transplant 2019 Dec 10;29(4):354-360 [FREE Full text] [doi: 10.1177/1526924819874699] [Medline: 31506000]
10. Kidney allocation in Canada: a Canadian forum. CCDT. URL: https://profedu.blood.ca/sites/msi/files/Kidney_Allocation_FINAL.pdf [accessed 2022-03-04]
11. Akl A, Ismail AM, Ghoneim M. Prediction of graft survival of living-donor kidney transplantation: nomograms or artificial neural networks? Transplantation 2008 Nov 27;86(10):1401-1406. [doi: 10.1097/TP.0b013e31818b221f] [Medline: 19034010]
12. Topuz K, Zengul FD, Dag A, Almehmi A, Yildirim MB. Predicting graft survival among kidney transplant recipients: a Bayesian decision support model. Decision Support Syst 2018 Feb;106:97-109. [doi: 10.1016/j.dss.2017.12.004]
13. Decruyenaere A, Decruyenaere P, Peeters P, Vermassen F. Validation in a single-center cohort of existing predictive models for delayed graft function after kidney transplantation. Ann Transplant 2015;20:544-552. [doi: 10.12659/aot.894034]
14. Patzer R, Basu M, Larsen CP, Pastan SO, Mohan S, Patzer M, et al. iChoose kidney: a clinical decision aid for kidney transplantation versus dialysis treatment. Transplantation 2016 Mar;100(3):630-639 [FREE Full text] [doi: 10.1097/TP.0000000000001019] [Medline: 26714121]
15. Reinaldo F, Rahman MA, Alves CF, Malucelli A, Camacho R. Machine learning support for kidney transplantation decision making. In: Proceedings of the International Symposium on Biocomputing. 2010 Presented at: ISB '10: International Symposium on BioComputing; Feb 15 - 17, 2010; Calicut Kerala India. [doi: 10.1145/1722024.1722079]
16. Luck M. Deep learning for patient-specific kidney graft survival analysis. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); Dec 4 - 9, 2017; Long Beach California USA.
17. Kilambi V, Bui K, Hazen GB, Friedewald JJ, Ladner DP, Kaplan B, et al. Evaluation of accepting kidneys of varying quality for transplantation or expedited placement with decision trees. Transplantation 2019 May;103(5):980-989 [FREE Full text] [doi: 10.1097/TP.0000000000002585] [Medline: 30720682]
18. Loupy A, Aubert O, Orandi BJ, Naesens M, Bouatou Y, Raynaud M, et al. Prediction system for risk of allograft loss in patients receiving kidney transplants: international derivation and validation study. BMJ 2019 Sep 17;366:l4923 [FREE Full text] [doi: 10.1136/bmj.l4923] [Medline: 31530561]
19. Schold JD, Arrigain S, Flechner SM, Augustine JJ, Sedor JR, Wee A, et al. Dramatic secular changes in prognosis for kidney transplant candidates in the United States. Am J Transplant 2019 Feb 14;19(2):414-424 [FREE Full text] [doi: 10.1111/ajt.15021] [Medline: 30019832]
20. Poggio ED, Augustine JJ, Arrigain S, Brennan DC, Schold JD. Long-term kidney transplant graft survival-making progress when most needed. Am J Transplant 2021 Aug 08;21(8):2824-2832. [doi: 10.1111/ajt.16463] [Medline: 33346917]
21. Lin RS, Horn SD, Hurdle JF, Goldfarb-Rumyantzev AS. Single and multiple time-point prediction models in kidney transplant outcomes. J Biomed Inform 2008 Dec;41(6):944-952 [FREE Full text] [doi: 10.1016/j.jbi.2008.03.005] [Medline: 18442951]
22. Mark E, Goldsman D, Gurbaxani B, Keskinocak P, Sokol J. Using machine learning and an ensemble of methods to predict kidney transplant survival. PLoS One 2019 Jan 9;14(1):e0209068 [FREE Full text] [doi: 10.1371/journal.pone.0209068] [Medline: 30625130]
23. Cox DR. Regression models and life-tables. J Royal Stat Soc Series B (Methodological) 2018 Dec 05;34(2):187-202. [doi: 10.1111/j.2517-6161.1972.tb00899.x]

24. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 2018 Feb 26;18(1):24 [FREE Full text] [doi: 10.1186/s12874-018-0482-1] [Medline: 29482517]
25. Lee C, Zame W, Yoon J, van der Schaar M. DeepHit: a deep learning approach to survival analysis with competing risks. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2018 Presented at: Proceedings of the AAAI Conference on Artificial Intelligence; Feb 2-7, 2018; New Orleans, Louisiana, USA.
26. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008 Sep 1;2(3):841-860. [doi: 10.1214/08-aoas169]
27. Breiman L. Random forests. Mach Learn 2001;45(1):5-32 [FREE Full text] [doi: 10.1023/A:1010933404324]
28. Ren K, Qin J, Zheng L, Yang Z, Zhang W, Qiu L, et al. Deep recurrent survival analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019 Presented at: Proceedings of the AAAI Conference on Artificial Intelligence; Jan 27-Feb 1, 2019; Honolulu, HI, USA. [doi: 10.1609/aaai.v33i01.33014798]
29. RNN-SURV: a deep recurrent model for survival analysis. In: Artificial Neural Networks and Machine Learning. Cham: Springer; 2018.
30. Moore D. Applied Survival Analysis Using R. Cham: Springer; 2016.
31. Harrell F, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 1982 May 14;247(18):2543-2546. [Medline: 7069920]
32. Antolini L, Boracchi P, Biganzoli E. A time-dependent discrimination index for survival data. Stat Med 2005 Dec 30;24(24):3927-3944. [doi: 10.1002/sim.2427] [Medline: 16320281]
33. Brier GW. Verification of forecasts expressed in terms of probability. Monthly Weather Rev 1950 Jan;78(1):1-3. [doi: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2]
34. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Stat Med 1999 Sep 15;18(17-18):2529-2545. [doi: 10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5]
35. Haider H, Hoehn B, Davis S, Greiner R. Effective ways to build and evaluate individual survival distributions. J Mach Learn Res 2020;21:1-63 [FREE Full text]
36. Austin PC, Harrell FE, van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for survival models. Stat Med 2020 Sep 20;39(21):2714-2742 [FREE Full text] [doi: 10.1002/sim.8570] [Medline: 32548928]
37. Harrell F. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Cham: Springer; 2006.
38. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med 2019 Sep 20;38(21):4051-4065 [FREE Full text] [doi: 10.1002/sim.8281] [Medline: 31270850]
39. Lofaro D, Maestripieri S, Greco R, Papalia T, Mancuso D, Conforti D, et al. Prediction of chronic allograft nephropathy using classification trees. Transplant Proc 2010 May;42(4):1130-1133. [doi: 10.1016/j.transproceed.2010.03.062] [Medline: 20534242]
40. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014 Dec 22;14(1). [doi: 10.1186/1471-2288-14-137]
41. Lasserre J, Arnold S, Vingron M, Reinke P, Hinrichs C. Predicting the outcome of renal transplantation. J Am Med Inform Assoc 2012 Mar 01;19(2):255-262 [FREE Full text] [doi: 10.1136/amiajnl-2010-000004] [Medline: 21875867]
42. Senanayake S, White N, Graves N, Healy H, Baboolal K, Kularatna S. Machine learning in predicting graft failure following kidney transplantation: a systematic review of published predictive models. Int J Med Inform 2019 Oct;130:103957. [doi: 10.1016/j.ijmedinf.2019.103957] [Medline: 31472443]
43. Goldfarb-Rumyantzev A, Scandling JD, Pappas L, Smout RJ, Horn S. Prediction of 3-yr cadaveric graft survival based on pre-transplant variables in a large national dataset. Clin Transplant 2003 Dec;17(6):485-497. [doi: 10.1046/j.0902-0063.2003.00051.x] [Medline: 14756263]
44. Nematollahi M, Akbari R, Nikeghbalian S, Salehnasab C. Classification models to predict survival of kidney transplant recipients using two intelligent techniques of data mining and logistic regression. Int J Organ Transplant Med 2017;8(2):119-122 [FREE Full text] [Medline: 28959387]

Abbreviations

DCD: donation after circulatory death
IBS: integrated Brier score
ICI: integrated calibration index
ML: machine learning
NDD: neurological determination of death
RNN: recurrent neural network
RSF: random survival forest

SRTR: Scientific Registry of Transplant Recipients
TRL-4: technology readiness level 4

Edited by C Lovis; submitted 28.10.21; peer-reviewed by C Ta, J Zhang; comments to author 16.01.22; revised version received 29.01.22; accepted 31.01.22; published 14.06.22.

Please cite as:
Paquette FX, Ghassemi A, Bukhtiyarova O, Cisse M, Gagnon N, Della Vecchia A, Rabearivelo HA, Loudiyi Y
Machine Learning Support for Decision-Making in Kidney Transplantation: Step-by-step Development of a Technological Solution
JMIR Med Inform 2022;10(6):e34554
URL: https://medinform.jmir.org/2022/6/e34554
doi:10.2196/34554
PMID:35700006

Original Paper

Multitask Learning With Recurrent Neural Networks for Acute Respiratory Distress Syndrome Prediction Using Only Electronic Health Record Data: Model Development and Validation Study

Carson Lam[1*], MD; Rahul Thapa[1*], BSc; Jenish Maharjan[1*], MSc; Keyvan Rahmani[1*], PhD; Chak Foon Tso[1], PhD; Navan Preet Singh[1], MSc; Satish Casie Chetty[1], PhD; Qingqing Mao[1], PhD

Dascena, Inc, Houston, TX, United States

[*]these authors contributed equally

Corresponding Author:
Satish Casie Chetty, PhD
Dascena, Inc
12333 Sowden Road
Suite B
Houston, TX, 77080
United States
Phone: 1 5132080270
Email: dchetty@dascena.com

Abstract

Background: Acute respiratory distress syndrome (ARDS) is a condition that is often considered to have broad and subjective diagnostic criteria and is associated with significant mortality and morbidity. Early and accurate prediction of ARDS and related conditions such as hypoxemia and sepsis could allow timely administration of therapies, leading to improved patient outcomes.

Objective: The aim of this study is to perform an exploration of how multilabel classification in the clinical setting can take advantage of the underlying dependencies between ARDS and related conditions to improve early prediction of ARDS in patients.

Methods: The electronic health record data set included 40,703 patient encounters from 7 hospitals from April 20, 2018, to March 17, 2021. A recurrent neural network (RNN) was trained using data from 5 hospitals, and external validation was conducted on data from 2 hospitals. In addition to ARDS, 12 target labels for related conditions such as sepsis, hypoxemia, and COVID-19 were used to train the model to classify a total of 13 outputs. As a comparator, XGBoost models were developed for each of the 13 target labels. Model performance was assessed using the area under the receiver operating characteristic curve. Heat maps to visualize attention scores were generated to provide interpretability to the neural networks. Finally, cluster analysis was performed to identify potential phenotypic subgroups of patients with ARDS.

Results: The single RNN model trained to classify 13 outputs outperformed the individual XGBoost models for ARDS prediction, achieving an area under the receiver operating characteristic curve of 0.842 on the external test sets. Models trained on an increasing number of tasks resulted in improved performance. Earlier prediction of ARDS nearly doubled the rate of in-hospital survival. Cluster analysis revealed distinct ARDS subgroups, some of which had similar mortality rates but different clinical presentations.

Conclusions: The RNN model presented in this paper can be used as an early warning system to stratify patients who are at risk of developing one of the multiple risk outcomes, hence providing practitioners with the means to take early action.

(JMIR Med Inform 2022;10(6):e36202) doi:10.2196/36202

KEYWORDS

deep learning; neural networks; ARDS; health care; multitask learning; clinical decision support; prediction model; COVID-19; electronic health record; risk outcome; respiratory distress; diagnostic criteria; recurrent neural network

Introduction

Background

Acute respiratory distress syndrome (ARDS) is a heterogeneous syndrome broadly characterized by noncardiogenic hypoxia, pulmonary edema, and the need for mechanical ventilation [1,2]. Despite advances made in the diagnosis and management of patients with ARDS, ARDS is present in approximately 10% of the patients admitted to intensive care units (ICUs) worldwide, and mortality is as high as 30% to 40% in most studies [1]. Tools such as the 2016 Kigali modification of the 2012 Berlin criteria have been developed to aid clinicians to diagnose patients with ARDS [3,4]. In addition, the Lung Injury Prediction Score and Early Acute Lung Injury score were developed to identify and stratify patients at risk of developing ARDS based on a collection of physiological variables and predisposing conditions [5-7]. However, ARDS only occurs in a small proportion of patients with a risk factor and currently there is no consensus on how or whether patients should be screened for ARDS. This becomes especially important in the context of patients who are critically ill in the ICU, where health care providers may experience challenges in continuous monitoring and processing large amounts of clinical data from patients. Early warning of impending ARDS should allow implementation of lower tidal volumes in breathing support and more careful fluid management, the 2 main strategies to prevent or reduce the severity of ARDS [8,9].

In the past decade, artificial intelligence has shown great promise in medicine, with potential applications across multiple domains in health care [10]. There have been significant advances in harnessing the power of big data from electronic health records (EHRs) to develop machine learning algorithms to predict the onset of a broad spectrum of medical conditions in patients. A wide variety of such algorithms have been studied and implemented by groups in both academia and industry [11-16]. Previous studies have suggested that the physiological states that exist early in the presentation of ARDS can be used to predict ARDS before the confirmatory tests required by gold standards such as the Berlin criteria [17,18], which requires radiology reports that are by nature subsequent to clinical suspicion. Early prediction is desirable because it would lead to earlier intervention, more time for the careful administration of treatment, or modification of the ongoing treatment, which in turn should lead to improved outcomes such as reduced morbidity and mortality.

Objectives

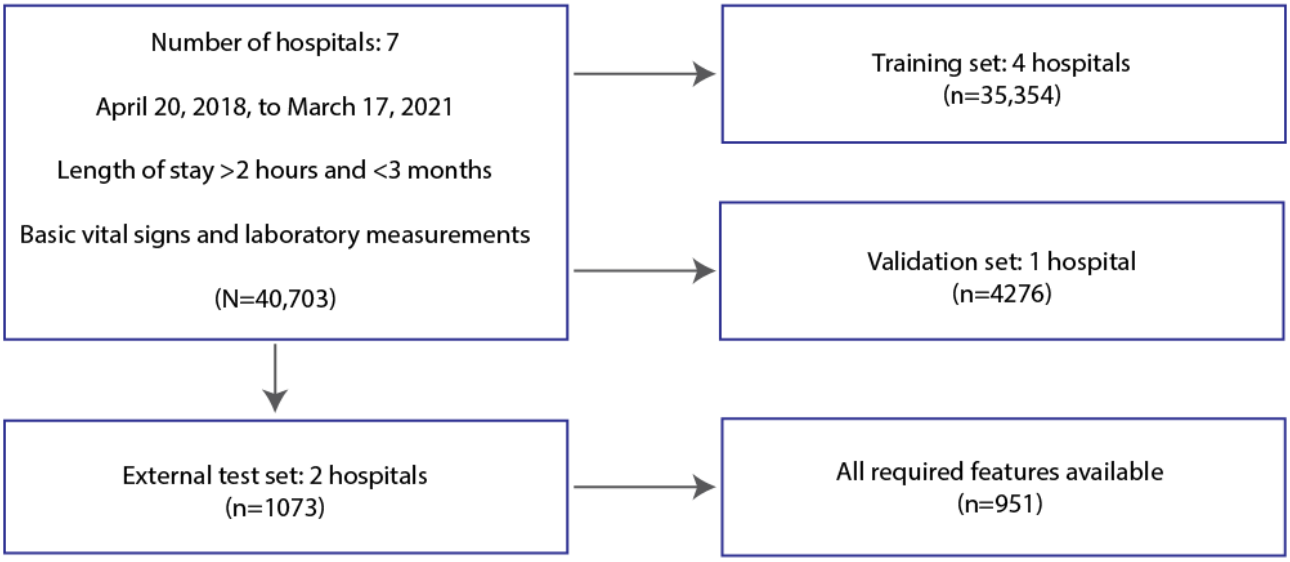
In real-world clinical settings, the task is to anticipate multiple diseases or clinical states. In this study, we aimed to demonstrate that multitask learning using deep learning models provides benefits over single-task machine learning models. To this end, we focused on the detection and early prediction of varying severities of ARDS together with sepsis, COVID-19, hypoxemia, and in-hospital mortality. Previous studies have developed multilabel classification models that predict multiple medical outcomes simultaneously. For example, Maxwell et al [19] and Zhang et al [20] used deep neural networks to predict multiple chronic diseases such as hypertension and diabetes and Lipton et al [21] used recurrent neural networks (RNNs) to classify 128 different diagnoses. Although research has been conducted on developing single-task learning models for ARDS prediction [22-25], thus far no studies have explored multilabel classification models for predicting ARDS. Here, we aimed to perform a deep analysis of how multilabel classification in the clinical setting can take advantage of the underlying dependencies among different diseases to allow for improved performance for the prediction of ARDS in patients over single-label classification models [26]. In addition, although research has been conducted showing that early disease prediction is possible, here we also present estimates supporting that early prediction of ARDS is beneficial. Finally, we explore an interesting additional benefit of using neural networks in hospitals and the identification of distinct disease phenotypes.

Methods

Data Description

The data set included 40,703 patient encounters whose care settings included the emergency department, inpatient facility, or ICU. All clinical information was drawn from patient EHR data from 7 different hospitals between April 20, 2018, and March 17, 2021, as shown in Figure 1**.** Data collection was passive, and all patient information was deidentified before the analysis performed in this study. Radiology data were not available. This prevented direct measurement of the Berlin criteria for ARDS [18]. The information collected from each hospital included discharge disposition, demographic data such as age and sex, and time-varying data, including vital signs, laboratory values, oxygen delivery method, medications, and diagnosis times of any conditions present in the health record. These data were extracted as an unordered record of data type, data value, data units, and data collection time, also known as datetime. Preprocessing of these data first involved reordering the data in chronological order under each data type, with 3 equal-length arrays representing the values, units, and datetimes of each measurement. The data were split into training, validation, and external test data sets based on the hospital sites. The training data set used data from 5 hospitals, the validation data set used data from 1 hospital, and the external test data set used data from 2 hospitals. The training and validation data sets were used during the development of the machine learning models, and the external test data set was used to evaluate the models trained on the combination of training and validation data sets.

Figure 1. Flowchart of patients. Among 7 hospitals, 40,703 patients met three criteria: (1) admission within the date range (April 20, 2018, to March 17, 2021), (2) length of stay within the range of 2 hours to 3 months, and (3) availability of basic vital signs (blood pressure, heart rate, temperature, respiratory rate, and peripheral oxygen saturation) and laboratory measurements (complete blood count and basic metabolic panel) in the electronic health record. These patients were separated into training, validation, and test sets based on their hospital sites. The test set was limited to those patients with the required features listed in Textbox 1 consisting of age, sex, and basic laboratory measurements, as well as complete blood count with differential.



Input Features

Model inputs were a defined set of data types, or features, across all hospitals, regardless of the data availability at a particular hospital. Textbox 1 includes all the features used to train the models in this study. The required features are the subset of features, including age, sex, and basic laboratory measurements, as well as complete blood count with differential, used to determine the time the algorithm makes its prediction. Next, these data values were organized into a matrix with features along the first dimension (rows) and discrete time in 20-minute intervals along the second dimension (columns). The first column, column index 0, contains the first time point of any vital sign or laboratory measurement and was considered to be the start of care. The first row was normalized age. The second and third rows were binary indicators for male and female sex. The remaining rows were the time-varying features and their corresponding mask to distinguish missing values from actual zeros (Table 1).

To normalize the features, we carried out a coarse approximation of the mean and SD based on the normal range of these features in the laboratory reports. The center of the normal range was used as the approximate mean value, and half of the difference between the 2 end points of the range was used to approximate the SD. If a feature was missing or not measured, it was set to 0. To let the model distinguish between null values and real values, a new set of features representing the availability mask was vertically appended to the matrix. Each feature row had a corresponding binary mask vector that contained 0s and 1s, representing null values and nonnull values, respectively. During batch training, these matrices were 0 padded on the left side into equal-sized tensors: (batch size, 58 features, 64 timesteps).

Textbox 1. Input features to the machine learning algorithm.

Demographics

- Age (required feature)
- Sex (required feature)

Vital signs

- Systolic blood pressure (required feature)
- Diastolic blood pressure (required feature)
- Heart rate (required feature)
- Arterial partial pressure of oxygen
- Respiratory rate (required feature)
- Peripheral oxygen saturation (required feature)
- Temperature (required feature)

Laboratory results

- Glucose
- Bilirubin
- White blood cell count (required feature)
- Red blood cell count
- Lymphocytes (required feature)
- Alanine transaminase
- International normalized ratio
- pH
- Blood urea nitrogen
- Creatinine (required feature)
- Platelet
- Neutrophils (required feature)
- Monocytes
- Hematocrit
- Lactate
- Aspartate aminotransferase

Other measurements

- Systemic inflammatory response syndrome (the systemic inflammatory response syndrome score is calculated as shown in Table S1 in Multimedia Appendix 1)

Table 1. Recurrent neural network features. The first 3 rows are the raw values, the next 3 rows are the corresponding normalized features, and the last 3 rows are the corresponding availability masks to distinguish missing values from actual zeros. The mask is a Boolean vector that is 0 if that measurement is missing and 1 if that measurement is present. It should be noted that this is a subset of the total features and timesteps used for example purposes only. The raw data are also for illustration and not a part of the input feature matrix, which consists of normalized features at 64 timesteps.

	0 minutes	20 minutes	40 minutes	60 minutes
Raw data				
SpO_2[a] (%)	90	100	99	80
Creatinine (mg/dL)	None	1.8	1.8	1.0
WBC[b] ($\times 10^9$/L)	None	None	None	12
Normalized				
SpO_2	–0.5	+0.02	0	–2.3
Creatinine	0	+0.2	+0.2	–0.11
WBC	0	0	0	+0.15
Mask				
SpO_2	1	1	1	1
Creatinine	0	1	1	1
WBC	0	0	0	1

[a]SpO_2: peripheral oxygen saturation.

[b]WBC: white blood cell.

Model Output and Targets

Additional target labels were chosen for the model that are distinct from ARDS, yet clinically related to it such that a collective representation in the neural network is justified. These labels are shown in Textbox 2 (these output labels will be used throughout the paper hereafter). The model was trained to predict these target labels, also referred to as outcomes, using a binary cross-entropy loss function. The descriptive statistics for the input features for each of the target outcomes are presented in Tables S2 and S3 in Multimedia Appendix 1 for training and test data.

Textbox 2. Clinical outcomes used as target labels for the machine learning algorithm. In total, 13 output labels were mapped to their respective definition.

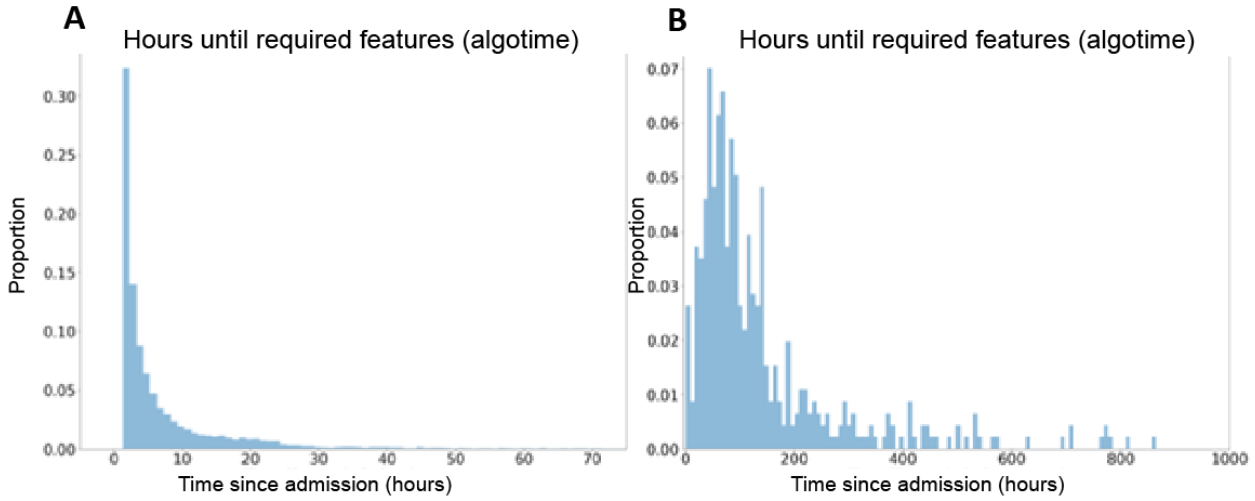
Output label and definition

- Acute respiratory distress syndrome (ARDS)-1: ARDS defined as having an International Classification of Diseases (ICD) code for ARDS as well as a drop in peripheral oxygen saturation (SpO_2) below 91%
- ARDS-2: ARDS defined as having an ICD code for ARDS as well as a drop in SpO_2 below 96%. A direct but broader criterion for ARDS
- ARDS-3: ARDS defined as having an ICD code for ARDS as well as a drop in SpO_2 below 91% and no mention of a heart failure–related ICD code among prior diagnoses
- ARDS-4: ARDS defined as having an ICD code for ARDS as well as a drop in SpO_2 below 96% and no mention of a heart failure–related ICD code among prior diagnoses
- ARDS-5: ARDS defined as having an ICD code for ARDS. A direct and simple definition of ARDS
- Sepsis-6: sepsis defined as having an ICD code for sepsis or septic shock as well as a systemic inflammatory response syndrome score >2
- Sepsis-7: sepsis defined as having an ICD code for sepsis or septic shock. A direct and simple definition of sepsis
- Hypoxemia-8: a drop in SpO_2 below 91% any time during hospitalization
- Hypoxemia-9: a drop in SpO_2 below 96% any time during hospitalization
- Hypoxemia-10: a drop in SpO_2 below 91% after algorithm evaluates
- Hypoxemia-11: a drop in SpO_2 below 96% after algorithm evaluates
- Death-12: in-hospital mortality
- Covid-13: COVID-19 positivity defined as in-hospital COVID-19 positive polymerase chain reaction test or new ICD diagnosis within 7 days before or after admission

Timing of Algorithm Evaluation

For simplicity, we evaluated the algorithm at a single point in time. This time is 2 timesteps (40 minutes) after the first time at which all required features have been measured at least once. At this time, which we refer to as the algotime, the model predicts all the target outcomes of Textbox 2. In the training and validation sets, we used the required features to determine algotime, but in the case of missing features, it defaults to 8 hours after admission. In the test set, we only included patients who had all the required features. With regard to padding, if there are <64 timesteps available before algotime, the input sequence is 0 padded on the left; if there are >64 timesteps available before algotime, only the most recent 64 are taken (no padding). As can be seen in Figure 2**,** on average, the algotime occurred 31 hours after admission and ARDS was clinically diagnosed 139 hours after admission. Thus, the average number of hours between the algotime and the clinical diagnosis of ARDS was 108 hours. It should be noted that the performance statistics reported in this paper correspond to the algotime, not the time of clinical ARDS diagnosis or the end of the hospital stay.

Figure 2. Timing of algorithm and diagnosis. (A) Histogram of number of hours between start of care and the first time point when all required features have been measured at least once (algotime). (B) Histogram of number of hours from start of care until new diagnosis of acute respiratory distress syndrome (ARDS) is entered into the electronic health record. Both histograms are based on the entire data (training+validation+test sets).



Benefit Estimation

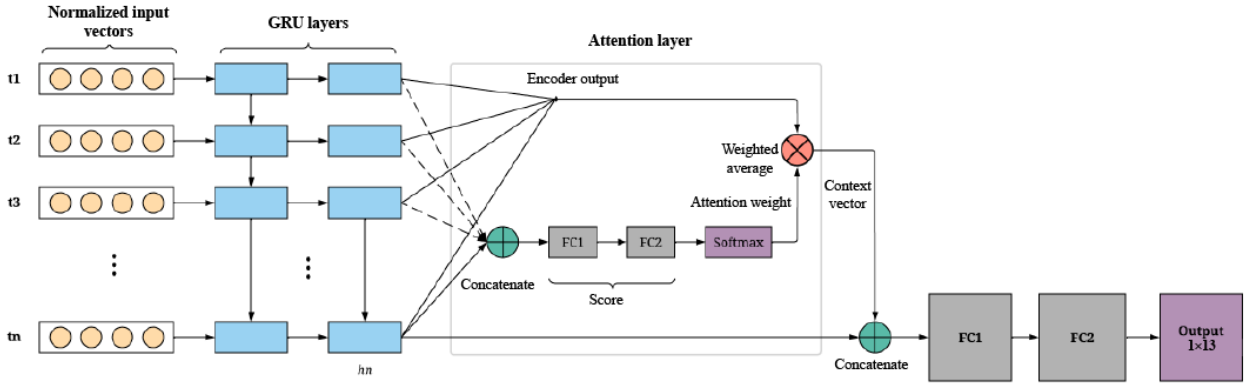
Thus far, we have defined our machine learning objective as the early prediction of conditions. However, the impact of early predictions on patient health outcomes is more important than the early prediction. To approximate this improvement in the outcome of mortality, we compared mortality rates between patients who received early and late clinical diagnoses of ARDS. We defined early diagnosis and late diagnosis based on when a patient was given a clinical diagnosis of ARDS compared with when the algorithm made a prediction. In other words, a diagnosis for ARDS (using the ARDS-1 definition in Textbox 2 of ARDS International Classification of Diseases [ICD] code and peripheral oxygen saturation [SpO_2] below 91%) is *early* if it is assigned before the algorithm makes a prediction and *late* if it is assigned after a prediction is made by the algorithm.

Machine Learning Models

We used an RNN as the main deep learning model for our research. RNNs are a class of artificial neural networks in which connections among nodes form a directed graph along a temporal sequence. RNNs can use their internal memory to process variable length sequences of inputs. The network is capable of learning a mapping function from the inputs over time to an output. It can even learn temporal dependence from the data. All these properties make RNNs a well-suited model for time series data, such as that which is used in this study. The model schema of the RNN used in this research is presented in Figure 3. We used a generic RNN with 4 gated recurrent unit (GRU) layers, an attention module, and 2 fully connected (FC) layers for all numbers of outputs. The RNN was implemented with the PyTorch package (version 1.40) in Python (version 3.6; Python Software Foundation) [27]. For the RNN, the sequence module that was used was a 4-layer GRU [28] with 128 hidden units. Before the sequence of vectors was fed to the GRU, it passed through a normalization layer:

$$n(v) = a(v - \mu / \sigma) + b \quad \textbf{(1)}$$

Figure 3. Recurrent neural network model schema. The inputs have been simplified for diagrammatic purposes. Different timesteps of normalized inputs are fed into gated recurrent unit (GRU) layers. The context vector from the attention layer and encoder output from the last GRU are concatenated before feeding them into the first fully connected (FC) layer.



Equation 1 is a normalization function that learns the parameters mean μ, SD σ, scaling a, and translation factor b used to normalize the sequence of vectors containing the inputs age, sex (Boolean), vital signs, laboratory measurements, and systemic inflammatory response syndrome (SIRS) score before entering the RNN. A soft attention module was used to assign scores to each timestep in the sequence. The scores are intended to be positively correlated with the importance of its respective timestep. A weighted sum of the sequence's hidden activations was called the context vector. We concatenated the context vector to the final GRU embedding and passed this to a 2-layer feed forward neural network for classification. The intermediate layer before the output logits was a 128D representation of each patient, referred to as the penultimate embedding. Similar to Bahdanau et al [29], the score (equation 2) of the attention neural network was parameterized by a feed forward neural network of the following form:

$$\text{score}(h_l,h_i) = K^T tanh(W_a \text{ Prelu}(W_b[h_n,h_i])) \quad \textbf{(2)}$$

where *tanh* and *Prelu* denote the hyperbolic tangent function and parameterized rectified linear unit nonlinearity functions, respectively; h_n denotes the last hidden activation in the GRU; h_i denotes each hidden activation in the sequence; i denotes the timestep; [,] denotes concatenation of separate vectors into 1 vector; and K, W_a, and W_b denote learned parameters of the neural network. The whole GRU-RNN, attention module, and classification module were end-to-end differentiable, which enabled optimization from input to output. The attention neural network was a mechanism of the RNN that allowed for higher-quality learning. Instead of summarizing a time series of vectors, the attention neural network assigned each vector a score according to how important the vector was in allowing the model to make a prediction. In this way, the attention network mechanism allowed the RNN to focus on specific parts of the input, thereby enabling improved model performance.

Each point in the RNN model schema represents a neuron. At each layer, the RNN combined the information from the current and previous timesteps to update the activations of the deepest GRU hidden layer. The activation of the last node of the deepest RNN layer is concatenated with the context vector provided by the attention network. The context vector is an importance-weighted average of the deepest layer activations generated by the attention neural network. This concatenated vector is passed through 2 FC layers to generate an output (eg, prediction of ARDS onset). With this RNN schema, the model was trained to predict several target labels simultaneously and to evaluate a loss function based on all targets. We implemented a deep learning method where a single network was trained to output 1 logit per label using a binary cross-entropy loss function [30]. The loss function averages binary cross-entropies against each of the targets in the model, effectively taking into account the output of all 13 tasks. Considering each label a task, this multilabel learning setup can be viewed as a case of multitask learning [31]. Specifically, because all hidden layer parameters are shared among all the targets, it is a hard parameter–sharing variant of multitask learning [32]. Each output logit was independently passed through a sigmoid activation function to produce the final multilabel output [33]. Early stopping was used, based on the ARDS-1 validation performance measured by the area under the receiver operating characteristic curve (AUROC). To explore the relationship between the objective function's number of targets and final model performance, the lowest 2 AUROC targets were removed successively from each version of the RNN such that the RNN was trained using 13, 11, 9, 7, and 5 targets.

Tree-based models frequently outperform deep learning models in many clinical applications [34]. To ensure that this was not the case in this instance, for comparison, XGBoost (XGB) models for each of the target labels were trained using XGBoost (version 0.81) [35] in Python (version 3.6) [36] and the same feature matrix as the RNN model. The XGB models were trained in a one-versus-all fashion for each target.

Model Interpretability

Heat maps were produced to visualize attention scores on each time series. Of the 959 patients in the test set, 50 (5.21%) were randomly selected for the following two interpretability analyses: (1) attention scores were visualized across timesteps as heat maps, and (2) the timestep with the highest attention weight generated by the attention network was then further analyzed to visualize each feature's deviation from the mean in this heavily attended timestep. This method implicitly describes the importance assigned to each feature by the model and

provides some insight into model interpretability. The feature vector at that timestep is interpreted as a z-score for the subset of features measured at that particular timestep. For example, a value of 0.5 for the respiratory rate indicates that the respiratory rate is half an SD above the mean.

In addition, Shapley additive explanations (SHAP) force plots for 4 different patients were also generated. The patients represent true positive, true negative, false positive, and false negative cases for ARDS as predicted by the RNN model trained using 13 targets. We used the ARDS-5 definition from Textbox 2 (ARDS defined as having an ICD code for ARDS) for this analysis.

Clustering

To explore the representations used by the model and to reveal distinct phenotypes among patients with ARDS, we collected the 64D activations produced by the first FC layer as a compressed representation, or embedding, for each patient. To visually display these embeddings, we used principal component analysis. We then used k-means clustering to group each embedding of a patient with ARDS into its unique cluster.

Statistics

To compare different algorithms and training objectives, we computed the 95% CI around the AUROC using the bootstrapping method [37-39]. These CIs are with respect to the test set (n=959).

Ethics Approval

All patient data were deidentified in compliance with the Health Insurance Portability and Accountability Act. This study was considered to be of minimal risk for human participants because data collection was passive and did not pose a threat to the participants involved. The project was approved with a waiver of informed consent (20-DASC-122) by an independent institutional review board, Pearl Institutional Review Board.

Results

Comparison of RNN Model With XGB Model

The XGB and RNN models were compared across all 13 outputs (Table 2). The performance metric used for comparison was the AUROC. The single RNN model trained to classify 13 outputs outperformed the XGB models trained separately to classify each of the outputs in ARDS and oxygen-related outcomes. The average receiver operating characteristic curves for all targets are also presented in Figure S1 in Multimedia Appendix 1, which further illustrates that the RNN model performs at least as well as the average of all XGB models, with the added advantage that the RNN model benefits from parameter sharing; that is, a single RNN model performs at least as well as the aggregate of 13 XGB models.

Table 2. Comparison of performances of the XGB[a] models and RNN-13[b] on each of the outcomes. The table provides the AUROC[c] of each model for each of the labels as well as the sensitivity and specificity. We note that the RNN-13 model outperforms the XGB model on 7 out of 13 outputs.

Labels (prevalence)	XGB			RNN-13		
	AUROC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUROC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
ARDS-1[d,e] (0.046)	0.797 (0.740-0.851)	0.659 (0.519-0.799)	0.729 (0.7-0.758)	0.842 (0.794-0.888)	0.659 (0.519-0.799)	*0.873 (0.852-0.852)*[f]
ARDS-2 (0.054)	0.700 (0.632-0.768)	0.654 (0.525-0.783)	0.657 (0.626-0.688)	0.791 (0.746-0.836)	0.673 (0.546-0.801)	*0.780 (0.753-0.753)*
ARDS-3 (0.044)	0.786 (0.714-0.856)	0.667 (0.524-0.809)	0.771 (0.744-0.798)	0.845 (0.795-0.894)	0.667 (0.524-0.809)	*0.826 (0.802-0.802)*
ARDS-4 (0.051)	0.748 (0.681-0.81)	0.653 (0.520-0.786)	0.714 (0.685-0.744)	0.812 (0.768-0.858)	0.653 (0.520-0.786)	*0.804 (0.778-0.778)*
ARDS-5 (0.055)	0.701 (0.629-0.77)	0.660 (0.533-0.788)	0.681 (0.651-0.711)	0.795 (0.751-0.839)	0.660 (0.533-0.788)	*0.795 (0.769-0.769)*
Sepsis-6 (0.023)	0.708 (0.604-0.803)	0.682 (0.487-0.876)	0.547 (0.516-0.579)	0.626 (0.533-0.714)	0.682 (0.487-0.876)	0.503 (0.471-0.471)
Sepsis-7 (0.023)	0.707 (0.599-0.798)	0.682 (0.487-0.876)	0.715 (0.686-0.744)	0.586 (0.481-0.681)	0.682 (0.487-0.876)	0.502 (0.469-0.469)
Hypoxemia-8 (0.268)	0.722 (0.684-0.760)	0.658 (0.600-0.716)	0.657 (0.622-0.692)	0.739 (0.708-0.770)	0.651 (0.592-0.709)	0.684 (0.649-0.649)
Hypoxemia-9 (0.799)	0.829 (0.802-0.856)	0.657 (0.623-0.690)	0.876 (0.829-0.922)	0.834 (0.810-0.855)	0.659 (0.625-0.692)	0.839 (0.786-0.786)
Hypoxemia-10 (0.182)	0.643 (0.601-0.688)	0.651 (0.581-0.722)	0.536 (0.501-0.571)	0.638 (0.601-0.673)	0.655 (0.585-0.726)	0.524 (0.489-0.489)
Hypoxemia-11 (0.38)	0.880 (0.861-0.901)	0.654 (0.605-0.703)	0.859 (0.831-0.887)	0.880 (0.862-0.897)	0.651 (0.602-0.700)	0.856 (0.828-0.828)
Death-12 (0.026)	0.761 (0.675-0.841)	0.680 (0.497-0.863)	0.700 (0.671-0.730)	0.700 (0.625-0.768)	0.680 (0.497-0.863)	0.625 (0.594-0.594)
Covid-13 (0.238)	0.805 (0.770-0.840)	0.654 (0.592-0.715)	0.814 (0.786-0.842)	0.673 (0.637-0.714)	0.654 (0.592-0.715)	0.622 (0.587-0.587)

[a]XGB: XGBoost.

[b]RNN-13: the recurrent neural network model that was trained using 13 targets.

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

[d]ARDS: acute respiratory distress syndrome.

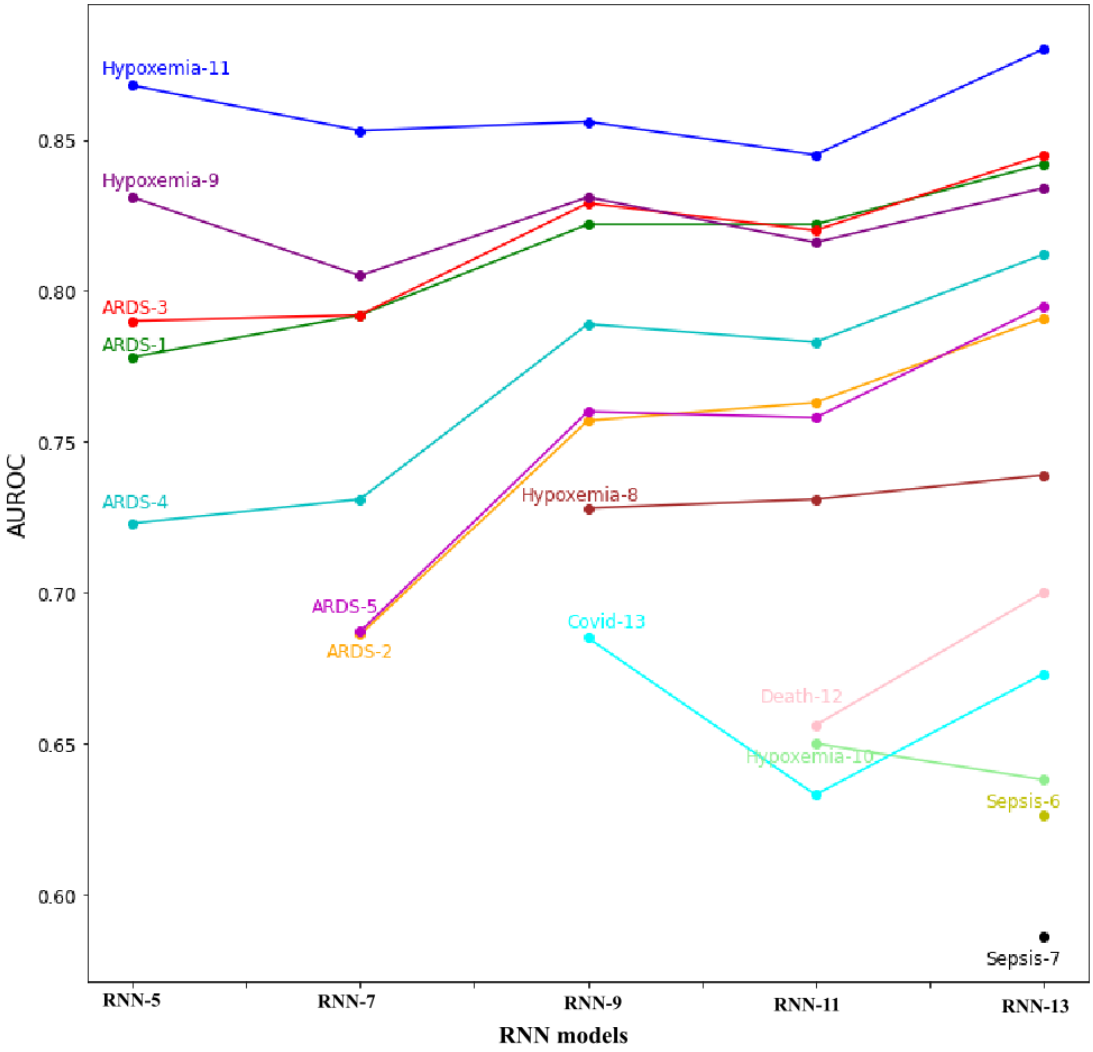
[e]Area under the receiver operating characteristic curve reported in the *Abstract*.

[f]Specificity of the RNN-13 model for the five ARDS labels.

Benefit of Multitask Learning

An intermediate number of output targets between 1 and 13 were also used to retrain the RNN. Figure 4 shows the maximum AUROC with different subsets of the 13 outcomes used as training targets. For most targets there is a general trend toward overall improvement of the AUROC. This demonstrates that there is some underlying dependency among some of the labels.

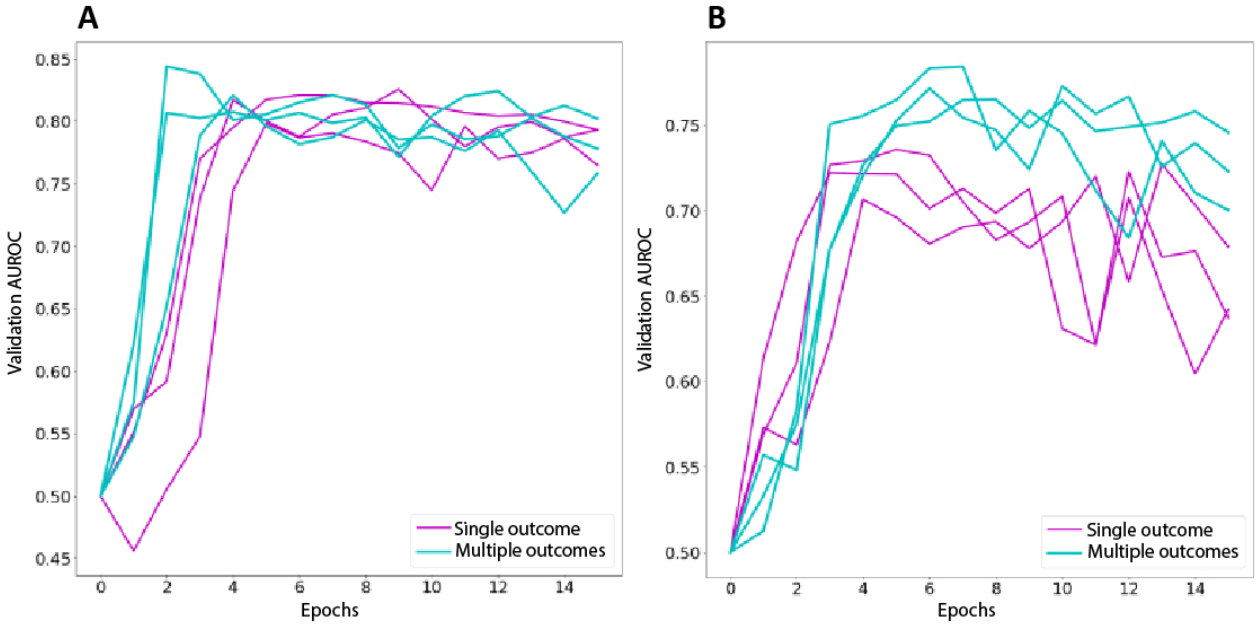
Figure 4. Model performance varies with the number of outcomes predicted during training. External test set area under the receiver operating characteristic curve (AUROC) plotted against the number of targets in the recurrent neural network (RNN) output (eg, RNN-9 refers to an RNN with 9 outputs). From right to left, the worst-performing 2 targets in terms of AUROC are removed to train the next RNN with a smaller number of targets. ARDS: acute respiratory distress syndrome.



Multitask Learning Converges Training in a Comparable Number of Epochs

The learning quality and efficiency of single- versus multiple-outcome models were evaluated in terms of the rate of improvement of the AUROC on the validation set per each stochastic gradient descent training epoch. We compared the rate of learning between RNNs trained with single targets and RNNs trained with multiple targets to demonstrate that multitask learning does not empirically require longer durations in training than single-learning objectives. The rate of learning was measured as the AUROC of the validation set for each epoch. In Figure 5, the plots of the AUROC of 3 separate randomly initialized training episodes for 15 epochs are shown for ARDS-1 and ARDS-2 (ARDS defined as having an ICD code for ARDS as well as a drop in SpO_2 below 96%). For these 2 outcomes, the time to reach the maximum validation AUROC in terms of the number of epochs is comparable between single- and multiple-target models.

Figure 5. Training on multiple targets converges in similar time to single targets. Learning progress measured using the area under the receiver operating characteristic curve (AUROC) on the validation set. Each line is a different training run with new randomized initial weights and training batches. (A) AUROC predicting ARDS-1 versus number of training epochs. (B) AUROC predicting ARDS-2 versus number of training epochs. ARDS-1: acute respiratory distress syndrome (ARDS) defined as having an International Classification of Diseases code for ARDS as well as a drop in peripheral oxygen saturation below 91%; ARDS-2: ARDS defined as having an International Classification of Diseases code for ARDS as well as a drop in peripheral oxygen saturation below 96%.



ARDS Clustering

Figure 6 shows the results of applying the k-means clustering algorithm to find clusters of patients with ARDS on the output of the first FC layer of the RNN model as mentioned in the *Machine Learning Models* section. The k-means algorithm was set to identify 3 clusters because this was the most distinguishable number of clusters in different visualizations of the embeddings. Figure 6 shows a 2D projection of the embeddings using principal component analysis in which the 3 clusters A, B, and C can be seen. Deeper analysis of these clusters is presented in Figure 7**,** which shows a fair amount of variability of the targets among clusters. It can be seen that a drop of SpO_2 below 91% is more likely to be characterized differently among the clusters than the drop of SpO_2 below 96% among ARDS targets, which is apparent in higher variability of ARDS-1 and ARDS-3 (ARDS defined as having an ICD code for ARDS as well as a drop in SpO_2 below 91% and no mention of a heart failure–related ICD code among prior diagnoses) among clusters as opposed to ARDS-2 and ARDS-4 (ARDS defined as having an ICD code for ARDS as well as a drop in SpO_2 below 96% and no mention of a heart failure–related ICD code among prior diagnoses). The clusters do not seem to be able to distinguish among ARDS symptoms when only the ICD code is used for ARDS prediction (as in ARDS-5). The mortality rate in cluster A is higher than that in the other 2 clusters, which is aligned with the fact that the rates of ARDS-1 and ARDS-3 are also higher in this cluster. Similar relative effects of SpO_2 <91% versus SpO_2 <96% can also be seen in hypoxemia targets. Table S4 in Multimedia Appendix 1 shows the distribution of the continuous input features among the 3 clusters. Cluster A shows more noticeable differences with the other 2 clusters when it comes to features such as systolic blood pressure, respiratory rate, neutrophils, lymphocytes, and SIRS, hinting at why the mortality is higher in this group. Note that targets 2, 5, and 9 are the most inclusive (general) of all targets; therefore, it is possible for one or more clusters to be completely enclosed by these targets, resulting in a rate of 100% (Figure 7). These inclusive targets are included to improve parameter sharing in the model for different outcomes.

Figure 6. Recurrent neural network representations separate into unique clusters. Clustering the population with acute respiratory distress syndrome (ARDS; n=1278) from the entire data set into 3 different groups A, B, and C by k-means clustering. The dimensions of the embedding vector were reduced using principal component analysis. PC1: principal component 1; PC2: principal component 2.

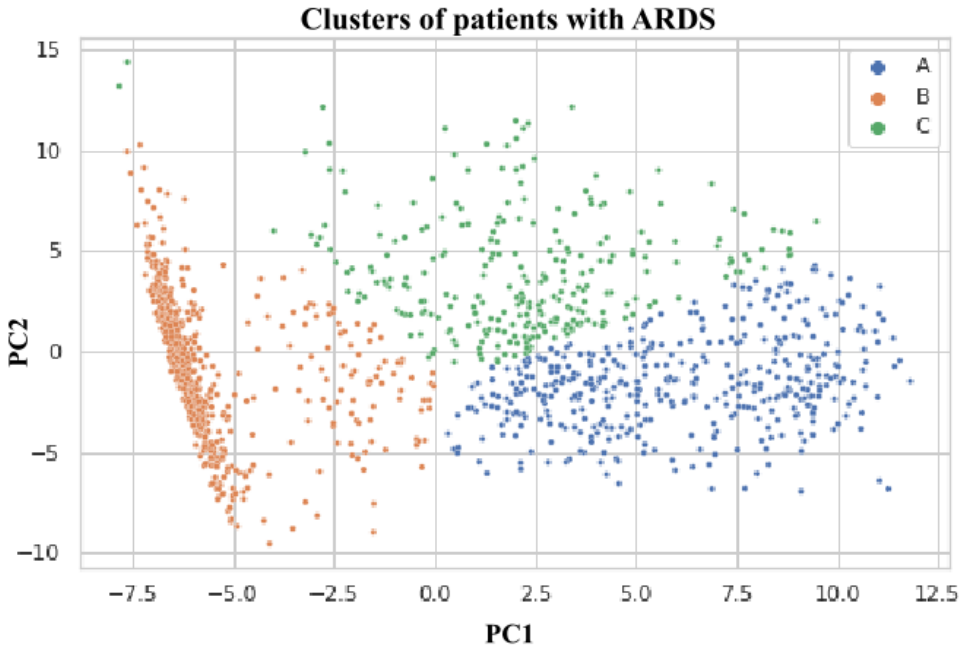
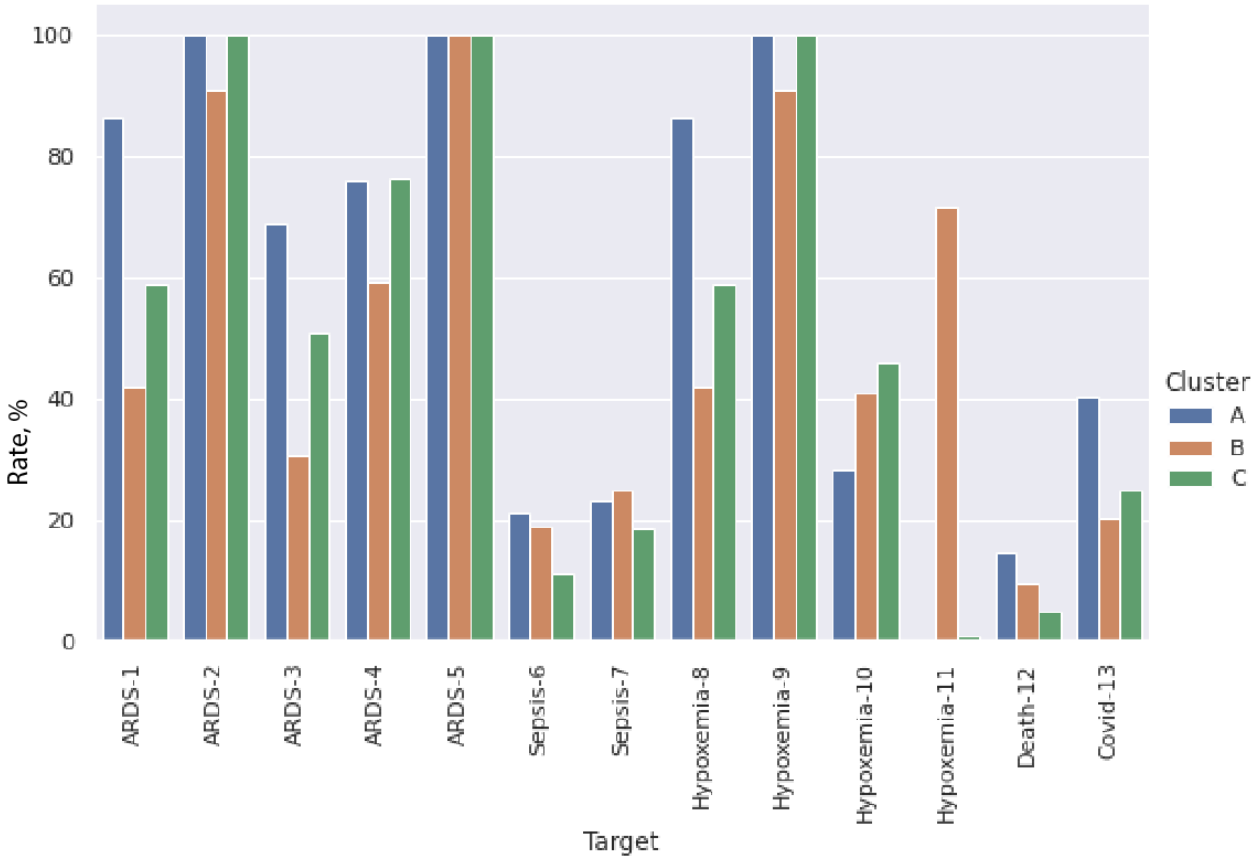


Figure 7. Incidence rates of different targets in each of the clusters. Target 12 is the mortality rate, which is 14.45%, 9.64%, and 4.73% for clusters A, B, and C, respectively. ARDS: acute respiratory distress syndrome.



Benefit Estimation

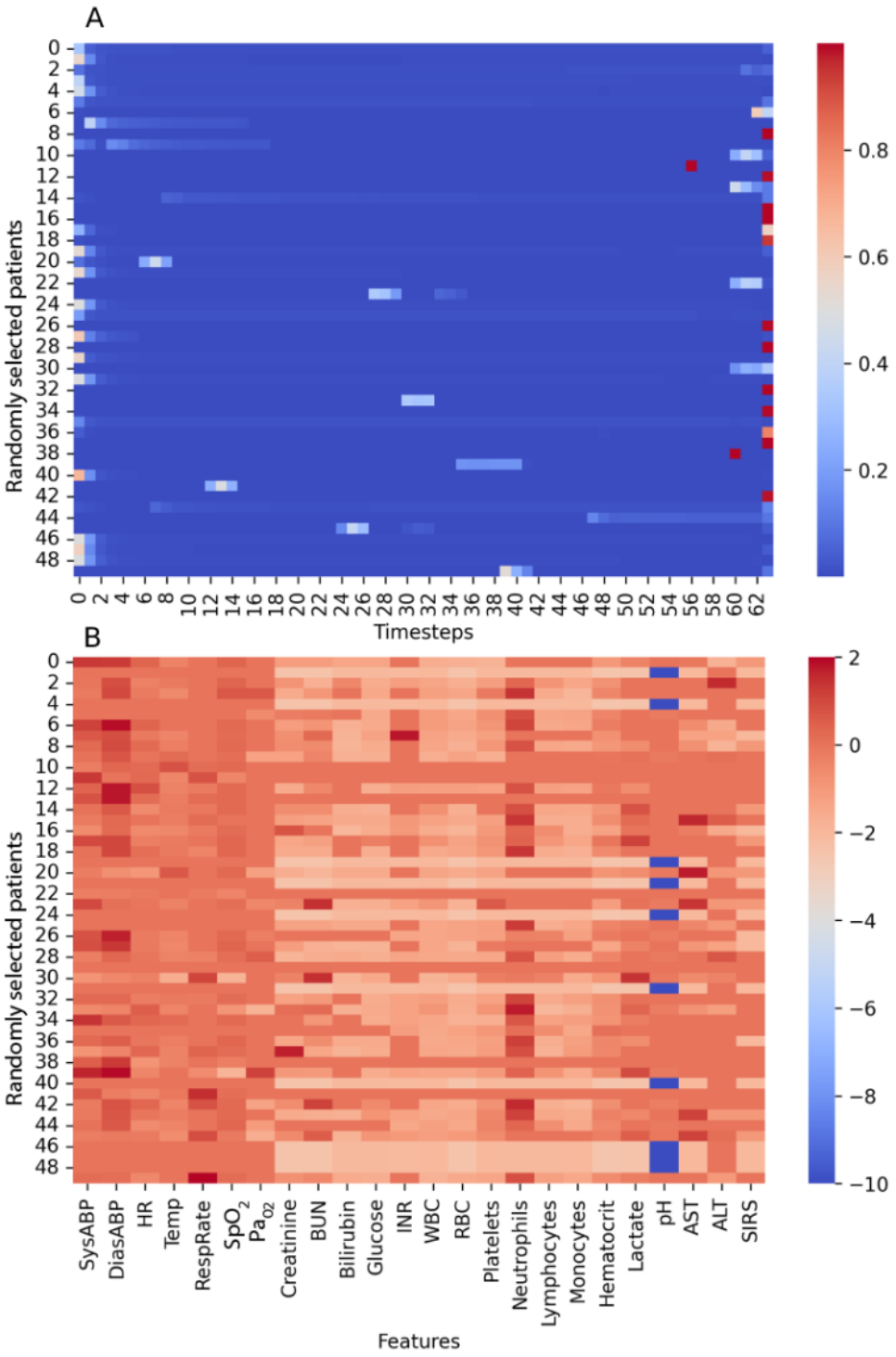
From our benefit estimation case study, we found that the mortality rate for patients who were diagnosed early with ARDS was 5.3% (14/266), whereas for those diagnosed late with ARDS, the mortality rate was 11.7% (116/995). The Fisher exact test statistic value was 0.002 for the 6.3% mortality benefit of using the algorithm's prediction for the early diagnosis group. For reference, the baseline mortality rate of patients without ARDS was 1.66% (656/39,442) and that of patients with ARDS was 10.31% (130/1261).

Model Interpretability

A visualization of the attention weights for different timesteps of the input sequence is shown in Figure 8A in which we observe variability in the distribution of the attention weights within the 64-timestep window. For some of the samples, the attention weights are higher toward the beginning of the sequence, whereas for others they are higher toward the end of the sequence. There are also cases where the attention weight is moderately higher in the middle of the sequence. The cases for which the attention weight is higher toward the end of the sequence represent the situation in which the most recent measurements with respect to the event of interest are more important. The cases for which the attention weights are higher toward the beginning of the sequence represent the situation in which the most relevant temporal data are near the beginning of, or before, the 64-timestep window. In this scenario, it is probably the attention network that is amplifying the signal from those early timesteps because without the attention network, GRUs alone will have a gradual decay of older timesteps. In the samples in which the attention weights are higher in the middle, there likely exists an intermediate timestep that has abnormal values that is emphasized more by the network. Figure 8B shows the calculated attention scores for specific features for the same patients as accounted for in Figure 8A. These scores were obtained by calculating the feature's z-score at the timestep with the highest attention weight. The figure is shown for all time-varying features. The figure reveals how every feature affects the model output. For example, high values of respiratory rate and low values of pH had a positive impact on the model.

From the SHAP force plots in Figure S2 in Multimedia Appendix 1, one can see the most influential features for a given patient. Red denotes the positive direction of influence on the ARDS-5 output, whereas blue denotes the negative direction of influence on the ARDS-5 output. The length of the arrow denotes the magnitude of SHAP values. The value in bold is the actual model output, which is then transformed into probability space to give the final output between 0 and 1. SpO_2 is an important feature, both to increase the probability of ARDS when SpO_2 is low (A) and to lower the probability of ARDS when SpO_2 is high (B). Low SpO_2 combined with high respiratory rate is the likely contributor to a false positive (C) in the context of a patient with poorly controlled diabetes (ie, blood glucose level=505 mg/dL and likely tachypnea of diabetic ketoacidosis) [40,41]. A normal SIRS score and normal neutrophil percentage in the absence of strongly positive features results in a false negative (D).

Figure 8. Attention heat maps. Each row along the y-axis is a patient. (A) Attention weights for timesteps. The heat map visualizes the attention weights on 50 randomly selected patients for the 64-timestep inputs to the recurrent neural network model. (B) Attention scores for features. The heat map visualizes the calculated z-score for every time-varying feature at the timestep with the greatest attention weight for the same set of patients as in part A. Red denotes a higher value or deviation in the positive direction; blue denotes a lower value or deviation in the negative direction. ALT: alanine transaminase; AST: aspartate aminotransferase; BUN: blood urea nitrogen; DiasABP: diastolic ambulatory blood pressure; HR: heart rate; INR: international normalized ratio; Pa_{O2}: arterial partial pressure of oxygen; RBC: red blood cell count; RespRate: respiratory rate; SIRS: systemic inflammatory response syndrome; SpO_2: peripheral oxygen saturation; SysABP: systolic ambulatory blood pressure; Temp: temperature; WBC: white blood cell count.



Discussion

Principal Findings

In this study, we described the development of a deep learning model for predicting multiple outcomes by simultaneously using the same set of input features. We showed that the RNN model trained to predict 13 outcomes simultaneously generalized better on ARDS outcomes than XGB models trained to predict individual outcomes. We showed that this improvement was proportional to the number of targets predicted by the RNN. This reinforces our conclusion that training the RNN model on a larger set of outcomes improves generalization. We hypothesize that multitask learning generalizes better in part because of parameter sharing, which has a regularizing effect, and information sharing across outcomes, which learns richer representations [42]. We would also like to emphasize that the intention of this paper was not to advance state-of-the-art multitask learning but to provide evidence that multitask learning is beneficial for early prediction of ARDS using only EHR data.

We used an RNN in this study because of its ability to use its internal memory to process variable length sequences of inputs, learn temporal dependence from the data, and share representations for an arbitrary number of outputs. We used a generic RNN with 4 GRU layers, an attention module, and 2 FC layers for all numbers of outputs. We experimented with various RNN architectures, varying the parameters such as the number of layers and hidden units. From our light grid search, we found that the RNN model architecture used in this paper performed best for our use case. In addition, the attention module seems to be an important part of the architecture in making the prediction because without it, the performance of the RNN dropped significantly for multiple targets as seen in Table S5 in Multimedia Appendix 1.

To compare the RNN with other algorithms, we used XGB because of its ability to handle missing or null values and its current dominance in industrial applications. As EHR data often have a high level of missing values because of variability in data acquisition and recording habits in the live clinical environment, this attribute of XGB is appealing. We trained multiple XGB models separately on the same input to classify different outcomes independently. We performed a grid search for hyperparameter optimization, tuning parameters such as tree depth and learning rate. Table S6 in Multimedia Appendix 1 shows the hyperparameters that were used for the grid search for the RNN and XGB models.

We also demonstrated an application of cluster analysis to probe deep learning models for clinical insights. Our analysis of the total population with ARDS uncovered 3 distinct populations, 2 of which have similarly high mortality rates but different clinical presentations. Recent studies have corroborated similar results in populations with COVID-19 in which 2 distinct phenotypes of ARDS were found with similar respiratory dynamics but 2-fold difference in odds of 28-day mortality [43]. With the methods outlined in this study, phenotype discovery would be an additional benefit that can be automatically applied to an arbitrarily large number of outcomes predicted.

To connect our machine learning findings with real-world clinical effects, we compared the mortality rates between patients diagnosed earlier with ARDS and patients diagnosed later with ARDS relative to the algorithm's prediction time. Our estimation showed that the mortality rate in the population diagnosed early with ARDS was almost half of that in the population diagnosed late with ARDS. Finally, to make our model more interpretable, we provide 2 heat maps attempting to visualize the attention score on each time series as well as 4 SHAP force plots presenting our case analyses regarding success and failure prediction.

Although we—and other researchers—have previously developed single-task machine learning models for predicting ARDS in different cohorts of hospitalized patients, to our knowledge, this is the first study to develop a multitask deep learning model for ARDS prediction [22-25]. Although previous studies have reported the development of high-performing ARDS prediction models, we intentionally do not make direct comparisons of model performance between our model and previous models for several reasons. The first is that to demonstrate that multitask learning improves performance over single-task learning, the models should ideally be trained and tested in a similar manner and on the same data sets. Comparisons with other published models may not provide any useful information on the direct benefit of using multitask learning models for ARDS prediction. Another point of consideration is that we used several subtypes of ARDS in our study; therefore, direct comparison against metrics from other studies that may use different ARDS definitions may not be fruitful.

Real-world clinical utility of such machine learning algorithms would need to be demonstrated through a multicenter prospective clinical study. We have previously developed and demonstrated the real-world impact of a sepsis prediction algorithm (InSight) on patient outcomes in a multicenter clinical validation study [44]. Although we performed retrospective validation on an external test set and demonstrated good performance of our algorithm in this study, ideally, the algorithm should be tested at multiple hospitals that vary by geographic location and patient demographic characteristics. Demonstrating a reduction in length of stay and improved outcomes of patients with ARDS through a clinical study would pave the way for deployment of the algorithm at medical institutions.

This study includes several limitations. In many hospital systems, radiology images and radiology reports are kept in a software system separate from the EHR. Ideally, we would prefer to confirm ARDS ICD codes by verifying the presence of bilateral lung infiltrates on chest imaging. Our inputs only included demographics, vital signs, and laboratory information. Future work should therefore incorporate EHR as well as imaging data. Our data set spans the emergency department, inpatient, and ICU settings and prescribes a single early time point for prediction. This could be a factor in the low AUROC for sepsis predictions, which prior studies have shown to be reliably accurate in the ICU setting [12,44]. This discrepancy warrants further investigation. In addition, we did not have reliable data on race and ethnicity of the patient population. Future studies would also benefit from training the models to

predict the additional output of respiratory support intervention beyond the level of a nonrebreather mask [45]. Finally, because this is a retrospective study, we are not able to determine the performance of our algorithm in a prospective clinical setting. Prospective testing is essential to determine how clinicians will respond to predictions of various outcomes. It is also important to determine whether our predictions can affect patient outcomes or resource allocation. Our work here is meant to serve as a reference for future research directions in establishing the most beneficial role for machine learning algorithms in the health care ecosystem and expanding the capabilities of machine learning in health care. Future research could also incorporate examining more state-of-the-art RNN architectures such as transformers that may have better performance for long sequence data processing.

Conclusions

We present a novel multitask deep learning model for predicting ARDS in hospitalized patients. Our results demonstrate that, based on the same input features, the higher the number of related outcomes predicted by our model, the better the performance on most outcomes. We demonstrate the clinical utility of our model by calculating the sensitivity, specificity, and AUROC of various iterations of the model on 2 external test sets and explore the interpretability of our model by visualizing attention weights using heat maps and SHAP for global and local model interpretability. Early prediction of ARDS, together with the stratification of patients into different subgroups based on different clinical presentations, will enable clinicians to take appropriate action to prevent the deterioration of a patient's condition, which should in turn improve patient outcomes and mortality or morbidity rates of ARDS.

Acknowledgments

The authors wish to thank Anna Siefkas for discussion, comments, and edits on this manuscript. The authors also wish to thank Zohora Iqbal, Gina Barnes, and Abigail Green-Saxena for comments and edits on this manuscript.

Authors' Contributions

CL contributed to data processing, designing the machine learning models and experiments, and drafting the paper. RT contributed to running machine learning experiments, designing clustering algorithms, generating plots and diagrams, and drafting the paper. JM contributed to running machine learning experiments, architecture search, analysis of results, generating plots, and drafting the paper. KR contributed to designing and analyzing clustering algorithms, evaluating machine learning algorithms and Shapley additive explanations, and drafting the paper. CFT contributed to data collection, analysis, and drafting the paper. SCC contributed to statistical analysis, literature review, and drafting the paper. QM contributed to experimental design, data collection, and drafting the paper. NPS contributed to experimental design, data collection, and drafting the paper.

Conflicts of Interest

All authors are employees of Dascena, Inc. CL, CFT, and QM have shares in Dascena, Inc. The company specializes in building machine learning algorithms, one of which is InSight, the sepsis prediction algorithm referenced in this paper.

Multimedia Appendix 1
Supplementary tables and figures.
[DOCX File , 405 KB - medinform_v10i6e36202_app1.docx]

References

1. Matthay MA, Zemans RL, Zimmerman GA, Arabi YM, Beitler JR, Mercat A, et al. Acute respiratory distress syndrome. Nat Rev Dis Primers 2019 Mar 14;5(1):18 [FREE Full text] [doi: 10.1038/s41572-019-0069-0] [Medline: 30872586]
2. Sweeney RM, McAuley DF. Acute respiratory distress syndrome. Lancet 2016 Nov;388(10058):2416-2430. [doi: 10.1016/s0140-6736(16)00578-x]
3. Ferguson ND, Fan E, Camporota L, Antonelli M, Anzueto A, Beale R, et al. The Berlin definition of ARDS: an expanded rationale, justification, and supplementary material. Intensive Care Med 2012 Oct;38(10):1573-1582. [doi: 10.1007/s00134-012-2682-1] [Medline: 22926653]
4. Riviello ED, Kiviri W, Twagirumugabe T, Mueller A, Banner-Goodspeed VM, Officer L, et al. Hospital incidence and outcomes of the acute respiratory distress syndrome using the kigali modification of the Berlin definition. Am J Respir Crit Care Med 2016 Jan;193(1):52-59. [doi: 10.1164/rccm.201503-0584oc]
5. Gajic O, Dabbagh O, Park PK, Adesanya A, Chang SY, Hou P, et al. Early identification of patients at risk of acute lung injury. Am J Respir Crit Care Med 2011 Feb 15;183(4):462-470. [doi: 10.1164/rccm.201004-0549oc]
6. Soto GJ, Kor DJ, Park PK, Hou PC, Kaufman DA, Kim M, et al. Lung injury prediction score in hospitalized patients at risk of acute respiratory distress syndrome. Crit Care Med 2016;44(12):2182-2191. [doi: 10.1097/ccm.0000000000002001]
7. Levitt J, Calfee C, Goldstein BA, Vojnik R, Matthay MA. Early acute lung injury. Crit Care Med 2013;41(8):1929-1937. [doi: 10.1097/ccm.0b013e31828a3d99]

8. Seitz K, Caldwell E, Hough CL. Fluid management in ARDS: an evaluation of current practice and the association between early diuretic use and hospital mortality. J Intensive Care 2020;8:78 [FREE Full text] [doi: 10.1186/s40560-020-00496-7] [Medline: 33062283]
9. Malhotra A. Low-tidal-volume ventilation in the acute respiratory distress syndrome. N Engl J Med 2007 Sep 13;357(11):1113-1120. [doi: 10.1056/nejmct074213]
10. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan 7;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
11. Radhachandran A, Garikipati A, Zelin NS, Pellegrini E, Ghandian S, Calvert J, et al. Prediction of short-term mortality in acute heart failure patients using minimal electronic health record data. BioData Min 2021 Mar 31;14(1):23 [FREE Full text] [doi: 10.1186/s13040-021-00255-w] [Medline: 33789700]
12. Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open 2018 Jan 26;8(1):e017833 [FREE Full text] [doi: 10.1136/bmjopen-2017-017833] [Medline: 29374661]
13. Mohamadlou H, Lynn-Palevsky A, Barton C, Chettipally U, Shieh L, Calvert J, et al. Prediction of acute kidney injury with a machine learning algorithm using electronic health record data. Can J Kidney Health Dis 2018 Jun 08;5:2054358118776326 [FREE Full text] [doi: 10.1177/2054358118776326] [Medline: 30094049]
14. Ryan L, Mataraso S, Siefkas A, Pellegrini E, Barnes G, Green-Saxena A, et al. A machine learning approach to predict deep venous thrombosis among hospitalized patients. Clin Appl Thromb Hemost 2021 Feb 24;27:1076029621991185 [FREE Full text] [doi: 10.1177/1076029621991185] [Medline: 33625875]
15. Giang C, Calvert J, Rahmani K, Barnes G, Siefkas A, Green-Saxena A, et al. Predicting ventilator-associated pneumonia with machine learning. Medicine (Baltimore) 2021 Jun 11;100(23):e26246 [FREE Full text] [doi: 10.1097/MD.0000000000026246] [Medline: 34115013]
16. Rahmani K, Garikipati A, Barnes G, Hoffman J, Calvert J, Mao Q, et al. Early prediction of central line associated bloodstream infection using machine learning. Am J Infect Control 2022 Apr;50(4):440-445 [FREE Full text] [doi: 10.1016/j.ajic.2021.08.017] [Medline: 34428529]
17. Coudroy R, Frat J, Boissier F, Contou D, Robert R, Thille AW. Early identification of acute respiratory distress syndrome in the absence of positive pressure ventilation. Crit Care Med 2018;46(4):540-546. [doi: 10.1097/ccm.0000000000002929]
18. ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, et al. Acute respiratory distress syndrome: the Berlin definition. JAMA 2012 Jun 20;307(23):2526-2533. [doi: 10.1001/jama.2012.5669] [Medline: 22797452]
19. Maxwell A, Li R, Yang B, Weng H, Ou A, Hong H, et al. Deep learning architectures for multi-label classification of intelligent health risk prediction. BMC Bioinformatics 2017 Dec 28;18(Suppl 14):523 [FREE Full text] [doi: 10.1186/s12859-017-1898-z] [Medline: 29297288]
20. Zhang X, Zhao H, Zhang S, Li R. A novel deep neural network model for multi-label chronic disease prediction. Front Genet 2019;10:351 [FREE Full text] [doi: 10.3389/fgene.2019.00351] [Medline: 31068968]
21. Lipton Z, Kale D, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. In: Proceedings of the 4th International Conference on Learning Representations, ICLR 2016. 2016 Presented at: 4th International Conference on Learning Representations, ICLR 2016; May 2-4, 2016; San Juan, Puerto Rico. [doi: 10.48550/arXiv.1511.03677]
22. Zeiberg D, Prahlad T, Nallamothu BK, Iwashyna TJ, Wiens J, Sjoding MW. Machine learning for patient risk stratification for acute respiratory distress syndrome. PLoS One 2019 Mar 28;14(3):e0214465 [FREE Full text] [doi: 10.1371/journal.pone.0214465] [Medline: 30921400]
23. Singhal L, Garg Y, Yang P, Tabaie A, Wong AI, Mohammed A, et al. eARDS: a multi-center validation of an interpretable machine learning algorithm of early onset Acute Respiratory Distress Syndrome (ARDS) among critically ill adults with COVID-19. PLoS One 2021;16(9):e0257056 [FREE Full text] [doi: 10.1371/journal.pone.0257056] [Medline: 34559819]
24. Lam C, Tso C, Green-Saxena A, Pellegrini E, Iqbal Z, Evans D, et al. Semisupervised deep learning techniques for predicting acute respiratory distress syndrome from time-series clinical data: model development and validation study. JMIR Form Res 2021 Sep 14;5(9):e28028 [FREE Full text] [doi: 10.2196/28028] [Medline: 34398784]
25. Le S, Pellegrini E, Green-Saxena A, Summers C, Hoffman J, Calvert J, et al. Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS). J Crit Care 2020 Dec;60:96-102 [FREE Full text] [doi: 10.1016/j.jcrc.2020.07.019] [Medline: 32777759]
26. Read J, Perez-Cruz F. Deep learning for multi-label classification. arXiv. Preprint posted online on December 17, 2014. [FREE Full text] [doi: 10.4018/978-1-4666-5202-6.ch142]
27. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, et al. Automatic differentiation in PyTorch. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); Dec 4 - 9, 2017; Long Beach, CA, USA URL: https://openreview.net/pdf?id=BJJsrmfCZ&source=post_page
28. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: Proceedings of the NIPS 2014 Workshop on Deep Learning. 2014 Presented at: NIPS 2014 Workshop on Deep Learning; Dec 13, 2014; Montreal, Canada.

29. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv. Preprint posted online on September 1, 2014. [FREE Full text] [doi: 10.48550/arXiv.1409.0473]
30. Good I. Some terminology and notation in information theory. Proc IEE C Monogr UK 1956 Mar;103(3):200-204. [doi: 10.1049/pi-c.1956.0024]
31. Zhang Y, Yang Q. A survey on multi-task learning. IEEE Trans Knowl Data Eng 2021 Mar 31:1. [doi: 10.1109/tkde.2021.3070203]
32. Ruder S. An overview of multi-task learning in deep neural networks. arXiv 2017. [doi: 10.48550/arXiv.1706.05098]
33. Ridnik T, Ben-Baruch E, Zamir N, Noy A, Friedman I, Protter M, et al. Asymmetric loss for multi-label classification. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021 Presented at: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); Oct 10-17, 2021; Montreal, QC, Canada. [doi: 10.1109/iccv48922.2021.00015]
34. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020 Jan;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
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. 2016 Presented at: KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13 - 17, 2016; San Francisco California USA. [doi: 10.1145/2939672.2939785]
36. XGBoost Python Package. dmlc XGBoost. URL: https://xgboost.readthedocs.io/en/stable/python/index.html [accessed 2021-01-19]
37. Lundberg S, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: 31st International Conference on Neural Information Processing Systems; Dec 4 - 9, 2017; Long Beach California USA. [doi: 10.5555/3295222.3295230]
38. Bertail P, Clémençcon S, Vayatis N. On Bootstrapping the ROC curve. In: Proceedings of the 21st International Conference on Neural Information Processing Systems. 2008 Presented at: 21st International Conference on Neural Information Processing Systems; Dec 8 - 10, 2008; Vancouver British Columbia Canada URL: https://dl.acm.org/doi/10.5555/2981780.2981798 [doi: 10.5555/2981780.2981798]
39. Liu H, Li G, Cumberland W, Wu T. Testing statistical significance of the area under a receiving operating characteristics curve for repeated measures design with bootstrapping. J Data Sci 2005;3(3):257-278 [FREE Full text] [doi: 10.6339/JDS.2005.03(3).206]
40. Gallo de Moraes A, Surani S. Effects of diabetic ketoacidosis in the respiratory system. World J Diabetes 2019 Jan 15;10(1):16-22 [FREE Full text] [doi: 10.4239/wjd.v10.i1.16] [Medline: 30697367]
41. Westerberg DP. Diabetic ketoacidosis: evaluation and treatment. Am Fam Physician 2013 Mar 01;87(5):337-346 [FREE Full text] [Medline: 23547550]
42. Nam J, Loza Mencía E, Kim H, Fürnkranz J. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In: Proceedings of theAdvances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. 2017 Presented at: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017; Dec 4-9, 2017; Long Beach, CA.
43. Ranjeva S, Pinciroli R, Hodell E, Mueller A, Hardin CC, Thompson BT, et al. Identifying clinical and biochemical phenotypes in acute respiratory distress syndrome secondary to coronavirus disease-2019. EClinicalMedicine 2021 Apr;34:100829 [FREE Full text] [doi: 10.1016/j.eclinm.2021.100829] [Medline: 33875978]
44. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR Med Inform 2016 Sep 30;4(3):e28 [FREE Full text] [doi: 10.2196/medinform.5909] [Medline: 27694098]
45. Wong A, Cheung P, Kamaleswaran R, Martin GS, Holder AL. Machine learning methods to predict acute respiratory failure and acute respiratory distress syndrome. Front Big Data 2020;3:579774 [FREE Full text] [doi: 10.3389/fdata.2020.579774] [Medline: 33693419]

Abbreviations

ARDS: acute respiratory distress syndrome
AUROC: area under the receiver operating characteristic curve
EHR: electronic health record
FC: fully connected
GRU: gated recurrent unit
ICD: International Classification of Diseases
ICU: intensive care unit
RNN: recurrent neural network
SHAP: Shapley additive explanations

SIRS: systemic inflammatory response syndrome
SpO_2: peripheral oxygen saturation
XGB: XGBoost

Edited by C Lovis; submitted 06.01.22; peer-reviewed by M Aczon, J Walsh, G Liu, Y Xie, B Puladi, A Bayani; comments to author 13.02.22; revised version received 07.04.22; accepted 02.05.22; published 15.06.22.

Please cite as:
Lam C, Thapa R, Maharjan J, Rahmani K, Tso CF, Singh NP, Casie Chetty S, Mao Q
Multitask Learning With Recurrent Neural Networks for Acute Respiratory Distress Syndrome Prediction Using Only Electronic Health Record Data: Model Development and Validation Study
JMIR Med Inform 2022;10(6):e36202
URL: https://medinform.jmir.org/2022/6/e36202
doi:10.2196/36202
PMID:35704370

Original Paper

Identifying the Risk of Sepsis in Patients With Cancer Using Digital Health Care Records: Machine Learning–Based Approach

Donghun Yang[1,2*], MSc; Jimin Kim[3*], PhD; Junsang Yoo[4], PhD; Won Chul Cha[4,5], MD, PhD; Hyojung Paik[2,3], PhD

[1]AI Technology Research Center, Division of S&T Digital Convergence, Korea Institute of Science and Technology Information, Daejeon, Republic of Korea

[2]Department of Data and High Performance Computing Science, University of Science and Technology, Daejeon, Republic of Korea

[3]Center for Supercomputing Applications, Division of National Supercomputing, Korea Institute of Science and Technology Information, Daejeon, Republic of Korea

[4]Department of Digital Health, Samsung Advanced Institute for Health Science & Technology, Sungkyunkwan University, Seoul, Republic of Korea

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

[*]these authors contributed equally

Corresponding Author:
Hyojung Paik, PhD
Center for Supercomputing Applications
Division of National Supercomputing
Korea Institute of Science and Technology Information
245 Daehak-ro
Yuseong-Gu
Daejeon, 34141
Republic of Korea
Phone: 82 428690791
Email: hyojungpaik@gmail.com

Abstract

Background: Sepsis is diagnosed in millions of people every year, resulting in a high mortality rate. Although patients with sepsis present multimorbid conditions, including cancer, sepsis predictions have mainly focused on patients with severe injuries.

Objective: In this paper, we present a machine learning–based approach to identify the risk of sepsis in patients with cancer using electronic health records (EHRs).

Methods: We utilized deidentified anonymized EHRs of 8580 patients with cancer from the Samsung Medical Center in Korea in a longitudinal manner between 2014 and 2019. To build a prediction model based on physical status that would differ between sepsis and nonsepsis patients, we analyzed 2462 laboratory test results and 2266 medication prescriptions using graph network and statistical analyses. The medication relationships and lab test results from each analysis were used as additional learning features to train our predictive model.

Results: Patients with sepsis showed differential medication trajectories and physical status. For example, in the network-based analysis, narcotic analgesics were prescribed more often in the sepsis group, along with other drugs. Likewise, 35 types of lab tests, including albumin, globulin, and prothrombin time, showed significantly different distributions between sepsis and nonsepsis patients ($P<.001$). Our model outperformed the model trained using only common EHRs, showing an improved accuracy, area under the receiver operating characteristic (AUROC), and F1 score by 11.9%, 11.3%, and 13.6%, respectively. For the random forest–based model, the accuracy, AUROC, and F1 score were 0.692, 0.753, and 0.602, respectively.

Conclusions: We showed that lab tests and medication relationships can be used as efficient features for predicting sepsis in patients with cancer. Consequently, identifying the risk of sepsis in patients with cancer using EHRs and machine learning is feasible.

(JMIR Med Inform 2022;10(6):e37689) doi:10.2196/37689

KEYWORDS

sepsis; cancer; EHR; machine learning; deep learning; mortality rate; learning model; electronic health record; network based analysis; sepsis risk; risk model; prediction model

Introduction

Sepsis is a life-threatening organ dysfunction in which a pathogen infection leads to a dysregulated host response to the infection [1]. Sepsis is diagnosed in millions of people every year globally, accounting for a high ratio of in-hospital mortality (25%-50%) [2]. In particular, the mortality rate increases dramatically when septic shock is established [3,4]. Although a timely diagnosis of sepsis is essential for a promising prognosis, only minor cold-like symptoms, such as fever, excessive breathing, and increased pulse rate, are presented in the early stage of sepsis [5]. Therefore, in hospitals, patients admitted to the ward may suffer from septic shock after clinicians have missed the signature symptoms of sepsis. Thus, it is important to stratify high-risk patients and provide appropriate treatment in a short amount of time [6].

Sepsis has shown a substantial incidence in patients with low immunity, such as patients with cancer, patients who are elderly, and newborns [7]. Patients with cancer are at high risk for sepsis, as many are immunosuppressed due to the cancer itself and chemotherapy treatment [8]. For example, leukocyte counts are lowered, especially when anticancer treatments decrease bone marrow function, suppressing immune response to the pathogen [9]. Although predicting sepsis in patients with cancer is essential, an early identification of the risk of sepsis remains an unmet medical need.

Various studies have been conducted to identify the risk of sepsis, including a statistical model–based approach for emergency room (ER) patients [10], a machine learning–based approach for inpatients [11], and an approach using unstructured clinical data [12]. The majority of previous studies have focused on patients with severe trauma in the intensive care unit (ICU). However, the stratification of sepsis risk among patients with cancer has scarcely been conducted.

Our study aimed to predict the risk of sepsis in patients with cancer at an early stage using clinical information and a machine learning approach. We utilized the deidentified electronic health records (EHRs) from the Samsung Medical Center (SMC) in Korea of 8580 patients with cancer, including inpatients, outpatients, ICU patients, and ER patients. Drug prescriptions and laboratory test results are known to reflect the physical status of patients [13]. In our previous study, we showed that distributions of lab test results recapitulate the physical states of patients, including disease signatures and drug-associated responses [14]. Prescriptions of medications for cancer are mainly determined based on the patients' medical conditions. Thus, we hypothesized that the patterns of prescribed medications and lab test results would be different between the sepsis and nonsepsis groups. To validate our hypothesis, we analyzed 2462 lab test results and 2266 medication prescriptions using network-based association rule [15] analysis and statistical analysis.

Based on the results of the analyses, we propose a machine learning–based sepsis predictive model that can reflect the physical conditions of patients with cancer and is trained on the prescribed drug and lab test patterns as well as EHRs, which are widely used in the reported sepsis prediction approaches [16,17].

Methods

Study Sample

Data were prepared from the Clinical Data Warehouse (CDW) and the SMC cancer registry, Seoul, South Korea, and deidentification was performed on the collected data. The study population included adult patients diagnosed with lung, liver, and breast cancer who visited the ER within 5 years of being diagnosed with cancer. The inclusion criteria were patients with cancer registered at the study sites. Patients were excluded from the study cohort if they met the following exclusion criteria: those under 18 years of age, those with multiple cancers, those who had not visited the emergency room within 5 years after the first cancer diagnosis, and those with ICD-10 codes not matched with C22, C34, and C50. The data were constructed by reflecting various EHR information such as hospitalization data, diagnosis code of cancer or other underlying disease, vital signs, genomic information, medication prescription, surgical history, radiation treatment, and lab test information for 5 years (2014-2019) before and after the cancer diagnosis of 8580 patients with cancer, including inpatients, outpatients, ICU patients, and ER patients. Most of the currently published sepsis prediction models use information within 48 hours before the onset of sepsis. However, due to the high risk of sepsis, it was considered necessary to predict in advance, so information 2 days prior to the ER visit was used. Data earlier than 7 days were somewhat difficult to consider as having an effect on the onset of sepsis, so the filtration criteria was set to 2-7 days.

Ethics Approval

The institutional ethics committee of SMC approved this study (Institutional Review Board File #2019-06-071).

Identifying Patients With Sepsis

We identified patients with sepsis using the Sequential Organ Failure Assessment (SOFA) scores of Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) guidelines [18] for a total of 18,610 ER visits by 8580 patients with cancer using the following procedures:

1. Nursing records, inspection records, clinical information, and medication prescription data were extracted from the CDW.
2. The variables were preprocessed to obtain the SOFA scores.
3. SOFA scores for each patient were calculated each time.
4. The time window was set by checking whether antibiotics were administered intravenously within 24 hours before and after the bacterial culture test.
5. Patients with sepsis were identified if their SOFA score changed by 2 points or more within the time window.
6. In accordance with the Sepsis-3 guidelines, if the SOFA score could not be measured in advance, it was considered 0. Consequently, if the change in the SOFA score was 2 or higher in the first visit to the ER, the patient was considered to be experiencing sepsis.

Data Filtering and Preprocessing

We aligned the collected EHRs of 8580 patients with cancer based on the date of the ER visit and filtered patients with information 2-7 days prior to the ER visit. Each patient's diagnostic code was recorded as an ICD-9 code and standardized to 3 digits for use as a categorical feature. Because there was a possibility of information leakage from giving hints to the machine learning–based predictive model, lab test results centered on specific disease groups were removed, and only the lab test information performed on over 60% of the patients was used. All categorical features were preprocessed using one-hot encoding, and all binary categorical features were encoded as 0 or 1. In addition, missing values were imputed with the mean value of patients with the same type of cancer, the same sex, and the same age, and extreme outlier values were removed.

Graph Network–Based Association Rule Analysis

Graph network–based association rules were performed on 2266 drug prescriptions. An association rule is a method for discovering frequent patterns and relationships between items from complicated data and can be employed to conceptualize complex dynamic systems comprising each interacting event [15]. Using the frequent pattern growth (FP-growth) algorithm [19], frequent relationships of drugs prescribed on the same day were analyzed. Next, only the group sets with a minimum support value of 0.05 or greater were selected. The support value ($S(D_i \rightarrow D_j)$), defined as in Equation (1), implied how often the sets go together when items are being tied up simultaneously, where $N(s)$ represents the total number of prescriptions, and $N(D_i,D_j)$ represents the number of events in which the *i-th* and *j-th* drugs were prescribed on the same day.

Finally, after designating each selected drug as a node, we plotted a graph network to visualize the result of the association rule analysis. The edges depicted the correlations of each drug.

Vectorization of Prescribed Medication Relationships

We vectorized the relationships found through graph network–based association rule analysis to be used as an input for the machine learning–based sepsis prediction model. After multiplying each one-hot encoded drug selected through the aforementioned analysis by the number of prescription days, the relationship for each pair of values was vectorized using the 3 formulas proposed in our previous study [20]. These 3 formulas ($r(I, H, T)$) comprised the interaction (I), the harmonized average (H), and the arctangent (T), in which (I) determined the level of interaction, (H) determined the overall intensity in a sensitive manner, and (T) determined the geometric angle difference as a single scalar value for each pair, defined as in Equation (2), where $D_i^{(p,s)}$ and $D_j^{(p,s)}$ indicate the *i-th* and *j-th* drug of the *s-th* prescriptions for the *p-th* patient, respectively. The D value represents the prescription frequency of each medication.

Prediction of Sepsis Using Machine Learning Approaches

We trained models on vectorized drug relationships and selected lab test types, along with the common EHRs that are widely used in the reported sepsis prediction models [16,17]. We considered 2 machine learning models comprising logistic regression (LR) [21] and random forest [22] and 3 deep learning models comprising artificial neural networks (ANNs) [23], residual convolutional neural networks (ResNet10) [24], and long short-term memory recurrent neural networks (RNN-LSTMs) [25]. When applied to the model, the data were reshaped to (1, 42, 42) for the ResNet10 model and padded to the maximum length of the sequence and reshaped to (number of patients, time sequence, number of features) for the LSTM model. We investigated the important features using Shapley Additive Explanations (SHAP) [26]. SHAP, one of the Explainable Artificial Intelligence (XAI) techniques, is a method used to interpret results from deep learning and machine learning models and is based on game theory. We used Tree SHAP explainer to calculate the Shapley values.

All proposed approaches were implemented using the Python 3.7 library, such as PyTorch 1.5, Scikit-learn, and SHAP, on an NVIDIA TITAN RTX 24 GB × 2. The source code is available on GitHub [27].

Results

Characteristics of the Filtered and Preprocessed Data Set From SMC

The overall process of our study is shown in Figure 1. We analyzed data from 8580 patients obtained from the CDW of SMC. Using the SOFA scores of the Sepsis-3 guidelines, of a total of 18,610 ER visits by 8580 patients with cancer, 2960 visits were identified as sepsis and 15,650 visits as nonsepsis. As a result of filtering the patients, the control group included 928 patients, and the sepsis group included 455 patients. The statistics of the filtered and preprocessed data set that was used to build the sepsis predictive model are shown in Table 1.

In the control group (ie, nonsepsis patients with cancer), there were 490 (52.8%) males and 438 (47.2%) females. The mean age was 58.2 (SD 11.0) years, and the average weight was 63.7 (SD 10.7) kg. In terms of the initial cancer diagnosis of each patient, 180 (19.4%) had liver cancer, 533 (57.4%) had lung cancer, and 215 (23.2%) had breast cancer. Meanwhile, in the sepsis group, there were 324 (71.2%) males and 131 (28.8%) females, with a relatively higher proportion of males than the control group. The mean age of the sepsis group was 60.3 (SD 0.5) years, and the average weight was 64.3 (SD 11.3) kg. In the sepsis group, 140 (30.8%) patients had liver cancer, 274 (60.2%) had lung cancer, and 41 (9%) had breast cancer. With these prepared data sets from SMC, we analyzed the differences in medication patterns by group.

Figure 1. Study overview. CDW: Clinical Data Warehouse; EHR: electronic health record; ER: emergency room; ER visits: total number of ER visits by the patients; SOFA: Sequential Organ Failure Assessment of the Sepsis-3 guidelines; SMC: Samsung Medical Center.

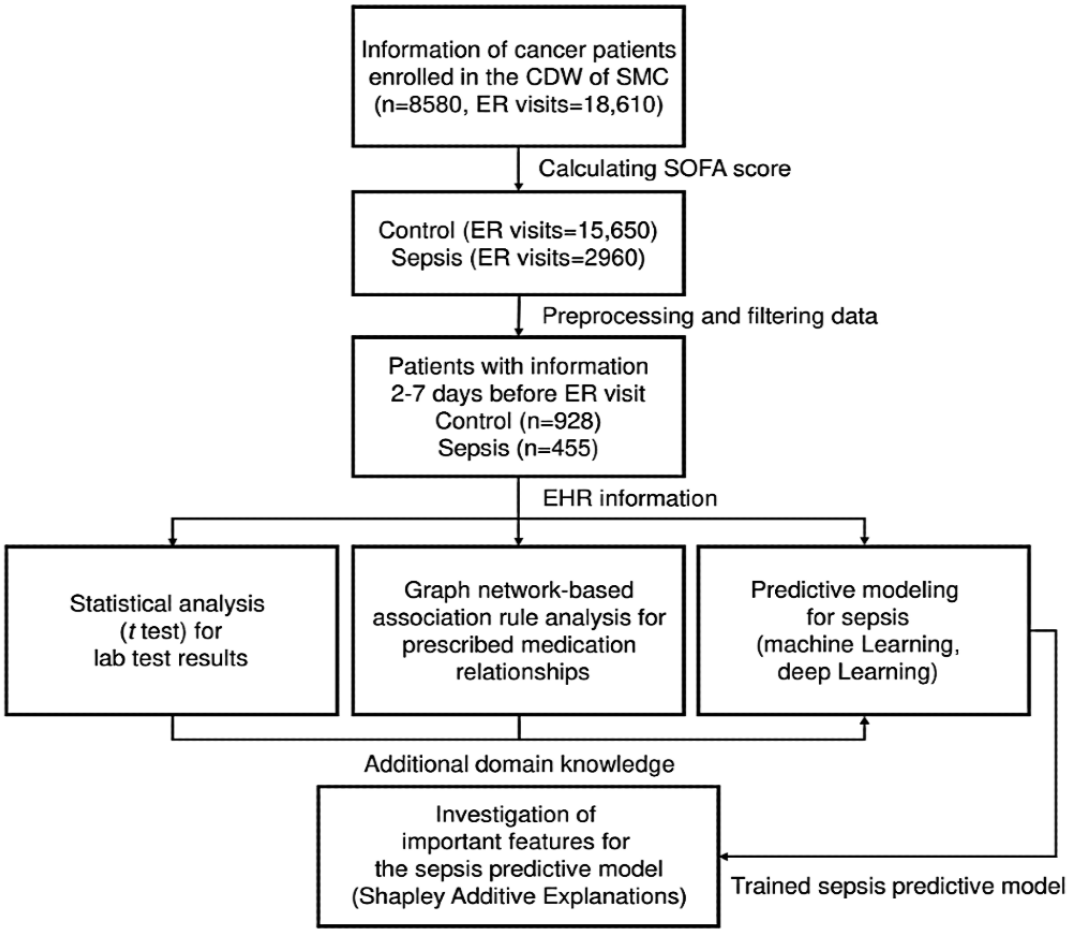


Table 1. Statistics of the input data used to build the sepsis predictive model.

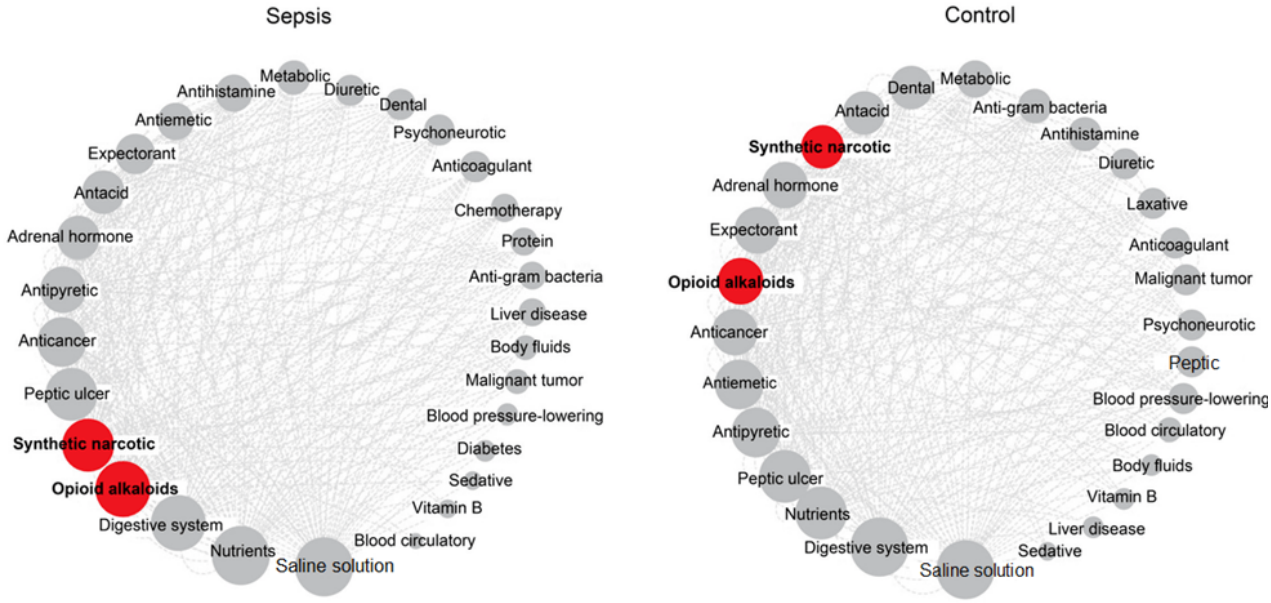
Patient characteristics	Total (N=1383)	Control group (n=928)	Sepsis group (n=455)
Sex, n (%)			
Male	814 (58.9)	490 (52.8)	324 (71.2)
Female	569 (41.1)	438 (47.2)	131 (28.8)
Age (years), mean (SD)	58.9 (10.9)	58.2 (11)	60.3 (0.5)
Weight (kg), mean (SD)	63.9 (0.9)	63.7 (10.7)	64.3 (11.3)
Cancer, n (%)			
Liver	320 (23.1)	180 (19.4)	140 (30.8)
Lung	807 (58.4)	533 (57.4)	274 (60.2)
Breast	256 (18.5)	215 (23.2)	41 (9.0)
Emergency room visits, n (%)	1466 (100)	991 (68)	475 (32)

Graph Network–Based Association Analysis for Prescribed Medications

Using the FP-growth algorithm, we analyzed patterns of the medications prescribed on the same day in 2666 prescriptions from the preprocessed and filtered EHR data. According to the analysis results, only group sets with a minimum support value of 0.05 or greater were selected. Of a total of 101 different drug types, 406 relationships among 29 drugs and 378 relationships among 28 drugs were selected for the sepsis group and nonsepsis group, respectively. To visualize the associations between the drug prescriptions, we constructed 2 graph networks with nodes representing the selected drugs and edges depicting the relationships among the nodes (Figure 2). The size of a node was determined by its average shortest path distance (Multimedia Appendix 1, graph A) and the number of edges (Multimedia Appendix 1, graph B), representing the topological properties of the network. A larger node meant that the corresponding drug was prescribed more often with other drugs compared to small nodes.

The patterns of medications between the sepsis and control groups were different. Nodes for medications such as "saline solution," which are commonly prescribed for most patients, showed similar patterns in both networks (Figure 2), whereas "opioid alkaloids" and "synthetic narcotic" nodes were ranked higher in the sepsis group than in the control group (Multimedia Appendix 1, graphs A and B). These kinds of narcotic analgesic nodes were the bottleneck and central nodes, meaning that they were prescribed more often with other drugs in the sepsis group. Thus, we were able to confirm our hypothesis that relationships and patterns of prescribed medications were distinct in the 2 groups.

Figure 2. Results of the graph network-based association analysis for prescribed medications. Graph networks for sepsis (n=29) and control (n=28) patients. Each node represents the medication selected through association rule analysis (support value ≥ 0.05), and edges depict that the linked nodes were prescribed together. Red nodes and bars represent the drugs with different patterns between the sepsis and control patients.

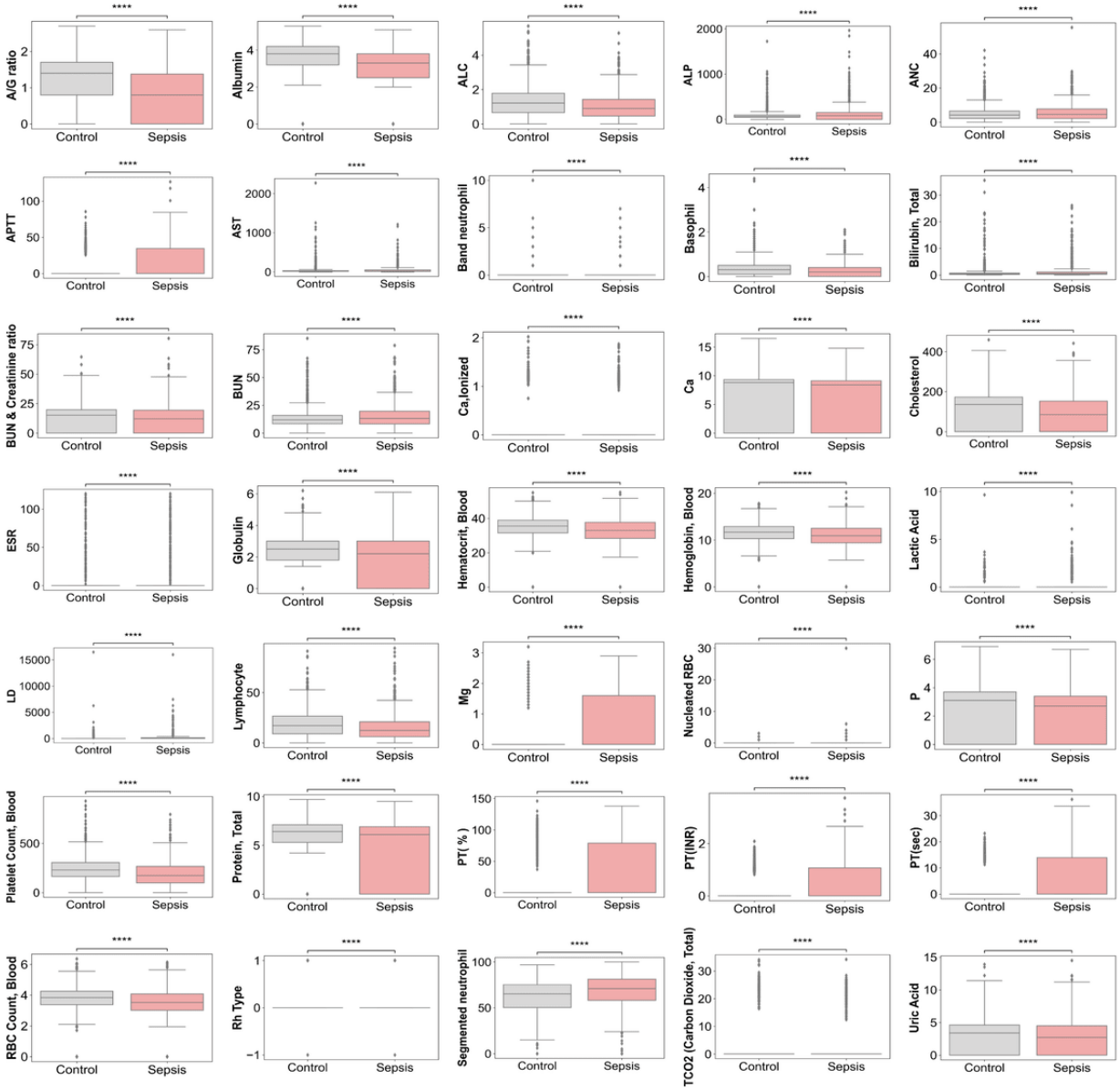


Statistical Analysis for Lab Tests

Using a *t* test, we analyzed the 2462 lab test results of the preprocessed and filtered EHR data to find lab test items with significantly different distributions between the sepsis and nonsepsis groups. Multimedia Appendix 2 presents the means and standard deviations of the 2 groups for all lab test types, where the *P* is symbolized (no significance: NS, *P*<.05: *, *P*<.001: ***, *P*<.005: **, *P*<.001: ****).

Figure 3 presents the distributions of the 2 groups for our selected lab test types, including predictors that are well-known hallmarks of sepsis. The changes in albumin, total protein, and cholesterol levels reflected the higher risk of mortality in patients with sepsis [28], and in our results, the values of these factors were significantly lower in patients with sepsis, with an albumin level of 2.73 (SD 1.57) versus 3.3 (SD 1.47), total protein of 4.77 (SD 2.98) versus 5.27 (SD 2.82), and a cholesterol level of 81.06 (SD 84.79) versus 111.46 (SD 84.3), respectively, with *P*<.001. A decreased albumin/globulin ratio (A/G ratio) was recently reported as a novel independent predictor of the development of postflexible ureteroscopic sepsis [29]. In this study, the A/G ratio in the sepsis group was lower than that in the nonsepsis group, at 0.76 (SD 0.7) versus 1.14 (SD 0.69), respectively, with *P*<.001. Furthermore, several studies found that activated partial thromboplastin time (APTT) and prothrombin time (PT) [30,31] are prognostic biomarkers for the specific identification of patients with sepsis. We observed that the occurrence of sepsis led to increased APTT and PT values, with an *APTT* value of 6.89 (SD 15.01) to 14.06 (SD 19.72), *PT(%)* value of 17.51 (SD 35.82) to 33.27 (SD 42.26), *PT(INR)* value of 0.23 (SD 0.46) to 0.49 (SD 0.63), and *PT(sec)* value of 2.9 (SD 5.84) to 6.14 (SD 7.69), with *P*<.001. With favorable consistency between the identified lab test differences of our candidates and known biomarker candidates of sepsis, we felt confident that the SMC EMRs successfully recapitulated the physical and biological signatures of the patients with sepsis. In other words, these results suggested that the 35 selected biomarkers could characteristically reflect the biological features of the patients with sepsis. Thus, we utilized these 35 lab test results as learning features to establish our prediction model for sepsis, as shown in Figure 3.

Figure 3. Distributions of the two groups for the selected lab test types. Of the 64 total lab test types (Multimedia Appendix 1), 35 lab test types showed significantly different distributions between the sepsis and control groups. ****: *P*<.001; A/G ratio: albumin/globulin ratio; ALC: absolute lymphocyte count; ALP: alkaline phosphatase; ANC: absolute neutrophil count; APTT: activated partial thromboplastin time; AST: aspartate aminotransferase; BUN: blood urea nitrogen; ESR: erythrocyte sedimentation rate; LD: lactate dehydrogenase; PT: prothrombin time; RBC: red blood cell.



Prediction of Sepsis Using Machine Learning Approaches

Using vectorized drug relationships and the values of the selected lab test types along with common EHRs, we trained 2 machine learning models (logistic regression and random forest) and 3 deep learning models (ANN, convolutional neural network [CNN], and RNN) to build a sepsis prediction model based on the physical status of patients with cancer. A total of 465 relationships between 31 drugs selected through association rule analysis were vectorized using the 3 formulas described in the Methods section. A total of 1395 (465 × 3) drug relationship vectors, the values of the selected 35 lab test types, and common EHR information including anonymized personal information, hospitalization data, and cancer diagnosis code were used as inputs for model training. We used simple logistic regression (LR) and RNN-LSTM for the logistic regression–based and RNN-LSTM–based models, respectively. The random forest–based model comprised 20 trees, and the ANN-based model comprised input and output layers, as well as hidden layers. In addition, we used ResNet10 consisting of 10 convolution layers, fully connected layers, and residual connections for the CNN-based model. All hyperparameters, such as the number of trees in the random forest model, batch size, learning rate, and number of layers in the deep learning models, were selected as optimal values for each model through grid searches [32]. The list of feature variables used in our proposed model is given in Multimedia Appendix 3. To verify the proposed sepsis prediction model, we compared the predictive performances with the models trained on only

common EHRs (ie, demography, diagnoses codes, and others) and the models trained on common EHRs and drug relationships by 5-fold cross-validation. Regarding performance evaluation metrics, the accuracy, area under the receiver operating characteristic (AUROC), area under the precision-recall curve (AUPRC), precision, recall, and F1 score were used. Multimedia Appendix 4 shows the performance evaluation results.

The overall performance of the proposed models with EHRs, lab data, and drug relationships were superior to that of the other models. The proposed random forest–based model showed the highest value in all the evaluation metrics except for recall (accuracy: 0.692, AUROC: 0.753, AUPRC: 0.573, precision: 0.518, recall: 0.718, and F1 score: 0.602). In the case of recall, the proposed ANN-based model showed the highest value (accuracy: 0.654, AUROC: 0.723, AUPRC: 0.522, precision: 0.477, recall: 0.721, and F1 score: 0.574). In particular, the proposed random forest–based proposed model recorded the largest performance improvement in all the metrics compared to the model trained on drug relationships and common EHRs (accuracy: 0.645 to 0.692, AUROC: 0.69 to 0.753, AUPRC: 0.487 to 0.573, precision: 0.465 to 0.518, recall: 0.629 to 0.718, and F1 score: 0.534 to 0.602). In addition, the proposed RNN-LSTM–based model showed the greatest performance improvement in accuracy, the AUROC, the AUPRC, and precision (accuracy: 0.603 to 0.675, AUROC: 0.655 to 0.729, AUPRC: 0.447 to 0.555, and precision: 0.43 to 0.504), and the ResNet10-based model showed the highest improvement in recall and the F1 score (recall: 0.577 to 0.689, F1 score: 0.499 to 0.567) compared to the model trained on only common EHRs. These findings suggest that the drug relationships and the selected lab test types were the main contributors to the proposed sepsis predictive models for patients with cancer.

Investigation of Important Features

To evaluate the contributions of the learning features, SHAP, an XAI technique, was utilized for the proposed random forest–based model, which showed the best performance when investigating important features that contributed to the prediction. Multimedia Appendix 5 shows the contribution ratios of the top 50 important features among 1738 features obtained through SHAP, where the x-axis denotes the feature contribution ratio, and the y-axis denotes the names of the features.

The top 50 important features include 26 lab test types and 15 drug relationships among the 31 drugs and the 35 lab test types selected by *t* test and association rule analysis, respectively. The 15 drug relationships contained narcotic analgesic drugs such as "*opioid alkaloids*" and "*synthetic narcotics*," which were prescribed more, along with other drugs, in the sepsis group. Among the characteristics of the patients with cancer, the number of cancer-infiltrating lymph nodes (Ca_LN_no), the degree of cancer extent (Extend_CD), and the size of the primary tumor (T_CD) were observed as decisive contributing factors.

As expected, prognostic biomarkers of sepsis, such as the albumin level, PT, A/G ratio, total protein level, and cholesterol level, ranked high. The blood platelet count has also been identified as a major contributor, and platelets are involved in mechanisms that promote immune responses and coagulation activation. Thrombocytopenia is common in ICU patients with sepsis and is reportedly associated with fatal outcomes [33]. The migration of neutrophils to infection sites is essential in the host's defense against invading pathogens during sepsis [34], which may have led to the absolute neutrophil count or segmented neutrophils improving the predictive performance of the model. Moreover, when expanded to the top 100, all selected lab test types except "*band neutrophil*," "*nucleated RBC*," and "*carbon dioxide, total*," as well as 49 drug relationships comprising 22 selected drugs, were included. These results show that the selected drug relationships and lab tests were important features in the proposed sepsis predictive model, suggesting that these features contributed to the accurate prediction of the model.

Discussion

Principal Findings

This study presents a machine learning–based approach to identify sepsis risk in patients with cancer at an early stage (2 days before onset). We elucidated that the relationships of prescribed medications and lab test patterns were distinct in the sepsis and control groups. Based on these analysis results, we built a machine learning–based sepsis prediction model trained on lab test items and vectorized drug relationships, along with EHRs. The proposed model outperformed the model trained on medication relationships or common EHRs. In particular, the proposed random forest–based model showed the best sepsis prediction performance (accuracy: 0.692, AUROC: 0.753, and F1 score: 0.602) and showed the greatest performance improvement. Furthermore, we demonstrated that the selected lab test results and drug relationships were indeed important features and mainly contributed to the accurate prediction of our proposed model. Therefore, lab tests and medication relationships can be used as efficient features for predicting sepsis. Consequently, it will be possible to use EHR information and deep learning methods to identify the risk of sepsis in patients with cancer.

Limitations

Several limitations of the study should be noted. First, health records are not intended specifically for research; nonbilling-related data, including self-reported data such as smoking status, would be partially inaccurate. As depicted in Table 1, a substantial portion of patients with cancer are diagnosed with liver or lung cancer. Although there is a fairly significant incidence of liver and lung cancer in Korea [35], characteristic signatures of lab results and medications (eg, a lower A/G ratio and usage of opioid alkaloids) among patients with sepsis should be addressed at the pan-cancer level in further studies. For the contribution of the relationships of medication pairs, we acknowledge that there are many stakeholders in the prescription of medications, including insurance coverage. In this study, patients in Korea were all covered by the National Health Insurance Service. Thus, there would be a limited utilization of the relationship of medication combinations for model training in further applications from different countries corresponding to the heterogeneous milieu of insurance coverages.

As we hypothesized, our network-based analysis disclosed distinct patterns of medications used between sepsis and nonsepsis patients with cancer. For example, synthetic narcotics and opioid agents appeared to be more frequently prescribed with other agents. These features (ie, lab test results and medication patterns) mainly contributed to the high performance of our prediction model. Because the usage of opioids is a known risk factor for sepsis [36], the possibility of iatrogenic effects for the medication pattern–based prediction of sepsis in patients with cancer remains unclear. Therefore, drug-drug interactions between synthetic narcotics and anticancer agents should be addressed to further understand sepsis in patients with cancer. The retrospective analysis of EHRs paves the way for future research to understand sepsis among patients with cancer.

Conclusion

To our knowledge, previous prognostic evaluation tools and models primarily use patient information obtained after admission to the ICU, and there are many limitations for medical interventions. However, since most patients with cancer are hospitalized through the emergency room for the initial diagnosis of sepsis, an appropriate evaluation tool is needed to identify the risk in advance. This study can be referenced as a baseline for efficiently predicting the onset of sepsis in patients with cancer, and the model is expected to be able to identify sepsis risk more accurately and earlier than before in the medical field.

Acknowledgments

This work was supported by the Korea Institute of Science and Technology Information (KISTI) (K-21-L02-C10, K-20-L02-C10-S01). Authors HP and JK were also supported by the Ministry of Science and ICT (N-21-NM-CA08-S01). This research was also supported by the Program of the National Research Foundation (NRF) funded by the Korean government (2021M3H9A203052011). The computational analysis was supported by the National Supercomputing Center, including the resources and technology. We also thank Samsung Medical Center for providing the data.

Data Availability

The data sets used and analyzed in this study are available from the corresponding author on reasonable request.

Authors' Contributions

HP conceptualized the study and its methodology, supervised the study, and was responsible for funding acquisition. WCC and JY acquired the data. DY and JK carried out the formal analysis, developed and validated the model, and carried out the visualization. WCC and HP were responsible for the project administration. JY performed a technical review of the Methods section. DY, JK, and HP wrote the original draft of the paper, and JK and HP reviewed and edited the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Topological properties of the network.
[PNG File , 233 KB - medinform_v10i6e37689_app1.png]

Multimedia Appendix 2
Comparison of lab test numerical distribution in the sepsis versus control groups.
[DOCX File , 23 KB - medinform_v10i6e37689_app2.docx]

Multimedia Appendix 3
Description of the feature variables used in the proposed sepsis prediction model.
[DOCX File , 22 KB - medinform_v10i6e37689_app3.docx]

Multimedia Appendix 4
Performance evaluation results. (A-F) Comparison of the sepsis prediction performance for the models trained on only common EHRs (only EHRs), the models trained on common EHRs and drug relationships (EHRs+Drug Rel), and the models trained on drug relationships and lab test types, along with common EHRs (EHRs+Drug Rel+Lab test types (proposed)) obtained by 5-fold cross-validation. AUROC: area under the receiver operating characteristic; AUPRC: area under the precision-recall curve; EHR: electronic health record; LR: logistic regression; RF: random forest; ANN: artificial neural network; ResNet10: residual convolutional neural network with 10 layers; LSTM: long short-term memory recurrent neural network.
[PNG File , 106 KB - medinform_v10i6e37689_app4.png]

Multimedia Appendix 5
Feature contribution ratio of the Shapley Additive Explanations. Bar chart shows the feature contribution ratio (%) of the top 50 features obtained by the Shapley Additive Explanations algorithm for the random forest-based sepsis prediction model. A/G ratio: albumin/globulin ratio; ALP: alkaline phosphatase; AM: taking medications in the morning (ante meridiem); APTT: activated partial thromboplastin time; AST: aspartate aminotransferase; BUN: blood urea nitrogen; Ca_LN_no: number of cancer-infiltrating lymph nodes; Extend_CD: degree of cancer extent; IV: intravenous administration; T_CD: size of primary tumor; PT: prothrombin time; RBC: red blood cell.
[PNG File , 132 KB - medinform_v10i6e37689_app5.png]

References

1. Gullo A, Bianco N, Berlot G. Management of severe sepsis and septic shock: challenges and recommendations. Crit Care Clin 2006 Jul;22(3):489-501, ix. [doi: 10.1016/j.ccc.2006.03.006] [Medline: 16893735]
2. Fleischmann C, Scherag A, Adhikari N, Hartog CS, Tsaganos T, Schlattmann P, International Forum of Acute Care Trialists. Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limitations. Am J Respir Crit Care Med 2016 Feb 01;193(3):259-272. [doi: 10.1164/rccm.201504-0781OC] [Medline: 26414292]
3. Martin GS. Sepsis, severe sepsis and septic shock: changes in incidence, pathogens and outcomes. Expert Rev Anti Infect Ther 2012 Jun 10;10(6):701-706 [FREE Full text] [doi: 10.1586/eri.12.50] [Medline: 22734959]
4. American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. Crit Care Med 1992;20(6):864-874. [doi: 10.1097/00003246-199206000-00025]
5. Hunt A. Sepsis: an overview of the signs, symptoms, diagnosis, treatment and pathophysiology. Emerg Nurse 2019 Sep 02;27(5):32-41. [doi: 10.7748/en.2019.e1926] [Medline: 31475503]
6. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock*. Crit Care Med 2006;34(6):1589-1596. [doi: 10.1097/01.ccm.0000217961.75225.e9]
7. Iskander KN, Osuchowski MF, Stearns-Kurosawa DJ, Kurosawa S, Stepien D, Valentine C, et al. Sepsis: multiple abnormalities, heterogeneous responses, and evolving understanding. Physiol Rev 2013 Jul;93(3):1247-1288 [FREE Full text] [doi: 10.1152/physrev.00037.2012] [Medline: 23899564]
8. Ménétrier-Caux C, Ray-Coquard I, Blay J, Caux C. Lymphopenia in Cancer Patients and its Effects on Response to Immunotherapy: an opportunity for combination with Cytokines? J Immunother Cancer 2019 Mar 28;7(1):85 [FREE Full text] [doi: 10.1186/s40425-019-0549-5] [Medline: 30922400]
9. Williams MD, Braun L, Cooper LM, Johnston J, Weiss RV, Qualy RL, et al. Hospitalized cancer patients with severe sepsis: analysis of incidence, mortality, and associated costs of care. Crit Care 2004 Oct;8(5):R291-R298 [FREE Full text] [doi: 10.1186/cc2893] [Medline: 15469571]
10. Brown SM, Jones J, Kuttler KG, Keddington RK, Allen TL, Haug P. Prospective evaluation of an automated method to identify patients with severe sepsis or septic shock in the emergency department. BMC Emerg Med 2016 Aug 22;16(1):31 [FREE Full text] [doi: 10.1186/s12873-016-0095-0] [Medline: 27549755]
11. Barton C, Chettipally U, Zhou Y, Jiang Z, Lynn-Palevsky A, Le S, et al. Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. Comput Biol Med 2019 Jun;109:79-84 [FREE Full text] [doi: 10.1016/j.compbiomed.2019.04.027] [Medline: 31035074]
12. Goh KH, Wang L, Yeow AYK, Poh H, Li K, Yeow JJL, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun 2021 Jan 29;12(1):711 [FREE Full text] [doi: 10.1038/s41467-021-20910-4] [Medline: 33514699]
13. Park MY, Yoon D, Lee K, Kang SY, Park I, Lee S, et al. A novel algorithm for detection of adverse drug reaction signals using a hospital electronic medical record database. Pharmacoepidemiol Drug Saf 2011 Jun 06;20(6):598-607. [doi: 10.1002/pds.2139] [Medline: 21472818]
14. Paik H, Chung A, Park H, Park RW, Suk K, Kim J, et al. Repurpose terbutaline sulfate for amyotrophic lateral sclerosis using electronic medical records. Sci Rep 2015 Mar 05;5(1):8580 [FREE Full text] [doi: 10.1038/srep08580] [Medline: 25739475]
15. Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules. In: Proceedings of the 20th International Conference on Very Large Data Bases. 1994 Presented at: VLDB '94; Sept 12-15; Santiago de Chile, Chile.
16. Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, et al. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. Crit Care Med 2020 Feb;48(2):210-217 [FREE Full text] [doi: 10.1097/CCM.0000000000004145] [Medline: 31939789]
17. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020 Mar 21;46(3):383-400 [FREE Full text] [doi: 10.1007/s00134-019-05872-y] [Medline: 31965266]

18. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016 Feb 23;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
19. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. SIGMOD Rec 2000 Jun;29(2):1-12. [doi: 10.1145/335191.335372]
20. Kim B, Kim Y, Park CHK, Rhee SJ, Kim YS, Leventhal BL, et al. Identifying the medical lethality of suicide attempts using network analysis and deep learning: nationwide study. JMIR Med Inform 2020 Jul 09;8(7):e14500 [FREE Full text] [doi: 10.2196/14500] [Medline: 32673253]
21. Cox DR. The Regression Analysis of Binary Sequences. J R Stat Soc Series B Stat Methodol 2018 Dec 05;21(1):238-238. [doi: 10.1111/j.2517-6161.1959.tb00334.x]
22. Breiman L. Random forests. Machine Learning 2001;45:5-32. [doi: 10.1023/A:1010933404324]
23. Tadeusiewicz R. Neural networks: A comprehensive foundation. Control Eng Pract 1995 May;3(5):746-747. [doi: 10.1016/0967-0661(95)90080-2]
24. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016 Jun Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 27-30; Las Vegas, NV. [doi: 10.1109/cvpr.2016.90]
25. Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Proceedings of the Annual Conference of the International Speech Communication Association. 2014 Feb Presented at: INTERSPEECH; Sep 14-18; Singapore.
26. Lundberg S, Lee S. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 2017.
27. Deep Sepsis in Cancer. URL: https://github.com/yangdonghun3/deep_sepsis_in_cancer [accessed 2022-05-24]
28. Takegawa R, Kabata D, Shimizu K, Hisano S, Ogura H, Shintani A, et al. Serum albumin as a risk factor for death in patients with prolonged sepsis: An observational study. J Crit Care 2019 Jun;51:139-144 [FREE Full text] [doi: 10.1016/j.jcrc.2019.02.004] [Medline: 30825787]
29. Lu J, Xun Y, Yu X, Liu Z, Cui L, Zhang J, et al. Albumin-globulin ratio: a novel predictor of sepsis after flexible ureteroscopy in patients with solitary proximal ureteral stones. Transl Androl Urol 2020 Oct;9(5):1980-1989 [FREE Full text] [doi: 10.21037/tau-20-823] [Medline: 33209662]
30. Benediktsson S, Frigyesi A, Kander T. Routine coagulation tests on ICU admission are associated with mortality in sepsis: an observational study. Acta Anaesthesiol Scand 2017 Aug 06;61(7):790-796. [doi: 10.1111/aas.12918] [Medline: 28681428]
31. Dempfle CH, Lorenz S, Smolinski M, Wurst M, West S, Houdijk WPM, et al. Utility of activated partial thromboplastin time waveform analysis for identification of sepsis and overt disseminated intravascular coagulation in patients admitted to a surgical intensive care unit. Crit Care Med 2004;32(2):520-524. [doi: 10.1097/01.ccm.0000110678.52863.f3]
32. Shekar B, Dagnew G. Grid search-based hyperparameter tuning and classification of microarray cancer data. 2019 Feb Presented at: International Conference on Advanced Computational and Communication Paradigms; Feb 25-28; Gangtok, Sikkim, India p. 1-8. [doi: 10.1109/icaccp.2019.8882943]
33. Vardon-Bounes F, Ruiz S, Gratacap M, Garcia C, Payrastre B, Minville V. Platelets are critical key players in sepsis. Int J Mol Sci 2019 Jul 16;20(14):3494 [FREE Full text] [doi: 10.3390/ijms20143494] [Medline: 31315248]
34. Sônego F, Castanheira FVES, Ferreira RG, Kanashiro A, Leite CAVG, Nascimento DC, et al. Paradoxical roles of the neutrophil in sepsis: protective and deleterious. Front Immunol 2016 Apr 26;7:155 [FREE Full text] [doi: 10.3389/fimmu.2016.00155] [Medline: 27199981]
35. Korean Liver Cancer Association K, National Cancer Center N. 2018 Korean Liver Cancer Association–National Cancer Center Korea Practice Guidelines for the Management of Hepatocellular Carcinoma. Gut Liver 2019 May 15;13(3):227-299. [doi: 10.5009/gnl19024]
36. Zhang R, Meng J, Lian Q, Chen X, Bauman B, Chu H, et al. Prescription opioids are associated with higher mortality in patients diagnosed with sepsis: A retrospective cohort study using electronic health records. PLoS One 2018 Jan 2;13(1):e0190362 [FREE Full text] [doi: 10.1371/journal.pone.0190362] [Medline: 29293575]

Abbreviations

A/G: albumin/globulin
ANN: artificial neural network
APTT: activated partial thromboplastin time
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic
CDW: Clinical Data Warehouse
CNN: convolutional neural network
EHR: electronic health record
ER: emergency room

FP-growth: frequent pattern growth
ICU: intensive care unit
KISTI: Korea Institute of Science and Technology Information
LR: logistic regression
NRF: National Research Foundation
PT: prothrombin time
ResNet10: residual convolutional neural networks
RF: random forest
RNN-LSTM: long short-term memory recurrent neural networks
Sepsis-3: Third International Consensus Definitions for Sepsis and Septic Shock
SHAP: Shapley Additive Explanations
SMC: Samsung Medical Center
SOFA: Sequential Organ Failure Assessment
XAI: Explainable Artificial Intelligence

Edited by G Eysenbach; submitted 03.03.22; peer-reviewed by S Molani, H Park; comments to author 24.03.22; revised version received 18.04.22; accepted 17.05.22; published 15.06.22.

Please cite as:
Yang D, Kim J, Yoo J, Cha WC, Paik H
Identifying the Risk of Sepsis in Patients With Cancer Using Digital Health Care Records: Machine Learning–Based Approach
JMIR Med Inform 2022;10(6):e37689
URL: https://medinform.jmir.org/2022/6/e37689
doi:10.2196/37689
PMID:35704364

Original Paper

Predicting Risk of Hypoglycemia in Patients With Type 2 Diabetes by Electronic Health Record–Based Machine Learning: Development and Validation

Hao Yang[1*], MSc; Jiaxi Li[2*], MSc; Siru Liu[3], PhD; Xiaoling Yang[4], MSc; Jialin Liu[1,5], MD

[1]Information Center, West China Hospital, Sichuan University, Chengdu, China

[2]Department of Clinical Laboratory Medicine, Jinniu Maternity and Child Health Hospital of Chengdu, Chengdu, China

[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

[4]West China School of Nursing, Endocrinology and Metabolism Department, West China Hospital, Sichuan University, Chengdu, China

[5]Department of Medical Informatics, West China Medical School, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Jialin Liu, MD
Information Center
West China Hospital
Sichuan University
No 37 Guoxue Road
Chengdu, 610041
China
Phone: 86 28 85422306
Fax: 86 28 85582944
Email: dljl8@163.com

Abstract

Background: Hypoglycemia is a common adverse event in the treatment of diabetes. To efficiently cope with hypoglycemia, effective hypoglycemia prediction models need to be developed.

Objective: The aim of this study was to develop and validate machine learning models to predict the risk of hypoglycemia in adult patients with type 2 diabetes.

Methods: We used the electronic health records of all adult patients with type 2 diabetes admitted to West China Hospital between November 2019 and December 2021. The prediction model was developed based on XGBoost and natural language processing. F1 score, area under the receiver operating characteristic curve (AUC), and decision curve analysis (DCA) were used as the main criteria to evaluate model performance.

Results: We included 29,843 patients with type 2 diabetes, of whom 2804 patients (9.4%) developed hypoglycemia. In this study, the embedding machine learning model (XGBoost3) showed the best performance among all the models. The AUC and the accuracy of XGBoost are 0.82 and 0.93, respectively. The XGboost3 was also superior to other models in DCA.

Conclusions: The Paragraph Vector–Distributed Memory model can effectively extract features and improve the performance of the XGBoost model, which can then effectively predict hypoglycemia in patients with type 2 diabetes.

(JMIR Med Inform 2022;10(6):e36958) doi:10.2196/36958

KEYWORDS

diabetes; type 2 diabetes; hypoglycemia; learning; machine learning model; EHR; electronic health record; XGBoost; natural language processing

Introduction

Diabetes is a serious long-term disease. The global prevalence of diabetes in people aged 20-79 years is estimated to be 10.5% (536.6 million) in 2021 and will rise to 12.2% (783.2 million) by 2045. Global health expenditures related to diabetes are estimated US $966 billion in 2021 and projected to reach US

$1054 billion by 2045 [1]. Diabetes continues to be a major clinical and public health concern [2].

Hypoglycemia (blood glucose<3.9 mmol/L or 70 mg/dL) is a common adverse event of diabetes treatment. Hospital hypoglycemia occurs in 3%-18% of hospitalized diabetic patients [3]. Severe hypoglycemia usually causes potentially life-threatening complications and is associated with an increase length of stay and mortality [4,5]. Hypoglycemia is especially common in older patients with diabetes [5], and the risk doubles every decade after the age of 60 years [6]. Many factors can lead to a high risk of hypoglycemia in older patients, including physiological changes in drug metabolism, age-related decline in renal function, cognitive decline, an increase in comorbidity, and potential overtreatment [7,8]. Since there are many risk factors that induce hypoglycemia in patients with diabetes, and some risk factors may also change during hospitalization, it is a challenge to identify and prevent hypoglycemia in people with diabetes [9,10].

In recent years, machine learning has been widely used for hypoglycemia prediction. For example, Schroeder et al [11] employed the Cox prediction model for the 6-month risk of hypoglycemia. Karter et al [12] developed a tool to identify patients with type 2 diabetes at a high risk of hypoglycemia. Plis et al [13] described a support vector regression model for predicting hypoglycemic events. Furthermore, Jin et al [14] have integrated deep learning with natural language processing (NLP) to automatically detect hypoglycemic events from electronic health record (EHR) notes.

Although numerous hypoglycemia prediction models have been developed, there is still a need to improve the accuracy and effectiveness of hypoglycemia prediction. In this study, we developed XGBoost ensembling NLP to predict the risk of hypoglycemia in hospitalized patients with type 2 diabetics, using data readily available in the EHRs.

Methods

Our cohort included patients with type 2 diabetes from West China Hospital of Sichuan University. All patient data were obtained from the hospital's EHR system.

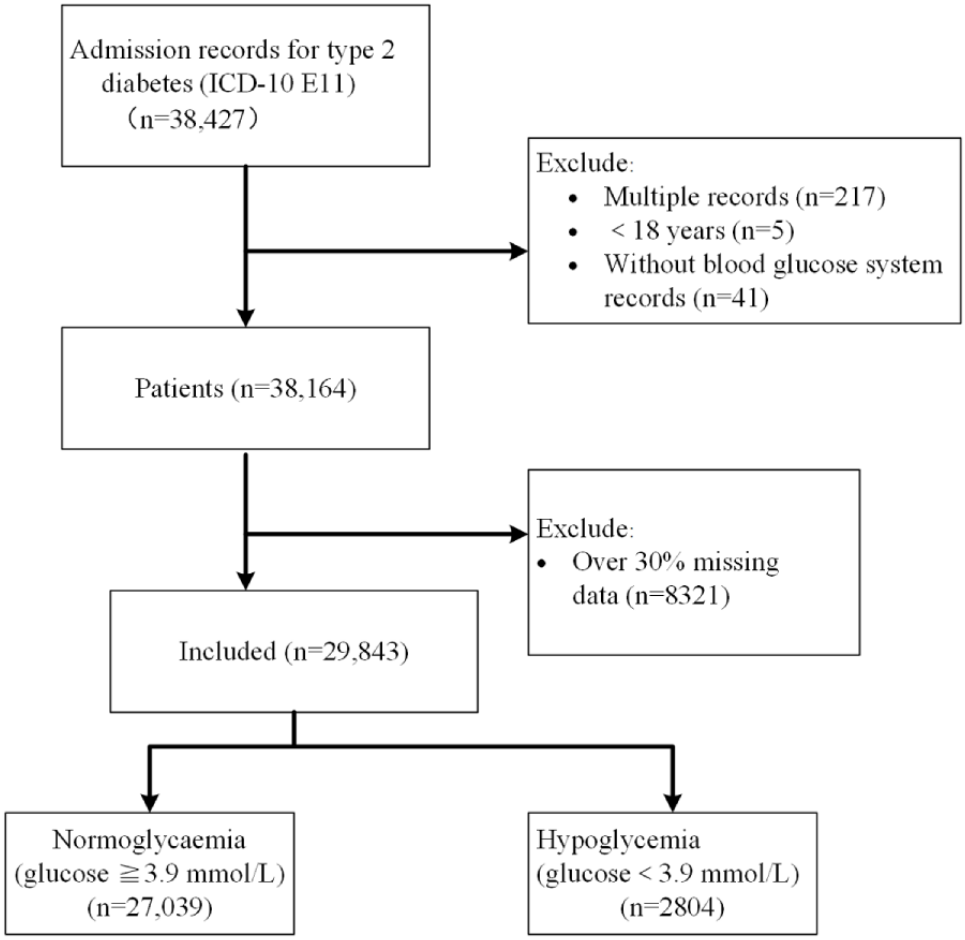
Ethics Approval

The study was approved by the Medical Ethics Committee of West China Hospital Sichuan University (2020-608). West China Hospital is a large teaching hospital with 4300 beds and a leading medical center of western China [15].

Patients

We performed a retrospective analysis of the available EHR of all patients with type 2 diabetes who were admitted to West China Hospital between November 2019 and December 2021. With the protection of patient privacy, only data related to the patient's hospitalization were retrieved, and the diagnosis was established based on International Classification of Diseases, 10th revision (ICD-10). The following inclusion criteria were used: (1) all patients with type 2 diabetes based on ICD-10, E11 (type 2 diabetes mellitus) with a hospital stay>24 hours; (2) patients aged 18 years or older. Patients with more than 30% missing values were excluded from the analysis [16]. The patient selection process is shown in Figure 1.

Figure 1. The patient selection process.



Variables Analyzed

The variables used to predict the risk of hypoglycemia in patients with type 2 diabetes included various demographic, laboratory, and clinical variables, as well as EHR notes. The extraction of variables was based on experts' opinion and our research [16-20]. These variables were collected during the first 24 hours of admission. Through data preprocessing, we analyzed some missing values (Table 1). Random forest regression was used to handle all missing numerical variables.

Table 1. Statistics of missing values (N=29,843).

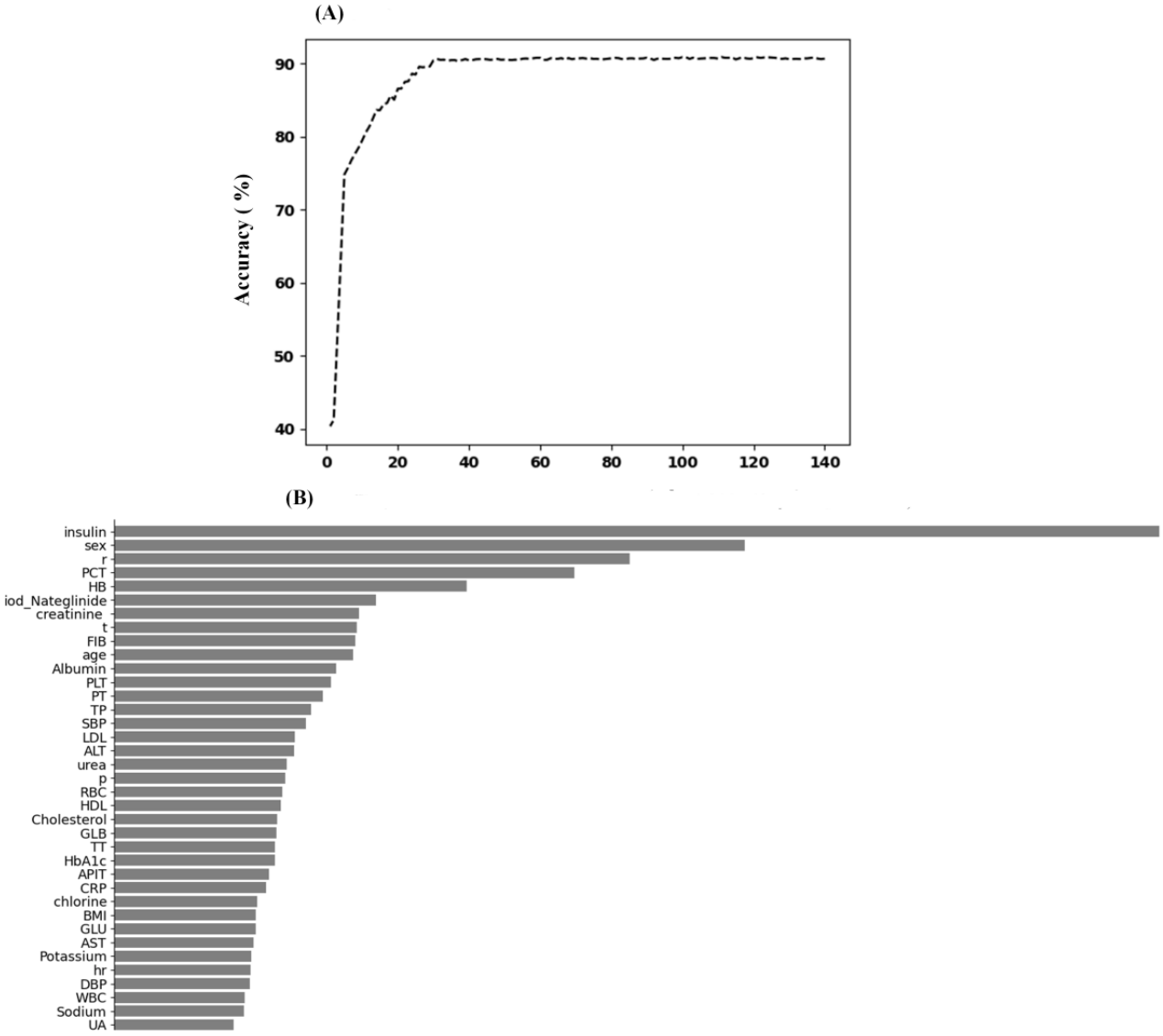
Features	Missing data, n (%)
Red blood cell count	1860 (6.2)
Hemoglobin	1858 (6.2)
Blood platelet count	1883 (6.3)
White blood cell count	1858 (6.2)
Total protein	1791 (6.0)
Albumin	1768 (5.9)
Globulin	1812 (6.1)
Urea	1755 (5.9)
Alanine aminotransferase	1821 (6.1)
Aspartate aminotransferase	1809 (6.1)
Cholesterol	2126 (7.1)
High-density lipoprotein	2128 (7.1)
Low-density lipoprotein	2131 (7.1)
Sodium	1516 (5.1)
Chlorine	1585 (5.3)
Thrombin time	3970 (13.3)
Creatinine	1749 (5.9)
Uric acid	1769 (5.9)
C-reactive protein	18,249 (61.1)
Procalcitonin	20,101 (67.3)
Glycosylated hemoglobin or HbA_{1c}[a]	14,410 (48.3)
Prothrombin time	3725 (12.5)
Activated partial thromboplastin time	3779 (12.7)

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

Variable Selection

After extracting all the variables, the parameter of feature importance in XGBoost was used to select and filter important variables [21]. The parameters were set as follows: the number of estimators was 100 and max depth was set to 6. Ultimately, 37 predictive variables and their weights were selected from 176 variables (Figure 2).

Figure 2. The weights of variables importance. ALT: alanine aminotransferase; APIT: activated partial thromboplastin time; AST: aspartate aminotransferase; CRP: C-reactive protein; DBP: diastolic blood pressure; FIB: fibrinogen; GLB: globulin; GLU: glucose; HB: hemoglobin; HbA_{1c}: glycated hemoglobin; HDL: high-density lipoprotein; hr: heart rate; iod-Nateglinide: iodine urea and Nateglinide; LDL: low-density lipoprotein; p: pulse; PCT: procalcitonin; PLT: blood platelet count; PT: prothrombin time; r: respiratory rate; RBC: red blood cell count; SBP: systolic pressure; t: body temperature; TP: total protein; TT: thrombin time; UA: uric acid; WBC: white blood cell count. (A) the curve between the number of features and accuracy. (B) the weights of variables importance (when accuracy is up to 90%).



Data Imbalance

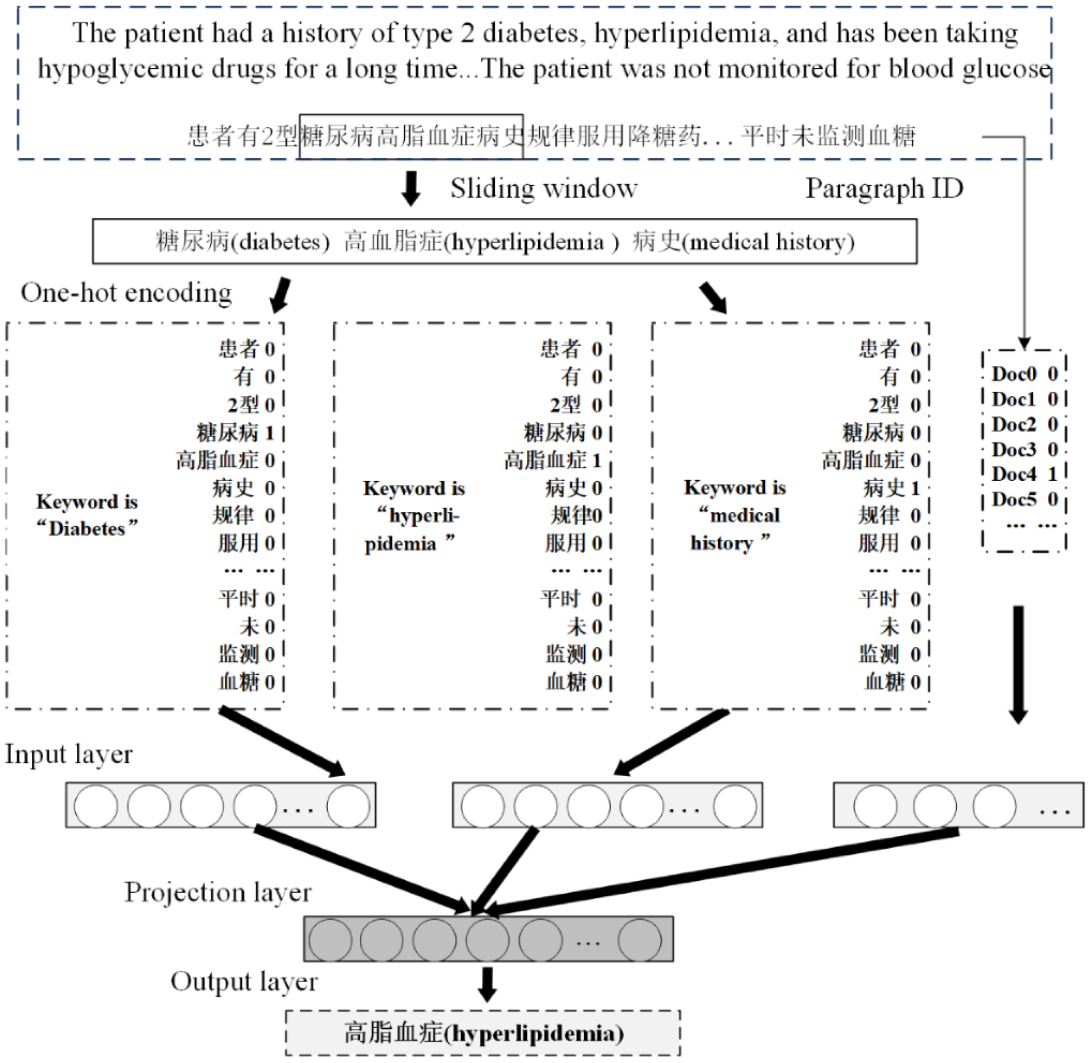
To overcome the data imbalance between the group with hypoglycemia and the normoglycemic control group, we used the Adaptive Synthetic (ADASYN) sampling method [22] to oversample the group with hypoglycemia to generate a portion of data that was comparable to the data from the normoglycemic group. The method for imbalanced learning was used to generate a low sample size to improve class imbalance. We used 5-fold cross-validation and sample balancing with ADASYN for each stratified training set. ADASYN was implemented using Imblearn in Python (Version 0.9.0; imbalanced-learn documentation) [23]. The sampling ratio was set to 1.

Embedding Models

We used a Python implementation of the Paragraph Vector model available at Gensim [24] and trained 100-dimensional vectors on our corpus. Due to the large computing time of training these corpora and because they are unlabeled corpora, we trained the distributed memory model of paragraph vectors or Paragraph Vector–Distributed Memory (PV-DM) [25] based on textual data from patients with diabetes (Figure 3).

The doc2vec model was used to train and feature map the chief complaints (CCs), history of present illness (HPI), and family history (FH) in the EHR. The results were feature fused into XGBoost model [21] to generate XGBoost1 (XGBoost+CC), XGBoost2 (XGBoost+CC+HPI), and XGBoost3 (XGBoost+CC+HPI+FH).

Figure 3. Training process of Paragraph Vector–Distributed Memory (PV-DM) model.



Statistical Analysis

As clinical indicators, categorical variables were shown as counts and percentages, and continuous variables were shown as means and SDs. Comparisons between groups were analyzed by a 2-tailed *t* test for continuous variables and chi-square test for categorical variables. All statistical analyses were carried out in R software (version 4.1.2; R Core Team). The statistical significance was considered as $P<.05$. The computing environment of this study includes central processing unit i7-7800x; memory 16 GB; operating system Windows 11, build 22598.200; and Python programming language.

Results

Participant Characteristics

The cohort included 29,843 patients with type 2 diabetes, of whom 2804 (9.4%) patients developed hypoglycemia. Among the 29,843 patients, the proportion of female patients in the group with hypoglycemia (n=1065, 38.0%) was higher than that in the normoglycemia group (n=9479, 35.1%; $P=.002$). The BMI of patients in the hypoglycemia and normoglycemia groups were 23.6 (SD 5.24) and 24.3 (SD 4.26), respectively. Statistically, the BMI of patients in the normoglycemia group was significantly higher than that of patients in the hypoglycemia group ($P<.001$). The proportion of insulin use in patients in hypoglycemia group (n=1575, 56.2%) was much higher than that for patients in the normoglycemia group (n=7306, 27.0%). In addition, the proportion of patients taking sulfonylureas or Nateglinide in the hypoglycemia group (n=1382, 49.3%) was also higher than that in the normoglycemia group (n=9273, 34.3%), which was a statistically significant difference ($P<.001$). The demographics of patients in normoglycemia and hypoglycemia groups are shown in Table 2.

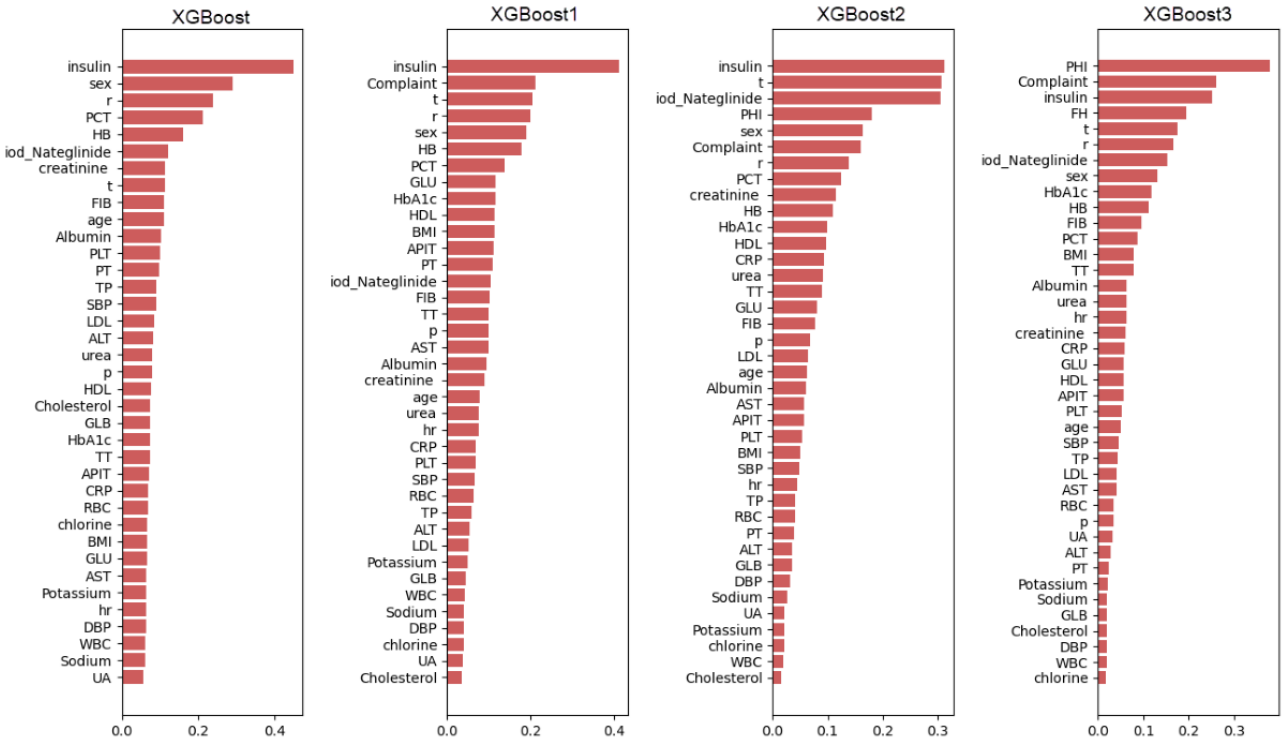
Table 2. Demographics of patients with diabetes (N=29,843).

Variables	Normoglycemia (blood glucose>3.9 mmol/L; n=27,039)	Hypoglycemia (blood glucose<3.9 mmol/L; n=2804)	*P* values
Sex, n (%)			.002
Female	9479 (35.1)	1065 (38)	
Male	17,560 (64.9)	1739 (62)	
Age (years), mean (SD; range)	64.2 (12.3; 18-104)	64.8 (12.6; 19-98)	.03
BMI, mean (SD)	24.3 (4.26)	23.6 (5.24)	<.001
Insulin, n (%)			<.001
No	19,733 (73)	1229 (43.8)	
Yes	7306 (27)	1575 (56.2)	
Sulfonylureas or Nateglinide, n (%)			<.001
No	17,766 (65.7)	1422 (50.7)	
Yes	9273 (34.3)	1382 (49.3)	

Feature Selection

We applied XGBoost and its ensemble models for feature selection to discard noninformative features and retain important features (Figure 4). Finally, 37 features were selected from 176 features. In the XGBoost model, insulin was the most important predictor variable among all the predictor variables, followed by sex, respiratory rate, Procalcitonin, and hemoglobin (Figure 4). However, these variables had different weights in XGBoost1, XGBoost2, and XGBoost3 (Figure 4).

Figure 4. Weight of the variables in the different models. ALT: alanine aminotransferase; APIT: activated partial thromboplastin time; AST: aspartate aminotransferase; CRP: C-reactive protein; DBP: diastolic blood pressure; FIB: fibrinogen; GLB: globulin; GLU: glucose; HB: hemoglobin; HDL: high-density lipoprotein; hr: heart rate; iod-Nateglinide: Iodine urea and Nateglinide; LDL: low-density lipoprotein; p: pulse; PCT: procalcitonin; PLT: blood platelet count; PT: prothrombin time; r: respiratory rate; RBC: red blood cell count; SBP: systolic pressure; t: body temperature; TP: total protein; TT: thrombin time; UA: uric acid; WBC: white blood cell count.



Model Performance

Table 3 shows the results of the 4 machine learning methods after 5-fold cross-validation. The area under the receiver operating characteristic curve (AUC=0.822) and accuracy (0.934) of the XGBoost3 were higher than all other models. The XGboost3 was superior to other models in terms of model

performance, which was evaluated using AUC and decision curve analysis [26] (Figure 5).

The oversample may have an impact on the accuracy of the test set. After completing the model training, we sampled 138 adult patients with type 2 diabetes (hypoglycemia=28, nonhypoglycemia=110) in West China Hospital from January to March 2022 for validation. The results showed that the prediction accuracy rate reaches 89.86%. The confusion matrix is shown in Figure 6.

Table 3. Accuracy and area under the receiver operating characteristic curve (AUC) of different models.

Model	Embedding method	AUC, mean (SD)	Accuracy, mean (SD)	*P* value
XGBoost	XGBoost	0.718 (0.0014)	0.892 (0.002)	• N/A[a]
XGBoost1	XGBoost+CC[b]	0.785 (0.0012)	0.919 (0.002)	• <.001 vs XGBoost
XGBoost2	XGBoost+CC+HPI[c]	0.817 (0.0023)	0.928 (0.001)	• <.001 vs XGBoost • <.001 vs XGBoost1
XGBoost3	XGBoost+CC+HPI+FH[d]	0.822 (0.0024)	0.934 (0.002)	• <.001 vs XGBoost • <.001 vs XGBoost1 • <.001 vs XGBoost2

[a]N/A: not applicable.

[b]CC: chief complaints.

[c]HPI: history of present illness.

[d]FH: family history.

Figure 5. Comparison between the change detection algorithm (CDA) and receiver operating characteristic (ROC) curve of different models. (A) The ROC curve of the 4 models. (B) The DCA curve of the 4 models.

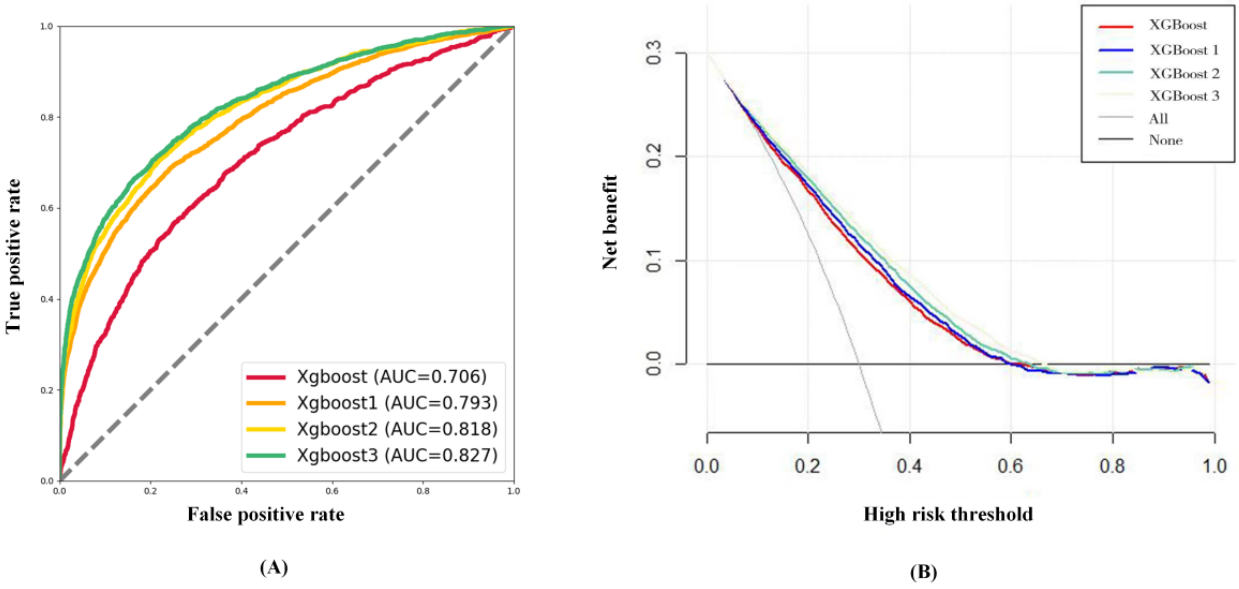
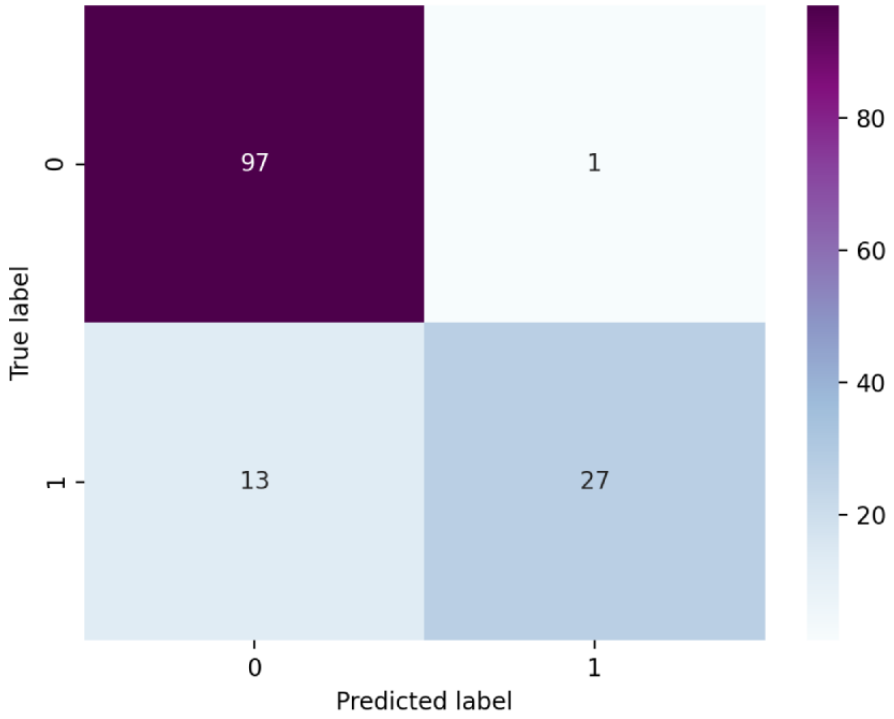


Figure 6. The confusion matrix of XGBoost3.



Discussion

Principal Findings

In this study, we used only some common types of features within 24 hours of patient admission to develop a hypoglycemia prediction model, because the earlier hypoglycemia is predicted or detected, the better we can avoid it. This study found that in women, patients at an older age, patients with low BMI, and those using insulin or various hypoglycemic drugs, the risk of hypoglycemia developing in type 2 diabetes increases. There was a statistical difference (P<.001). Some studies have shown that these factors increase significantly in the incidence of hypoglycemia in patients with type 2 diabetes [27-29]. This may be related to the higher risk of sulfonylurea-related hypoglycemia in women compared to men [30]. One possible reason for this is the pharmacokinetics and pharmacodynamics of sulfonylureas in women [31]. In patients with type 2 diabetes, low BMI may be associated with reduced insulin resistance [32]. Patients with obesity can benefit from the same type of antidiabetic drugs that patients with low or normal weight use [33]. This phenomenon is known as the “obesity paradox,” but the mechanism is unknown [34]. This indicates that a standard BMI or overweight are key determinants in reducing the risk of severe hypoglycemia in patients with type 2 diabetes [35].

We developed a hypoglycemia risk prediction model based on XGBoost integrated PV-DM, which can be applied to patients with type 2 diabetes. The result showed that XGBoost3 has the largest AUC and highest accuracy to predict hypoglycemia. There is a significant difference between this model and other models (P<.001). Consistent with previous research [36], combining numerical variables with textual data from EHR can effectively improve the predictive performance of the model. Applying this model to clinical practice could help physicians adjust hypoglycemic drugs based on patient characteristics and hypoglycemia risk factors. This study demonstrates that the inclusion of EHR increases the prognostic accuracy of hypoglycemia in patients with diabetes, providing a more comprehensive and optimized method for predicting hypoglycemic events.

This study also has some limitations. First, the study was carried out in a single institution, and the performance of the model and the distribution of covariates may differ when applied to a sample from a different institution. Second, this study involved Chinese patients. Due to ethnic differences, the results of this study need to be further verified in other ethnic groups.

Conclusions

We developed a multivariate risk prediction model to predict the occurrence of hypoglycemia in patients with type 2 diabetes. In this prediction model, the PV-DM model can effectively extract the EHR notes and improve the performance of the XGBoost model.

The predictive model can help predict the occurrence of hypoglycemia in patients with type 2 diabetes and provide clinicians with an effective way to prevent hypoglycemia in patients with diabetes. In future research, we will focus on external validation of this model in a larger cohort of patients with type 2 diabetes and explore combining state-of-the-art methods in NLP with deep learning to enhance the model’s predictive power.

Authors' Contributions

J Liu, HY, and J Li conceptualized the study. J Liu, HY, J Li, SL, and XY carried out the collection and analysis of the literature and data, and drafted the manuscript. Both HY and J Li acted as first authors for this study. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Res Clin Pract 2022 Jan;183:109119. [doi: 10.1016/j.diabres.2021.109119] [Medline: 34879977]
2. Khan MAB, Hashim MJ, King JK, Govender RD, Mustafa H, Al Kaabi J. Epidemiology of type 2 diabetes - global burden of disease and forecasted trends. J Epidemiol Glob Health 2020 Mar;10(1):107-111 [FREE Full text] [doi: 10.2991/jegh.k.191028.001] [Medline: 32175717]
3. Ruan Y, Tan GD, Lumb A, Rea RD. Importance of inpatient hypoglycaemia: impact, prediction and prevention. Diabet Med 2019 Apr;36(4):434-443. [doi: 10.1111/dme.13897] [Medline: 30653706]
4. Yun J, Ko S. Avoiding or coping with severe hypoglycemia in patients with type 2 diabetes. Korean J Intern Med 2015 Jan;30(1):6-16 [FREE Full text] [doi: 10.3904/kjim.2015.30.1.6] [Medline: 25589828]
5. Pathak RD, Schroeder EB, Seaquist ER, Zeng C, Lafata JE, Thomas A, SUPREME-DM Study Group. Severe hypoglycemia requiring medical intervention in a large cohort of adults with diabetes receiving care in U.S. integrated health care delivery systems: 2005-2011. Diabetes Care 2016 Mar;39(3):363-370 [FREE Full text] [doi: 10.2337/dc15-0858] [Medline: 26681726]
6. Huang ES, Laiteerapong N, Liu JY, John PM, Moffet HH, Karter AJ. Rates of complications and mortality in older patients with diabetes mellitus: the diabetes and aging study. JAMA Intern Med 2014 Feb 01;174(2):251-258 [FREE Full text] [doi: 10.1001/jamainternmed.2013.12956] [Medline: 24322595]
7. Ligthelm RJ, Kaiser M, Vora J, Yale J. Insulin use in elderly adults: risk of hypoglycemia and strategies for care. J Am Geriatr Soc 2012 Aug;60(8):1564-1570. [doi: 10.1111/j.1532-5415.2012.04055.x] [Medline: 22881394]
8. American Diabetes Association. 12. Older Adults: standards of Medical Care in Diabetes-2020. Diabetes Care 2020 Jan;43(Suppl 1):S152-S162. [doi: 10.2337/dc20-S012] [Medline: 31862755]
9. Hulkower RD, Pollack RM, Zonszein J. Understanding hypoglycemia in hospitalized patients. Diabetes Manag (Lond) 2014 Mar;4(2):165-176 [FREE Full text] [doi: 10.2217/DMT.13.73] [Medline: 25197322]
10. Mathioudakis NN, Abusamaan MS, Shakarchi AF, Sokolinsky S, Fayzullin S, McGready J, et al. Development and Validation of a Machine Learning Model to Predict Near-Term Risk of Iatrogenic Hypoglycemia in Hospitalized Patients. JAMA Netw Open 2021 Jan 04;4(1):e2030913 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.30913] [Medline: 33416883]
11. Schroeder EB, Xu S, Goodrich GK, Nichols GA, O'Connor PJ, Steiner JF. Predicting the 6-month risk of severe hypoglycemia among adults with diabetes: development and external validation of a prediction model. J Diabetes Complications 2017 Jul;31(7):1158-1163 [FREE Full text] [doi: 10.1016/j.jdiacomp.2017.04.004] [Medline: 28462891]
12. Karter AJ, Warton EM, Lipska KJ, Ralston JD, Moffet HH, Jackson GG, et al. Development and validation of a tool to identify patients with type 2 diabetes at high risk of hypoglycemia-related emergency department or hospital use. JAMA Intern Med 2017 Oct 01;177(10):1461-1470 [FREE Full text] [doi: 10.1001/jamainternmed.2017.3844] [Medline: 28828479]
13. Plis K, Bunescu R, Marling C, Shubrook J, Schwartz F. A machine learning approach to predicting blood glucose levels for diabetes management. AAAI Workshop: Modern Artificial Intelligence for Health Analytics 2014 Jun 18 [FREE Full text]
14. Jin Y, Li F, Vimalananda VG, Yu H. Automatic detection of hypoglycemic events from the electronic health record notes of diabetes patients: empirical study. JMIR Med Inform 2019 Nov 08;7(4):e14340 [FREE Full text] [doi: 10.2196/14340] [Medline: 31702562]
15. West China Hospital of Sichuan University. URL: http://www.wchscu.cn/Home.html [accessed 2022-03-25]
16. Liu J, Wu J, Liu S, Li M, Hu K, Li K. Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model. PLoS One 2021;16(2):e0246306 [FREE Full text] [doi: 10.1371/journal.pone.0246306] [Medline: 33539390]
17. Silbert R, Salcido-Montenegro A, Rodriguez-Gutierrez R, Katabi A, McCoy RG. Hypoglycemia among patients with type 2 diabetes: epidemiology, risk factors, and prevention strategies. Curr Diab Rep 2018 Jun 21;18(8):53 [FREE Full text] [doi: 10.1007/s11892-018-1018-0] [Medline: 29931579]
18. Ravaut M, Sadeghi H, Leung KK, Volkovs M, Kornas K, Harish V, et al. Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data. NPJ Digit Med 2021 Feb 12;4(1):24 [FREE Full text] [doi: 10.1038/s41746-021-00394-8] [Medline: 33580109]
19. Arvind V, Kim JS, Oermann EK, Kaji D, Cho SK. Predicting surgical complications in adult patients undergoing anterior cervical discectomy and fusion using machine learning. Neurospine 2018 Dec;15(4):329-337 [FREE Full text] [doi: 10.14245/ns.1836248.124] [Medline: 30554505]

20. Li K, Shi Q, Liu S, Xie Y, Liu J. Predicting in-hospital mortality in ICU patients with sepsis using gradient boosting decision tree. Medicine (Baltimore) 2021 May 14;100(19):e25813 [FREE Full text] [doi: 10.1097/MD.0000000000025813] [Medline: 34106618]
21. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD '16; August 13-17; San Francisco, CA p. 785-794.
22. Haibo H, Yang B, Garcia E. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). 2008 Presented at: IEEE; 1-8 June; Hong Kong p. 1322-1328.
23. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. JMLR 2017 Jan;18(1):559-563 [FREE Full text]
24. Řehůřek R, Sojka P. Gensim-statistical semantics in Python. 2011 Presented at: EuroScipy; 25-28 August; Paris URL: https://www.fi.muni.cz/usr/sojka/posters/rehurek-sojka-scipy2011.pdf
25. Le Q, Mikolov T. Distributed representations of sentences and documents. In: PMLR. 2014 Presented at: 31st International Conference on Machine Learning; June 21-26; Beijing, China p. 1188-1196 URL: http://proceedings.mlr.press/v32/le14.pdf
26. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565-574 [FREE Full text] [doi: 10.1177/0272989X06295361] [Medline: 17099194]
27. Han K, Yun J, Park Y, Ahn Y, Cho J, Cha S, et al. Development and validation of a risk prediction model for severe hypoglycemia in adult patients with type 2 diabetes: a nationwide population-based cohort study. Clin Epidemiol 2018;10:1545-1559 [FREE Full text] [doi: 10.2147/CLEP.S169835] [Medline: 30425585]
28. Lee SE, Kim K, Son KJ, Song SO, Park KH, Park SH, et al. Trends and risk factors in severe hypoglycemia among individuals with type 2 diabetes in Korea. Diabetes Res Clin Pract 2021 Aug;178:108946 [FREE Full text] [doi: 10.1016/j.diabres.2021.108946] [Medline: 34252506]
29. Gonzalez C, Monti C, Pinzon A, Monsanto H, Ejzykowicz F, Argentinean Recap Group. Prevalence of hypoglycemia among a sample of sulfonylurea-treated patients with type 2 diabetes mellitus in Argentina: The real-life effectiveness and care patterns of diabetes management (RECAP-DM) study. Endocrinol Diabetes Nutr (Engl Ed) 2018 Dec;65(10):592-602. [doi: 10.1016/j.endinu.2018.05.014] [Medline: 30076124]
30. Kajiwara A, Kita A, Saruwatari J, Oniki K, Morita K, Yamamura M, et al. Higher risk of sulfonylurea-associated hypoglycemic symptoms in women with type 2 diabetes mellitus. Clin Drug Investig 2015 Sep;35(9):593-600. [doi: 10.1007/s40261-015-0314-6] [Medline: 26293520]
31. Soldin OP, Mattison DR. Sex differences in pharmacokinetics and pharmacodynamics. Clin Pharmacokinet 2009;48(3):143-157 [FREE Full text] [doi: 10.2165/00003088-200948030-00001] [Medline: 19385708]
32. Tsai T, Lee C, Cheng B, Kung C, Chen F, Shen F, et al. Body Mass Index-mortality relationship in severe hypoglycemic patients with type 2 diabetes. Am J Med Sci 2015 Mar;349(3):192-198 [FREE Full text] [doi: 10.1097/MAJ.0000000000000382] [Medline: 25526505]
33. Cai X, Yang W, Gao X, Zhou L, Han X, Ji L. Baseline Body Mass Index and the efficacy of hypoglycemic treatment in type 2 diabetes: a meta-analysis. PLoS One 2016;11(12):e0166625 [FREE Full text] [doi: 10.1371/journal.pone.0166625] [Medline: 27935975]
34. Gravina G, Ferrari F, Nebbiai G. The obesity paradox and diabetes. Eat Weight Disord 2021 May;26(4):1057-1068. [doi: 10.1007/s40519-020-01015-1] [Medline: 32954485]
35. Plečko D, Bennett N, Mårtensson J, Bellomo R. The obesity paradox and hypoglycemia in critically ill patients. Crit Care 2021 Nov 01;25(1):378 [FREE Full text] [doi: 10.1186/s13054-021-03795-z] [Medline: 34724956]
36. Arnaud E, Elbattah M, Gignon M, Dequen G. Deep learning to predict hospitalization at triage: integration of structured data and unstructured text. In: IEEE. 2020 Presented at: 2020 IEEE International Conference on Big Data (Big Data); December 10-13; Atlanta, GA p. 4836-4841.

Abbreviations

ADASYN: adaptive synthetic
AUC: area under the receiver operating characteristic curve
CC: chief complaints
DCA: decision curve analysis
EHR: electronic health record
FH: family history
HPI: history of present illness
ICD-10: International Classification of Diseases, 10th revision
NLP: natural language processing
PV-DM: Paragraph Vector–Distributed Memory

Edited by C Lovis; submitted 31.01.22; peer-reviewed by M Elbattah, A Staffini, E Sükei; comments to author 27.04.22; revised version received 08.05.22; accepted 31.05.22; published 16.06.22.

Please cite as:
Yang H, Li J, Liu S, Yang X, Liu J
Predicting Risk of Hypoglycemia in Patients With Type 2 Diabetes by Electronic Health Record–Based Machine Learning: Development and Validation
JMIR Med Inform 2022;10(6):e36958
URL: https://medinform.jmir.org/2022/6/e36958
doi:10.2196/36958
PMID:35708754

Original Paper

Vaccine Adverse Event Mining of Twitter Conversations: 2-Phase Classification Study

Sedigheh Khademi Habibabadi[1,2*], PhD; Pari Delir Haghighi[3*], PhD; Frada Burstein[3], Prof Dr; Jim Buttery[1,4*], Prof Dr

[1]Centre for Health Analytics, Melbourne Children's Campus, Melbourne, Australia

[2]Department of General Practice, University of Melbourne, Melbourne, Australia

[3]Department of Human-Centred Computing, Faculty of Information Technology, Monash University, Melbourne, Australia

[4]Department of Paediatrics, University of Melbourne, Melbourne, Australia

[*]these authors contributed equally

Corresponding Author:
Sedigheh Khademi Habibabadi, PhD
Centre for Health Analytics
Melbourne Children's Campus
50 Flemington Rd
Melbourne, 3052
Australia
Phone: 61 0383416200
Email: sedigh.khademi@gmail.com

Abstract

Background: Traditional monitoring for adverse events following immunization (AEFI) relies on various established reporting systems, where there is inevitable lag between an AEFI occurring and its potential reporting and subsequent processing of reports. AEFI safety signal detection strives to detect AEFI as early as possible, ideally close to real time. Monitoring social media data holds promise as a resource for this.

Objective: The primary aim of this study is to investigate the utility of monitoring social media for gaining early insights into vaccine safety issues, by extracting vaccine adverse event mentions (VAEMs) from Twitter, using natural language processing techniques. The secondary aims are to document the natural language processing techniques used and identify the most effective of them for identifying tweets that contain VAEM, with a view to define an approach that might be applicable to other similar social media surveillance tasks.

Methods: A VAEM-Mine method was developed that combines topic modeling with classification techniques to extract maximal VAEM posts from a vaccine-related Twitter stream, with high degree of confidence. The approach does not require a targeted search for specific vaccine reaction–indicative words, but instead, identifies VAEM posts according to their language structure.

Results: The VAEM-Mine method isolated 8992 VAEMs from 811,010 vaccine-related Twitter posts and achieved an F_1 score of 0.91 in the classification phase.

Conclusions: Social media can assist with the detection of vaccine safety signals as a valuable complementary source for monitoring mentions of vaccine adverse events. A social media–based VAEM data stream can be assessed for changes to detect possible emerging vaccine safety signals, helping to address the well-recognized limitations of passive reporting systems, including lack of timeliness and underreporting.

(JMIR Med Inform 2022;10(6):e34305) doi:10.2196/34305

KEYWORDS

immunization; vaccines; natural language processing; vaccine adverse effects; vaccine safety; social media; Twitter; machine learning

Introduction

Background

Vaccines belong to the broad category of medicines, in a subcategory known as *biologicals* [1]. Unlike medicines that are prescribed to limited populations as a course of *treatment* for a disease, vaccines are given to both healthy and vulnerable populations at large, sometimes over a short period, to enhance their immune systems' ability to combat a pathogen. In contrast to those who are taking a medicine to help to cure a disease or to treat unwanted symptoms, most people receiving a vaccine are not ill. Therefore, there is a deferred individual benefit to taking a vaccine, and, consequently, a very low acceptance of risk regarding vaccines [2]. In addition, the pathophysiology of vaccine-related adverse events is not as well defined as those of adverse drug reactions—a reaction triggered by a vaccine could be caused by any of its multiple ingredients, its underlying technology (eg, messenger RNA–based vs protein-based delivery), or even an error in administration [3]—and some people are particularly prone to reacting to vaccine ingredients [4]. Furthermore, a vaccine's *time to market* may be curtailed, such as has occurred during the COVID-19 pandemic, and so provide less opportunities for studying potential vaccine side effects over a large population for a long time.

Vaccine safety relies upon rigorous compliance to development and manufacturing standards, well conducted clinical trials, thorough assessment, licensing, control, and administration of vaccines. Postlicensure vaccine safety surveillance is a key component of ensuring vaccine safety [5] and continues in a variety of forms after regulatory approval or emergency use authorization. It is the primary mechanism to identify serious or rare adverse events following immunization (AEFI) that are unlikely to have been exposed by prelicensure trials, and it allows surveillance in populations that were unable to be included in the trials [6]. Identification of minor AEFI is potentially as important as those of severe adverse events, as minor AEFI may act as a surrogate warning for more severe sequelae (eg, increased rates of fever may be a marker for increased febrile seizures [7])—that is, increased incidences of even minor events could indicate larger problems.

Traditional passive (spontaneous) surveillance systems, where a voluntary reporting of AEFI is made by individuals or by their treating health professionals, are the main method of vaccine safety monitoring and have proven to be useful in early detection of vaccine-related and drug-related safety issues [8,9]. Although these systems are the backbone of drug safety monitoring, they suffer from major disadvantages, including underreporting, incomplete data, and time lag between an event happening and subsequent reporting of it [10]. Active surveillance systems survey vaccine recipients and vaccine administrators to determine the outcomes of recent vaccinations, irrespective of any AEFI experience. Increasingly, alternate data sources are being added to surveillance systems, as they offer the potential to capture timely and additional measurements of the quantity of possible adverse events.

Extensive use of social media has provided a platform for sharing and seeking health-related information. Social media data have consequently become a widely used source of data for public health research [11]. In comparison with established traditional surveillance systems, social media monitoring is inexpensive and near to real time and covers large populations [12], thus offering an easily accessible wide-ranging data source for tracking emerging trends—which may be unavailable or less noticeable in data gathered by traditional reporting systems [13].

Many researchers have used social media as a pharmacovigilance source [14]. However, there is relative deficit in the use of social media for AEFI detection. Many investigations of vaccine and vaccination-related social media posts are related to sentiments, attitudes, and opinions [15-21]. Studies on using social media for detection of adverse drug reaction have included vaccine-related words in keyword searches used for collecting data. An example is an annotated data set of tweets containing 250 drug-related keywords, including *vaccine*, for over a period of 4 months [22]. We downloaded and assessed these data sets, but they did not contain any AEFI data. A total of 2 recent studies have focused on detecting influenza [23] and COVID-19 [24] vaccine adverse events from Twitter. However, the emphasis of both these studies were on identifying specific vaccine adverse events using a lexicon of adverse reactions.

Objectives

In this paper, we use the term *vaccine adverse event mention* (VAEM) to refer to *any* vaccine-related personal health mention, that is, VAEMs are conversations that contain personal health mentions in a vaccine context. This distinguishes VAEM from the AEFI and adverse drug reaction signals used in previous studies on the use of social media for vaccine and drug reaction surveillance, as these are searching for specific adverse vaccine events and drug reactions.

Although vaccine safety surveillance systems monitor for unexpected, rare, and late-onset events, they also aim to observe changes in the rate of known and expected events, because "while rare but particularly serious events can be detected through review of each individual report or active surveillance, an increased incidence in a more common AEFI is often more difficult to detect, and has been described as akin to 'finding a needle in the haystack'" [13]. VAEM are conversations, ideally gathered in volume, that contain information that may be the common AEFI that are so elusive to traditional reporting, while also allowing the detection of previously unknown severe events.

This paper presents the VAEM-Mine method, which encapsulates the workflow and techniques required to enable detection of VAEM by applying natural language processing techniques to a relatively unfocused social media stream, consisting of any vaccine-related Twitter conversation. The VAEM-Mine method detects likely VAEM based on their characteristics of being *personal health mentions* in a vaccination context. VAEM-Mine has 2 components—a topic modeling process that initially detects and filters for VAEM (described in a previous publication [25]) and a classification task that accurately identifies VAEM in the filtered data—which is described in detail in this paper.

Methods

Ethics Approval

Ethics approval for this study was granted by Monash University Human Research Ethics Committee (project ID 11767).

Data Collection

The Twitter application program interface was used to collect English tweets with search terms *vaccination, vaccinations, vaccine, vaccines, vax, vaxx, vaxine, vaccinated, vaccinated, flushot,* and *flu shot*. These were general terms that were designed to collect a broadly representative sample of vaccine-related conversations. We included *flu shot* as a keyword because we found that this was most often used, rather than the term *flu vaccine*, whereas other vaccines were usually mentioned in conjunction with the word *vaccine*—and thus, for them, we only needed to search for *vaccine* keywords. Upon examining the downloaded data for specific vaccine names, we found more records mentioning other vaccines than those mentioning the influenza vaccine. No specific reaction mentions were used.

A total of 400,000 tweets were initially collected across 5 months, from February 7, 2018, to June 7, 2018, which were used for an initial training and evaluation of topic models and classifiers. An additional 411,010 tweets were collected from August 9, 2018, to July 20, 2019, which were used to verify the trained topic models and classifiers and to train more powerful classifiers. The resulting data consisted of a total of 811,010 tweets and a daily average of 2906 tweets.

The data were prepared by removing URLs and by converting to lower case. Duplicates were removed based on tweet ID and text. Other preparation included removing hashtags, usernames, punctuation, and numbers. Tweets with <5 words were removed. N-grams were created for topic modeling; preparation for classification is explained in the following section. The final cleaned tweets were 82.21% (328,822/400,000) of the initial collection and 87.48% (359,535/411,010) of the second collection—a total of 688,357.

Table 1 illustrates a sample of tweets that mention receiving vaccinations or vaccines. The first 3 examples contain genuine VAEM, but the others do not—even when the language is similar. Our goal was to first isolate the most likely records describing personal experiences of vaccination and then to refine that selection to those that are genuine adverse reaction mentions.

Table 1. Sample of vaccine-related tweets.

Tweet	Type
"Aw wtf my poor arm is dead af from my flu shot."	VAEM[a]
"Cannot lie on belly, baby gets squished; cannot lie on back, baby squishes; cannot lie on right side, i get heartburn; cannot lie on left side, vax arm is sore; let the third trimester moaning begin!"	VAEM
"2 people recently, including my 88yo father, had flu shot and really bad reaction afterwards. both said it was probably as bad as getting the flu!!! flu2018 maybe undercooked the vaccine."	VAEM
"I got vaccinated as a kid. As a result, I'm now starting to gray and bald. My balding got so bad I had to shave my head. I've also gained weight. Because of vaccines I've started aging instead of dying as a baby."	Non-VAEM
"Urgent vaccination plea after measles outbreak in West Yorkshire."	Non-VAEM
"Researchers are developing a personalized vaccine which they hope could tackle ovarian cancer."	Non-VAEM

[a]VAEM: vaccine adverse event mention.

The topic modeling showed that VAEM and similar personal health mentions were a distinct topic (among 13 vaccine-related topics), and therefore, that topic models could be used to filter for the tweets that were most similar to VAEM. Taking tweets from only that topic meant that relatively homogenous data sets could be created for labeling and subsequent training of classifiers. The use of topic modeling for filtering data before classification was adopted as a core component of the VAEM-Mine method. A previous publication [25] described the process of choosing the best performing topic models for the method, including a detailed description of the scoring method used to identify the best models.

Classification

Overview

As described in the previous section, data were collected in 2 phases. Topic models were trained on the first-phase data and were used to filter that data and the subsequent second-phase data into likely VAEM-containing data sets, which were then used for classification. Classifiers were trained and assessed with the filtered first-phase data set and the combined (filtered) first-phase and second-phase data sets. The following section describes the creation of these data sets; the subsequent section describes the classifiers.

Classification Data Sets

The original prepared (cleaned) data collections of 328,822 and 359,535 tweets were reduced, by applying topic model–based filtering, to data sets containing 18,801 (5.72%) and 80,372 (22.35%) tweets that were more likely to contain VAEMs—a total of 99,173 tweets, which was only 14.41% (99,173/688,357) of the total original cleaned data.

Therefore, filtering eliminated approximately 85.59% (589,184/688,357) of the data, which did not contain any significant numbers of VAEM. These more VAEM-focused data sets were binary labeled by the author (SKH), as either VAEM or non-VAEM. All the labels were verified by the

domain expert. Although only 10.07% (9991/99,173) of the tweets were identified as VAEM, this was a considerably better proportion of VAEM compared with the original cleaned data, which contained VAEM in only 1.45% (9991/688,357) of the tweets.

Balanced data sets of 18.72% (3519/18,801) and 19.57% (15,730/80,372) of the tweets were created from these imbalanced data sets together with holdout test data sets—these were an imbalanced test set of 3.27% (614/18,801) of the tweets and a balanced test set of 1.03% (828/80,372) of the tweets. The main data sets were named *Phase-One* and *Phase-Two* data sets, and the test data sets were referred to as *Phase-One Test* and *Phase-Two Test* data sets.

The imbalanced Phase-One Test data set of 3.27% (614/18,801) of the tweets were obtained from Victoria, Australia, in the period preceding and during the 2018 influenza immunization period. These tweets were assembled to enable comparison of tweet trends with statistics from the Australian Victorian vaccine authority, Surveillance of Adverse Events Following Vaccination In the Community. With 90 VAEM and 524 non-VAEM, the test set was imbalanced but reflected how the data were obtained through the topic model filtering process, without any subsequent balancing. The Phase-One Test data set was used as a benchmark throughout the classification testing. The data sets (Table 2) were combined to retrain classifiers and train transformer-based classifiers—becoming a *Combined* data set of 19,249 tweets and a *Combined Test* data set of 1442 tweets. The training data were split into training and validation data with a 75:25 ratio.

Table 2. Data set numbers.

Stage	Phase-One data, n (%)	Phase-Two data, n (%)	Total, n
Topic modeling	328,822 (47.77)	359,535 (52.23)	688,357
Filtering out by topic modeling	−310,021 (52.62)	−279,163 (47.38)	−589,184
After topic modeling	18,801 (18.96)	80,372 (81.04)	99,173
Filtering out by data preparation and balancing	−14,668 (18.69)	−63,814 (81.31)	−78,482
For classification training	4133 (19.97)	16,558 (80.03)	20,691
For training and validation	3519 (18.28)	15,730 (81.72)	19,249
For testing	614 (42.58)	828 (57.42)	1442

Classifiers

Our default data approach with traditional models (ie, not neural network–based) was *bag-of-words* [26], represented via compressed sparse matrices. We used SKLearn (Scikit-learn) [27] vectorizing libraries such as TfidfTransformer [28] for tokenizing lowercase text for the standard classifiers. A grid or random search was used to ascertain the best combinations of vectorizer, removal of stop words and numbers, and n-grams. The neural networks used dense word embedding vectors via a Word2Vec skip-gram corpus [29] for Convolutional Neural Networks (CNNs) and Long Short-Term Memories (LSTMs), and the Word2Vec corpus used Gensim library functions [30] using all the Twitter data. The transformer models used byte-pair-encoding [31]; the byte-pair-encoding tokens were derived only from the filtered texts we had retained from topic modeling. The classifiers are listed in Table 3, and details of their definitions and parameters are listed in Multimedia Appendix 1.

Table 3. List of classifiers.

Models	Library or GitHub source
LR CV[a]	sklearn.linear_model [32]
SGD[b] Classifier	sklearn.linear_model [32]
Linear SVC[c]	sklearn.svm.SVC [33]
RF[d]	sklearn.ensemble [34]
Extra Trees	sklearn.ensemble [34]
Multinomial NB[e]	sklearn.naive_bayes [35]
NB SVM[f] (combined NB and Linear SVM)	GitHub Joshua-Chin/nbsvm [36]
XGBoost[g]	GitHub dmlc/xgboost [37]
Ensemble (NB SVM, LR CV, SGD, Linear SVC, and RF)	Majority voting [38]
CNN,[h] LSTM,[i] BiLSTM,[j] GRU,[k] BiGRU,[l] CNN-BiLSTM, and CNN-BiGRU	Pytorch [39], RaRe-Technologies [30], Shawn1993 [40], and bamtercelboo [41]
RoBERTa,[m] RoBERTa Large, BERT,[n] XLNet,[o] XLNet Large, and XLM[p]	Pytorch; huggingface transformers [42]

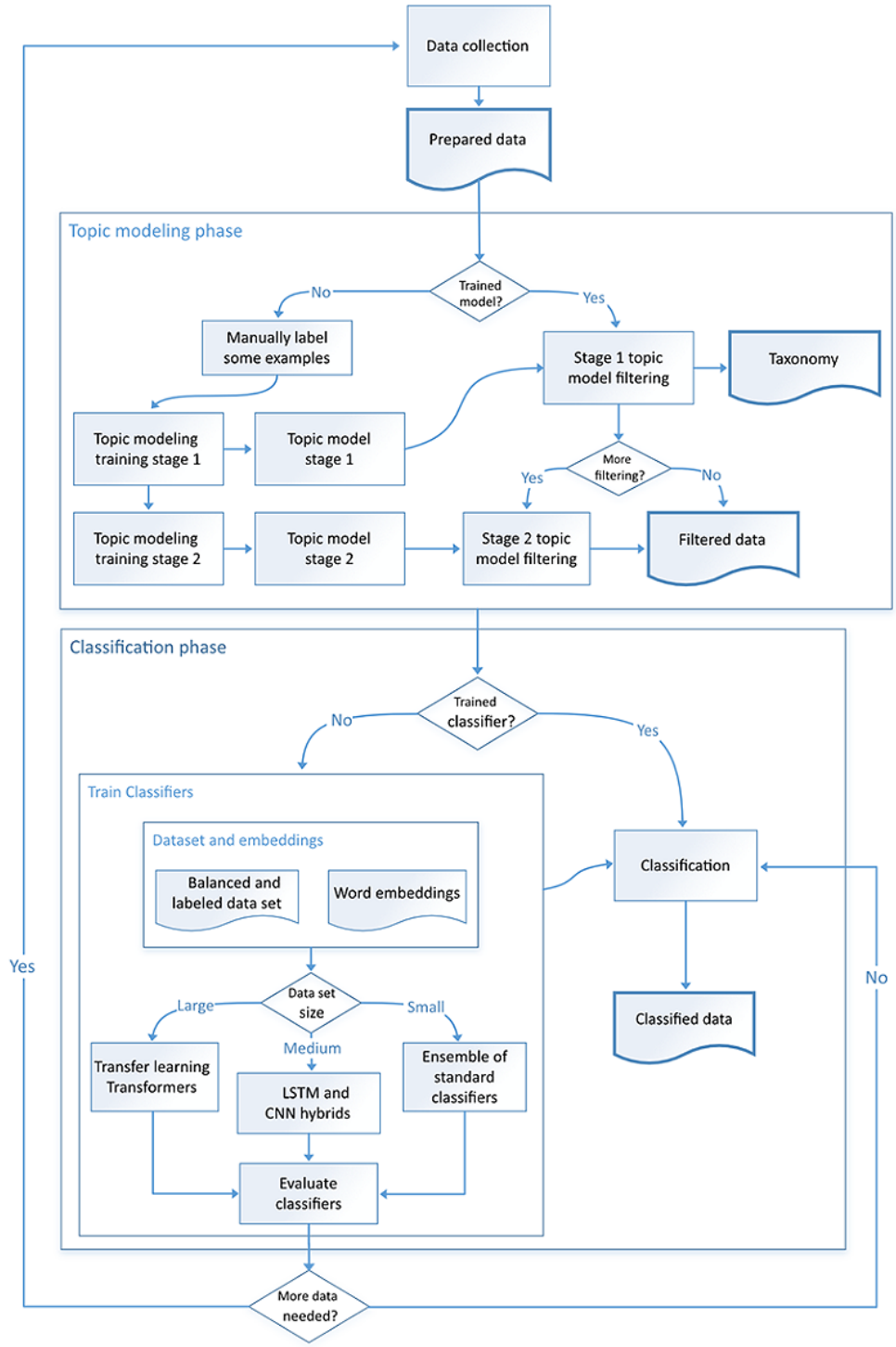
[a]LR CV: Logistic Regression Cross Validation.
[b]SGD: Stochastic Gradient Descent.
[c]SVC: Support Vector Classification.
[d]RF: Random Forest.
[e]NB: Naïve Bayes.
[f]SVM: Support Vector Machine.
[g]XGBoost: Extreme Gradient Boosting.
[h]CNN: Convolutional Neural Network.
[i]LSTM: Long Short-Term Memory.
[j]BiLSTM: Bidirectional LSTM.
[k]GRU: Gated Recurrent Unit.
[l]BiGRU: Bidirectional Gated Recurrent Unit.
[m]RoBERTa: Robustly Optimized Bidirectional Encoder Representations Pretraining Approach.
[n]BERT: Bidirectional Encoder Representations.
[o]XLNet: Generalized Autoregressive Pretraining for Language Understanding.
[p]XLM: Cross-Lingual Language Model.

VAEM-Mine Method

The classification models were the final component of a pipeline named the VAEM-Mine method (Figure 1), consisting of processes that started with data collection and cleaning, followed by processing through topic models to filter for data that were as close as possible to the VAEM, and then, a focused binary classification approach for isolating VAEM.

The method included decision points to determine the appropriate direction, either the training process or the application of the trained models to incoming data. At the beginning of the topic modeling phase, a trained model did not exist; thus, the work of training the topic models began. The first step was to label some examples of the subject of interest (in this case, VAEM) and additional examples of other subjects. This enabled the application of a topic modeling scoring, which measured how the VAEM-label of interest was distributed in the topics, compared with other labeled topics. A topic model was considered to score well if the VAEM were concentrated in only a few topics, and ideally in only 1 topic, with minimum data belonging to the other labels. Further refinement of the data was possible by a second stage of topic modeling on the data obtained from the top model of the first stage. The second stage identified topics that had a high ratio of VAEM to other subjects in the texts, but at the expense of losing some texts containing VAEM. Having trained the models, they could be applied to filter the incoming data, and it was up to the user whether they take only the output of the best topic (or topics) of the first-stage topic model or further refine the data by taking it from selected topics of the second-stage topic model. The topics of the first stage of topic modeling were also potentially useful to obtain a domain taxonomy.

Figure 1. The vaccine adverse event mention–mine method. CNN: Convolutional Neural Network; LSTM: Long Short-Term Memory.



The filtered data were handled by the classification phase, which also had the decision point for either training classifiers or using trained classifiers. When training, the choice of classifiers should relate to the quantity of available data, and if results are not as expected, a decision may be made to obtain more data. The method required the incoming filtered data to be labeled for the creation of data sets suitable to train the classifiers. It additionally required the creation of domain-specific embeddings. The VAEM-Mine method can be adopted as a workflow to tackle any similar task of identifying personal health mentions.

Results

Classification Analysis

Classification training and evaluation was conducted twice; first, with the filtered data that were obtained from applying topic modeling to the initial phase of data collection and then, with the data obtained through topic model filtering over all the

collected data. The following sections describe these as Phase-One and Phase-Two classification.

Phase-One Classification

The first phase of classification experiments used a training set of 2639 records, a validation set of 880 records, and the imbalanced holdout Phase-One Test data set of 614 tweets. The F_1 scores for the models evaluated in this phase are listed in Table 4.

Table 4. Phase-One F1 scores.

Model	Validation	Imbalanced test	Balanced test	Combined test
CNN[a]-BiGRU[b]	0.842	0.762	0.846	0.825
BERT[c]	N/A[d]	0.767	0.841	0.824
BiGRU	0.807	0.793	0.828	0.822
CNN–LSTM[e]	0.805	0.777	0.815	0.808
BiLSTM[f]	0.815	0.807	0.807	0.807
GRU[g]	0.820	0.730	0.822	0.804
CNN-BiLSTM	0.816	0.766	0.810	0.802
CNN	0.816	0.787	0.800	0.798
LSTM	0.796	0.767	0.803	0.796
Ensemble	0.815	0.726	0.829	0.810
Logistic Regression CV[h]	0.812	0.730	0.820	0.803
Linear SVC[i]	0.814	0.693	0.824	0.797
SGD[j]	0.805	0.636	0.825	0.785
Naïve Bayes SVM[k]	0.792	0.767	0.789	0.785
Random Forest	0.814	0.694	0.801	0.779
Extra Trees	0.833	0.688	0.801	0.777
XGBoost[l]	0.811	0.704	0.791	0.774
Naïve Bayes	0.798	0.605	0.799	0.756

[a]CNN: Convolutional Neural Network.
[b]BiGRU: Bidirectional Gated Recurrent Unit.
[c]BERT: Bidirectional Encoder Representations.
[d]N/A: not applicable.
[e]LSTM: Long Short-Term Memory.
[f]BiLSTM: Bidirectional Long Short-Term Memory.
[g]GRU: Gated Recurrent Unit.
[h]CV: Cross Validation.
[i]SVC: Support Vector Classification.
[j]SGD: Stochastic Gradient Descent.
[k]SVM: Support Vector Machine.
[l]XGBoost: Extreme Gradient Boosting.

Table 4 includes subsequent tests of the models against the Phase-Two *Balanced test* data set and a *Combined Test* data set that uses all the test data. F_1 scores were measured for the positive, VAEM class, rather than for both classes. The models are arranged in order of the best F_1 score over the test data sets; validation scores are also included, where available. Validation F_1 scores are not available for models using transfer learning—they used a cross-validation approach, and thus, were given combined training and validation data and were evaluated only against test data sets.

The Ensemble model shown in the middle of Table 4 was scored based on a maximum voting of the predictions of 5 traditional classifiers on the test data set—consisting of the Naïve Bayes Support Vector Machine, Linear Regression Cross Validation, Stochastic Gradient Descent, Linear Support Vector Classification, and Random Forest classifiers. It had the overall best score among the traditional classifiers on the large test data.

All the deep learning models outperformed the best traditional classifier on the *Imbalanced Test* data set, by at least 6% and almost as much as 10%—the improvement was mostly owing to great capacity to correctly distinguish non–VAEM-related tweets, and thus obtain a greater precision. However, when evaluated against the *Balanced* and *Combined Test* sets, the results differed—here, the traditional classifiers outperformed many of the deep learning models, especially the Ensemble, which was only surpassed by the top 3 deep learning models.

Phase-Two Classification

The second phase of classification used 5 times as many records to train the models, by combining the 3519 training records from the first phase with another 15,730 records, resulting in a total of 19,249. Phase Two also introduced a large, more balanced test data set of 828 records. The greater amount of data allowed a proper assessment of neural networks, but it also improved model performance across the board (Table 5). The *imbalanced change* and *combined change* columns show the percentage increase in the models' F_1 score over the *Imbalanced Test* and *Combined Test* data sets, compared with their Phase-One equivalents.

There was a much greater consistency of scoring over all the test data sets, and the top models scored best over all the test data sets. The highest score was from the Robustly Optimized Bidirectional Encoder Representations Pretraining Approach (RoBERTa) Large Transformer model, with an F_1 score of 0.919 on the Imbalanced data set; the standard RoBERTa model was placed second.

One of the most noteworthy effects of having more data was that the previously strong combinations of CNN with Bidirectional Gated Recurrent Unit and Bidirectional LSTM models were surpassed by the LSTM on the *Imbalanced Test* data set, both when combined with a CNN but most significantly as a stand-alone model. The LSTM in fifth position on the imbalanced test scoring was only 2.5% behind the score of the RoBERTa Large model. One can fairly conclude that a CNN or hybrid CNN approach performs well when limited data are available but will likely be surpassed by architectures designed for sequential language processing as more data become available.

A detailed analysis of the classifiers' performance is provided in Multimedia Appendix 2.

Table 5. Phase-Two F1 scores.

Model	Validation	Imbalanced test	Balanced test	Combined test	Imbalanced change, %	Combined change, %
RoBERTa[a] Large	N/A[b]	0.919	0.908	0.910	—[c]	—
RoBERTa	N/A	0.901	0.905	0.904	—	—
XLNet[d] Large	N/A	0.884	0.906	0.902	—	—
XLNet	N/A	0.870	0.903	0.897	—	—
XLM[e]	N/A	0.910	0.894	0.897	—	—
BERT[f]	N/A	0.863	0.892	0.887	12.6	7.7
BiGRU[g]	0.877	0.855	0.896	0.890	7.9	8.2
CNN[h]-BiGRU	0.874	0.849	0.890	0.884	11.4	7.1
LSTM[i]	0.866	0.875	0.879	0.878	14.1	10.3
CNN-LSTM	0.866	0.862	0.876	0.873	10.9	8.1
BiLSTM[j]	0.872	0.847	0.884	0.878	5	8.8
GRU[k]	0.869	0.825	0.876	0.868	13.1	7.9
CNN-BiLSTM	0.872	0.824	0.879	0.871	7.6	8.6
CNN	0.864	0.805	0.866	0.856	2.4	7.2
Ensemble	0.870	0.818	0.874	0.865	12.6	6.8
Logistic RCV[l]	0.866	0.807	0.873	0.861	10.5	7.3
SGD[m]	0.865	0.806	0.873	0.861	26.7	9.7
Linear SVC[n]	0.864	0.802	0.869	0.857	15.7	7.5
Random Forest	0.857	0.796	0.864	0.853	14.7	9.5
Extra Trees	0.857	0.789	0.862	0.849	14.7	9.2
NB[o] SVM[p]	0.838	0.798	0.838	0.832	3.9	5.9
XGBoost[q]	0.845	0.714	0.854	0.831	1.3	7.4
NB	0.835	0.735	0.841	0.822	21.5	8.7

[a]RoBERTa: Robustly Optimized Bidirectional Encoder Representations Pretraining Approach.

[b]N/A: not applicable.

[c]Change calculation was not performed because no previous figures existed.

[d]XLNet: Generalized Autoregressive Pretraining for Language Understanding.

[e]XLM: Cross-Lingual Language Model.

[f]BERT: Bidirectional Encoder Representations.

[g]BiGRU: Bidirectional Gated Recurrent Unit.

[h]CNN: Convolutional Neural Network.

[i]LSTM: Long Short-Term Memory.

[j]BiLSTM: Bidirectional Long Short-Term Memory.

[k]GRU: Gated Recurrent Unit.

[l]RCV: Regression Cross Validation.

[m]SGD: Stochastic Gradient Descent.

[n]SVC: Support Vector Classification.

[o]NB: Naïve Bayes.

[p]SVM: Support Vector Machine.

[q]XGBoost: eXtreme Gradient Boosting.

VAEM-Mine Method Performance

Here, we assess the overall effectiveness of the method, regarding the quantities of tweets having VAEMs that were progressively filtered out by the method. The values presented are the total numbers of tweets collected and processed via the method, with estimates where appropriate.

Topic Modeling Phase

Table 6 depicts the numbers obtained from after data collection to the completion of the topic modeling. From the original 811,010 records, 122,653 (15.12%) records were removed by data cleaning, and topic modeling was used to process 688,357 (84.87%) records. Stage 1 of topic modeling filtered out 82.86% (570,383/688,357) of the records to retain 17.14% (117,974/688,357) of the records likely to contain VAEM. The data were approximately 14.55% (117,974/811,010) of the original total and contained >99% of all the available VAEM (Multimedia Appendix 3).

Table 6. Summary of topic modeling counts (N=811,010).

Steps	Counts, n (% of initial data)
Tweets collected	811,010 (100)
Cleaned	–122,653 (–15.12)
Tweets after cleaning	688,357 (84.88)
Discarded (stage 1)	–570,383 (–70.33)
Tweets after stage 1	117,974 (14.55)
Discarded (stage 2)	–19,083 (–2.35)
Tweets after stage 2[a,b]	98,891 (12.19)

[a]Stage 2 proportions—non–vaccine adverse event mention: 88,900 and vaccine adverse event mention: 9991 (10.10% of stage 2 data; 1.45% of tweets after cleaning; 1.23% of initial data).

[b]Vaccine adverse event mention proportions—in other stage 2 topics: 2367 and in best stage 2 topic: 7624 (76.31% of vaccine adverse event mention).

To prepare for the first round of classification, additional 19,083 records were discarded—those which were not in the top 3 topics of the stage 2 topic model. Subsequent labeling of the discarded topic most likely to contain VAEM (based on the distribution of topic model labels) showed only 1.49% (94/6274) of VAEM in the data, which was approximately 5.15% (94/1826) of the VAEM in the first round.

For the second round of classification, all the records that were identified as likely VAEM by the topic model were retained. The resulting 12.19% (98,891/811,010) records retained over both rounds of topic modeling were labeled, and VAEM were found to be 10.10% (9991/98,891) of the retained data. The stage 2 topic models' topic numbers were assessed, and it was found that the best stage 2 topic of 14,498 tweets contained 76.31% (7624/9991) of the retained VAEM, and there were approximately 11.10% (7624/6874) more VAEM than non-VAEM in the topic.

From these figures, we conclude that topic modeling is an effective filtering mechanism, as it identified approximately all the VAEM, while removing a lot of unwanted data. The filtered data were more manageable for labeling for classification than it would have otherwise been, and if needed, the filtered output of the stage 2 topic model can be used as it is, with the understanding that it discards some VAEM and still contains a small but similar number of non-VAEM. However, as discussed previously, classification is a more precise final step to obtain VAEM from the filtered records.

Classification Phase

To assess classifier effectiveness regarding the total data, the recall and precision of the best classifier, the RoBERTa Large model, were applied to the total VAEM to obtain an *estimate* of its performance on the total VAEM. These were a precision score of 0.874 and a recall score of 0.948 for the combined test data:

1. Applying the recall score of 0.948 to the total 9991 VAEM-containing tweets, we estimate that 94.81% (9472/9991) of the VAEM tweets would be correctly classified and 5.19% (519/9991) of the VAEM would be missed.
2. We find that 1.54% (1370/88,900) of the non-VAEM tweets would be added to the 9472 tweets to match to the precision score of 0.874 (9472/10,842).
3. These results of 94.81% (9472/9991) of VAEM together with 1.54% (1370/88,900) of the non-VAEM in the predicted positive class were clearly superior to those obtained with the best topic of stage 2 topic modeling, where we saw the proportion of VAEM in the best topic was 76.31% (7624/9991) and the almost equal number of non-VAEM in the topic was approximately 7.70% (6847/88,900) of the non-VAEM.

Combined Topic Modeling and Classification Effectiveness

By measuring the combined effectiveness of topic modeling and classification, the following results are estimated:

1. As explained in Multimedia Appendix 3, counts of VAEM identified via topic modelling were estimated to be 99% of all likely VAEM; therefore, with 99% being represented as a count of 9991 VAEM, it is estimated that 10,090 VAEM originally existed.

2. A total of 8992 VAEM are estimated to be identified via the combined effects of cleaning, topic modelling, and classification from the original 811,010 records, being at least 89.11% (8992/10,090) of all likely VAEM and 1.11% (8992/811,010) of the original data.
   - A total of 98.89% (802,018/811,010) of the data were eliminated through cleaning, topic modeling, and classification.
   - Totally, around 11% (1098/10,090) of the VAEM were also eliminated during this processing; the attrition is a consequence of the filtering and classification required to capture the estimated 89.12% (8992/10,090).
3. Overall, 98.89% (802,018/811,010) of data were eliminated as not containing VAEM, with a very small amount misidentified, to identify 1.11% (8992/811,010) of the data as having VAEM, with 90% success.

The results indicate that the combined approach of topic modeling followed by classification effectively identifies and isolates VAEMs from approximately all other vaccine-related Twitter posts. The VAEM-Mine method enables us to identify the most effective topic models and classifiers for the core task of isolating VAEM. In particular, the key to the method's success is the topic modeling phase, which drastically reduces the amount of irrelevant data and thus delivers manageable data to the classification phase. As natural language processing technologies improve and new topic models and classifiers can be introduced, we assume that even these results will improve.

Discussion

The key objective of this study was to contribute to research on vaccine safety surveillance, by illustrating that social media monitoring has the potential to augment existing surveillance systems. We have demonstrated a topic modeling and classification VAEM-Mine method for identifying VAEM with high degree of sensitivity and specificity following vaccination.

Principal Findings

The VAEM-Mine method approached the problem of finding sparse VAEMs by using topic modeling followed by classification. Topic modeling identified texts based on their semantic and syntactic nature. Then, it was used to extract those tweets that predominantly describe personal health issues in relation to vaccines. Classification identified VAEMs from the filtered texts with high degree of accuracy. Neither of the machine learning components were explicitly trained on specific reaction keywords, instead they identified texts owing to their innate capacity to detect patterns in language structure.

Other studies on detecting influenza [23] and COVID-19 [24] have required purpose-built machine learning classifiers that identify specific adverse event reactions from tweets. Their classifiers were trained to identify known reaction keywords derived from medical databases. Our approach relies on language features of the tweets to elicit the likely cohort and the power of modern transformer classifiers to determine the true signals. By tackling the problem of finding adverse events through the lens of the language used in personal health mentions, we conclude that social media can provide a wealth of useful data.

The VAEM-Mine method has significant capability to successively isolate VAEMs from the massive amount of other vaccine-related Twitter posts. The topic modeling phase could isolate up to 99.02% (9991/10,090 [estimated]) of the Twitter posts that contained VAEM. The data identified by Stage 1 topic modelling as likely containing VAEM were only 14.55% (117,974/811,010) of the original data, thereby eliminating 85.45% (693,306/811,010) of mostly irrelevant posts. The classification phase identified 8992 (90%) of the 9991 VAEM with an F_1 score of 0.91. The combination of topic modelling and classification resulted in the identification of 89.12% (8992/10,090 [estimated]) of the VAEM.

Training the topic modeling component of the method is enabled by identifying the most effective topic models by using F_1 scoring over a small number of labeled posts—the scoring identifies when topic models are most effective at grouping labeled VAEM into a topic. The topic modeling scoring method is an important contribution of this study.

This study also presents detailed reporting, including comparisons, on a range of classification models, including traditional machine learning models and deep neural (deep learning) networks. Their effectiveness was measured against different-sized data sets, emulating data sizes that are likely to be available to other researchers [43], and we used charts (Multimedia Appendix 2) to illustrate how the amount of training data affects model recall and precision.

Limitations

There are unavoidable issues and potential biases that result from using any social media data. A limitation of this study is the use of only English-language tweets as data source; the approach needs to be validated by using other social media data sources and other languages. Although the data collection for this study spanned a year and included some potential trend patterns during influenza seasons, a long-term data collection would be better for any analysis of trends. At the time of the study, a full year's data were required to properly train and evaluate the classifiers—this was in part because of the limited pipeline of the Twitter application program interface and because data collection was from a period before the COVID-19 pandemic and signals were correspondingly less frequent compared with those found during the COVID-19 vaccines rollout.

However, the proposed VAEM-Mine method can identify VAEM with F_1 score of 0.91 and is applicable to any similar problem of detecting personal health mentions in social media posts based on the language of conversations.

Conclusions and Future Research

We have determined that the VAEM-Mine method is an effective approach for both identifying and applying the topic models and classifiers that, when combined, can filter out the vast amount of irrelevant vaccine-related conversations and isolate VAEMs.

A key contribution of this study is that appropriately scored topic modeling is highly effective for identifying social posts that might contain VAEM. The technique of F_1 scoring of topic models based on a small number of labeled posts, identified in this study, is practical and easily implementable and can be used by other researchers to assist with identifying topic models that group texts on specific language features.

The volume of social media posts regarding the current COVID-19 pandemic is immense, but those that are related to personally experiencing illness owing to the virus or vaccines are a small portion of these; however, they contain similar language. Currently, we are applying the VAEM-Mine method to both internally gathered and published [44] COVID-19 vaccine–related Twitter data sets to examine trends in VAEM reporting. There are several ways in which the identified VAEM posts can be used for vaccine safety signal detection. Among them are (1) examining individual posts by domain experts; (2) further classifying the posts to identify adverse events of special interest, which include vascular, neurological, or allergic disorders and enhanced disease; and (3) measuring changes of post volumes that might indicate unfolding events.

This paper interprets the success of the VAEM-Mine method in terms of percentages of data captured by the method and compares classifiers in terms of F_1 scores. Future studies can analyze the method's success in terms of model explainability [45].

Acknowledgments

The authors would like to thank Christopher Palmer for providing technical advice for the project. This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Model definitions and parameters.
[DOCX File , 18 KB - medinform_v10i6e34305_app1.docx]

Multimedia Appendix 2
Classification performance analysis.
[DOCX File , 4495 KB - medinform_v10i6e34305_app2.docx]

Multimedia Appendix 3
Verification of best topic model.
[DOCX File , 94 KB - medinform_v10i6e34305_app3.docx]

References

1. Milstien JB, Batson A, Wertheimer AI. Vaccines and drugs: characteristics of their use to meet public health goals. Health, Nutrition, and Population, The World Bank. 2015. URL: http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2005/04/14/000090341_20050414151834/Rendered/PDF/320400MilstienVaccinesDrugsFinal.pdf [accessed 2022-05-12]
2. Budhiraja S, Akinapelli R. Pharmacovigilance in vaccines. Indian J Pharmacol 2010 Apr;42(2):117 [FREE Full text] [doi: 10.4103/0253-7613.64488] [Medline: 20711383]
3. Almenoff J, Tonning JM, Gould AL, Szarfman A, Hauben M, Ouellet-Hellstrom R, et al. Perspectives on the use of data mining in pharmaco-vigilance. Drug Saf 2005;28(11):981-1007. [doi: 10.2165/00002018-200528110-00002] [Medline: 16231953]
4. Agmon-Levin N, Paz Z, Israeli E, Shoenfeld Y. Vaccines and autoimmunity. Nat Rev Rheumatol 2009 Nov;5(11):648-652. [doi: 10.1038/nrrheum.2009.196] [Medline: 19865091]
5. Griffin MR, Braun MM, Bart KJ. What should an ideal vaccine postlicensure safety system be? Am J Public Health 2009 Oct;99 Suppl 2:S345-S350. [doi: 10.2105/AJPH.2008.143081] [Medline: 19797747]
6. Chen RT, Shimabukuro TT, Martin DB, Zuber PL, Weibel DM, Sturkenboom M. Enhancing vaccine safety capacity globally: a lifecycle perspective. Vaccine 2015 Nov 27;33 Suppl 4(0 4):D46-D54 [FREE Full text] [doi: 10.1016/j.vaccine.2015.06.073] [Medline: 26433922]
7. Mesfin YM, Cheng AC, Enticott J, Lawrie J, Buttery JP. Use of telephone helpline data for syndromic surveillance of adverse events following immunization in Australia: a retrospective study, 2009 to 2017. Vaccine 2020 Jul 22;38(34):5525-5531. [doi: 10.1016/j.vaccine.2020.05.078] [Medline: 32593607]
8. Härmark L, van Grootheest AC. Pharmacovigilance: methods, recent developments and future perspectives. Eur J Clin Pharmacol 2008 Aug;64(8):743-752. [doi: 10.1007/s00228-008-0475-9] [Medline: 18523760]

9. Clothier HJ, Crawford N, Russell MA, Buttery JP. Allergic adverse events following 2015 seasonal influenza vaccine, Victoria, Australia. Euro Surveill 2017 May 18;22(20):30535 [FREE Full text] [doi: 10.2807/1560-7917.ES.2017.22.20.30535] [Medline: 28552101]
10. Pal SN, Duncombe C, Falzon D, Olsson S. WHO strategy for collecting safety data in public health programmes: complementing spontaneous reporting systems. Drug Saf 2013 Feb;36(2):75-81 [FREE Full text] [doi: 10.1007/s40264-012-0014-6] [Medline: 23329541]
11. Conway M, Hu M, Chapman WW. Recent advances in using natural language processing to address public health research questions using social media and consumergenerated data. Yearb Med Inform 2019 Aug;28(1):208-217 [FREE Full text] [doi: 10.1055/s-0039-1677918] [Medline: 31419834]
12. Paul MJ, Dredze M. Social Monitoring for Public Health. In: Dredze M, Paul MJ, editors. Synthesis Lectures on Information Concepts, Retrieval, and Services. Williston, VT, USA: Morgan and Claypool Publishers; Aug 31, 2017:1-183.
13. Clothier HJ, Lawrie J, Russell MA, Kelly H, Buttery JP. Early signal detection of adverse events following influenza vaccination using proportional reporting ratio, Victoria, Australia. PLoS One 2019 Nov 1;14(11):e0224702 [FREE Full text] [doi: 10.1371/journal.pone.0224702] [Medline: 31675362]
14. Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N, et al. Adverse drug reaction identification and extraction in social media: a scoping review. J Med Internet Res 2015 Jul 10;17(7):e171 [FREE Full text] [doi: 10.2196/jmir.4304] [Medline: 26163365]
15. Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 2011 Oct;7(10):e1002199 [FREE Full text] [doi: 10.1371/journal.pcbi.1002199] [Medline: 22022249]
16. Larson HJ, Smith DM, Paterson P, Cumming M, Eckersberger E, Freifeld CC, et al. Measuring vaccine confidence: analysis of data obtained by a media surveillance system used to analyse public concerns about vaccines. Lancet Infect Dis 2013 Jul;13(7):606-613. [doi: 10.1016/S1473-3099(13)70108-7] [Medline: 23676442]
17. Du J, Xu J, Song HY, Tao C. Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. BMC Med Inform Decis Mak 2017 Jul 05;17(Suppl 2):69 [FREE Full text] [doi: 10.1186/s12911-017-0469-6] [Medline: 28699569]
18. Lama Y, Hu D, Jamison A, Quinn SC, Broniatowski DA. Characterizing trends in human papillomavirus vaccine discourse on Reddit (2007-2015): an observational study. JMIR Public Health Surveill 2019 Mar 27;5(1):e12480 [FREE Full text] [doi: 10.2196/12480] [Medline: 30916662]
19. Radzikowski J, Stefanidis A, Jacobsen KH, Croitoru A, Crooks A, Delamater PL. The measles vaccination narrative in Twitter: a quantitative analysis. JMIR Public Health Surveill 2016 Jan 4;2(1):e1 [FREE Full text] [doi: 10.2196/publichealth.5059] [Medline: 27227144]
20. Mollema L, Harmsen IA, Broekhuizen E, Clijnk R, De Melker H, Paulussen T, et al. Disease detection or public opinion reflection? Content analysis of tweets, other social media, and online newspapers during the measles outbreak in The Netherlands in 2013. J Med Internet Res 2015 May 26;17(5):e128 [FREE Full text] [doi: 10.2196/jmir.3863] [Medline: 26013683]
21. Surian D, Nguyen DQ, Kennedy G, Johnson M, Coiera E, Dunn AG. Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. J Med Internet Res 2016 Aug 29;18(8):e232 [FREE Full text] [doi: 10.2196/jmir.6045] [Medline: 27573910]
22. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform 2015 Feb;53:196-207 [FREE Full text] [doi: 10.1016/j.jbi.2014.11.002] [Medline: 25451103]
23. Wang J, Zhao L, Ye Y. Semi-supervised multi-instance interpretable models for flu shot adverse event detection. In: Proceedings of the 2018 IEEE International Conference on Big Data. 2018 Presented at: BigData '18; December 10-13, 2018; Seattle, WA, USA p. 851-860. [doi: 10.1109/bigdata.2018.8622434]
24. Lian AT, Du J, Tang L. Using a machine learning approach to monitor COVID-19 Vaccine Adverse Events (VAE) from Twitter data. Vaccines (Basel) 2022 Jan 11;10(1):103 [FREE Full text] [doi: 10.3390/vaccines10010103] [Medline: 35062764]
25. Khademi Habibabadi S, Haghighi PD. Topic modelling for identification of vaccine reactions in Twitter. In: Proceedings of the Australasian Computer Science Week Multiconference. 2019 Presented at: ACSW '19; January 29-31, 2019; Sydney, Australia p. 1-10. [doi: 10.1145/3290688.3290735]
26. Zhai CX, Massung S. Text Data Management and Analysis. San Rafael, CA, USA: Morgan & Claypool Publishers; Jun 30, 2016:88-94.
27. 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(2011):2825-2830 [FREE Full text] [doi: 10.1007/978-1-4842-5373-1_1]
28. sklearn.feature_extraction.text.TfidfTransformer — scikit-learn 0.24.2 documentation. scikit-learn. 2021. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html [accessed 2021-05-23]
29. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of the International Conference on Learning Representations. 2013 Jan 16 Presented at: ICLR '13; May 2-4, 2013; Scottsdale, AZ, USA URL: https://arxiv.org/abs/1301.3781v3

30. Řehůřek R, Sojka P. Software framework for topic modeling with large corpora. In: Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. 2010 Presented at: LREC '10; May 22, 2010; Valletta, Malta p. 46-50 URL: http://www.muni.cz/research/publications/884893
31. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016 Presented at: ACL '16; August 7-12, 2016; Berlin, Germany p. 1715-1725. [doi: 10.18653/v1/p16-1162]
32. sklearn.linear_model. scikit-learn. 2022. URL: https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model [accessed 2022-05-25]
33. sklearn.svm.SVC. Scikit-learn. 2022. URL: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html# [accessed 2022-05-25]
34. sklearn.ensemble. Scikit-learn. 2022. URL: https://scikit-learn.org/stable/modules/classes.html#module-sklearn.ensemble [accessed 2022-05-25]
35. sklearn.naive_bayes. Scikit-learn. 2022. URL: https://scikit-learn.org/stable/modules/classes.html#module-sklearn.naive_bayes [accessed 2022-05-25]
36. Chin J. NBSVM. GitHub. 2012. URL: https://github.com/Joshua-Chin/nbsvm [accessed 2022-06-02]
37. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD '16; August 13-17, 2016; San Francisco, CA, USA p. 785-794. [doi: 10.1145/2939672.2939785]
38. sklearn.ensemble.VotingClassifier. Scikit-learn. 2022. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html#sklearn.ensemble.VotingClassifier [accessed 2022-05-25]
39. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of the Advances in Neural Information Processing Systems 32. 2019 Presented at: NeurIPS '19; December 8 - 14, 2019; Vancouver, Canada.
40. Shawn1993/cnn-text-classification-pytorch: CNNs for Sentence Classification. GitHub. 2020 Oct 14. URL: https://github.com/Shawn1993/cnn-text-classification-pytorch [accessed 2022-02-07]
41. bamtercelboo / cnn-lstm-bilstm-deepcnn-clstm-in-pytorch – In PyTorch Learing Neural Networks Likes CNN(Convolutional Neural Networks for Sentence Classification (Y.Kim, EMNLP 2014)、LSTM、BiLSTM、DeepCNN、CLSTM、CNN and LSTM. GitHub. 2019 Apr 23. URL: https://github.com/bamtercelboo/cnn-lstm-bilstm-deepcnn-clstm-in-pytorch [accessed 2022-02-07]
42. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's transformers: state-of-the-art natural language processing. arXiv (forthcoming) 2019 Oct 9 [FREE Full text] [doi: 10.18653/v1/2020.emnlp-demos.6]
43. Magge A, Klein A, Miranda-Escalada A, Al-Garadi MA, Alimova I, Miftahutdinov Z, et al. Overview of the Sixth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at NAACL 2021. In: Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task. 2021 Presented at: NAACL '21; June 10, 2021; Mexico City, Mexico p. 21-32. [doi: 10.18653/v1/2021.smm4h-1.4]
44. DeVerna MR, Pierri F, Truong BT, Bollenbacher J, Axelrod D, Loynes N, et al. CoVaxxy: a collection of English-language Twitter posts about COVID-19 vaccines. In: Proceedings of the 15th International AAAI Conference on Web and Social Media. 2021 Jan Presented at: AAAI '21; June 7-10, 2021; Virtual p. 992-999.
45. Burkart N, Huber MF. A survey on the explainability of supervised machine learning. J Artif Intell Res 2021 Jan 19;70:245-317. [doi: 10.1613/jair.1.12228]

Abbreviations

AEFI: adverse events following immunization
CNN: Convolutional Neural Network
LSTM: Long Short-Term Memory
RoBERTa: Robustly Optimized Bidirectional Encoder Representations Pretraining Approach
VAEM: vaccine adverse event mention

Edited by C Lovis; submitted 16.10.21; peer-reviewed by H Ayatollahi, F Velayati, M Elbattah, D Huang; comments to author 02.01.22; revised version received 22.02.22; accepted 11.04.22; published 16.06.22.

Please cite as:
Khademi Habibabadi S, Delir Haghighi P, Burstein F, Buttery J
Vaccine Adverse Event Mining of Twitter Conversations: 2-Phase Classification Study
JMIR Med Inform 2022;10(6):e34305
URL: https://medinform.jmir.org/2022/6/e34305
doi:10.2196/34305
PMID:35708760

Original Paper

Predicting 30-Day Readmission Risk for Patients With Chronic Obstructive Pulmonary Disease Through a Federated Machine Learning Architecture on Findable, Accessible, Interoperable, and Reusable (FAIR) Data: Development and Validation Study

Celia Alvarez-Romero[1], MSc; Alicia Martinez-Garcia[1], PhD; Jara Ternero Vega[2], MSc; Pablo Díaz-Jimènez[2], MSc; Carlos Jimènez-Juan[2], MSc; María Dolores Nieto-Martín[2], PhD; Esther Román Villarán[1], MSc; Tomi Kovacevic[3,4], PhD; Darijo Bokan[3], PhD; Sanja Hromis[3,4], PhD; Jelena Djekic Malbasa[3,4], PhD; Suzana Beslać[3], MD; Bojan Zaric[3,4], PhD; Mert Gencturk[5], MSc; A Anil Sinaci[5], PhD; Manuel Ollero Baturone[2], PhD; Carlos Luis Parra Calderón[1], MSc

[1]Computational Health Informatics Group, Institute of Biomedicine of Seville, Virgen del Rocío University Hospital, Consejo Superior de Investigaciones Científicas, University of Seville, Seville, Spain

[2]Internal Medicine Department, Virgen del Rocío University Hospital, Seville, Spain

[3]Institute for Pulmonary Diseases of Vojvodina, Sremska Kamenica

[4]Medical Faculty, University of Novi Sad, Novi Sad

[5]Software Research & Development and Consultancy Corporation, Ankara, Turkey

Corresponding Author:
Celia Alvarez-Romero, MSc
Computational Health Informatics Group
Institute of Biomedicine of Seville, Virgen del Rocío University Hospital
Consejo Superior de Investigaciones Científicas, University of Seville
Avda Manuel Siurot s/n
Seville
Spain
Phone: 34 955013313
Email: celia.alvarez@juntadeandalucia.es

Abstract

Background: Owing to the nature of health data, their sharing and reuse for research are limited by legal, technical, and ethical implications. In this sense, to address that challenge and facilitate and promote the discovery of scientific knowledge, the Findable, Accessible, Interoperable, and Reusable (FAIR) principles help organizations to share research data in a secure, appropriate, and useful way for other researchers.

Objective: The objective of this study was the FAIRification of existing health research data sets and applying a federated machine learning architecture on top of the FAIRified data sets of different health research performing organizations. The entire FAIR4Health solution was validated through the assessment of a federated model for real-time prediction of 30-day readmission risk in patients with chronic obstructive pulmonary disease (COPD).

Methods: The application of the FAIR principles on health research data sets in 3 different health care settings enabled a retrospective multicenter study for the development of specific federated machine learning models for the early prediction of 30-day readmission risk in patients with COPD. This predictive model was generated upon the FAIR4Health platform. Finally, an observational prospective study with 30 days follow-up was conducted in 2 health care centers from different countries. The same inclusion and exclusion criteria were used in both retrospective and prospective studies.

Results: Clinical validation was demonstrated through the implementation of federated machine learning models on top of the FAIRified data sets from different health research performing organizations. The federated model for predicting the 30-day hospital readmission risk was trained using retrospective data from 4.944 patients with COPD. The assessment of the predictive model was performed using the data of 100 recruited (22 from Spain and 78 from Serbia) out of 2070 observed (records viewed) patients during the observational prospective study, which was executed from April 2021 to September 2021. Significant accuracy

(0.98) and precision (0.25) of the predictive model generated upon the FAIR4Health platform were observed. Therefore, the generated prediction of 30-day readmission risk was confirmed in 87% (87/100) of cases.

Conclusions: Implementing a FAIR data policy in health research performing organizations to facilitate data sharing and reuse is relevant and needed, following the discovery, access, integration, and analysis of health research data. The FAIR4Health project proposes a technological solution in the health domain to facilitate alignment with the FAIR principles.

(JMIR Med Inform 2022;10(6):e35307) doi:10.2196/35307

KEYWORDS

FAIR principles; research data management; clinical validation; chronic obstructive pulmonary disease; privacy-preserving distributed data mining; early predictive model

Introduction

Overview

FAIR4Health is a project that received funding from the European Union's (EU) Horizon 2020 research and innovation program under grant 824666. This project started in December 2018 and ended in November 2021. The main objective of this European project was to promote and encourage the EU health research community to apply the Findable, Accessible, Interoperable, and Reusable (FAIR) principles [1] in their data sets derived from publicly funded research initiatives through the implementation of an effective outreach strategy at the EU level, the production of a set of guidelines to set the foundations for a FAIR data certification road map, the development of an intuitive platform, and the demonstration of the potential impact on health research and health outcomes through the validation of 2 pathfinder case studies. At a high level, this project aimed to facilitate health research data sharing and reuse. This project brought together expertise from the key stakeholders involved in properly addressing this main objective: health research, data managers, medical informatics, software developers, standards, and lawyers. The FAIR4Health Consortium accounted for 17 partners from 11 EU and non-EU countries.

Despite strong concerns and challenges regarding data sharing in health research [2,3] and following efforts to distinguish between the concepts of *open data* [4,5] and *FAIR data* [6,7], it is evident that data sharing is one of the pillars of scientific progress. Cooperation between different countries and cultures is the fastest way to gather valuable knowledge and address challenges such as the current pandemic [8,9]. Given the strong global focus on scientific research and international cooperation, the adoption and implementation of a FAIR data policy in health research organizations is a strong requirement. Therefore, the implementation of FAIR data initiatives and lessons learned in the FAIRification process in the health field is paramount to support evidence-based clinical practice and research transparency in the era of big data and open research publishing [10].

The purpose of the FAIR4Health project [11] was to design a workflow [12] and develop a framework to reach the FAIRification of health research data sets addressing the relevant legal, technical, and ethical considerations and requirements of sensitive data. For that, FAIR4Health FAIRification tools were implemented and deployed in different health research performing organizations of the FAIR4Health Consortium. Then, 2 pathfinder case studies were carried out to demonstrate the potential impact of the application of a FAIR strategy on health outcomes and health and social care research, making use of a privacy-preserving distributed data mining (PPDDM) architecture implemented on the FAIR4Health platform. The PPDDM architecture used a federated machine learning approach in which health research data do not leave its premises while the models travel between the data-hosting sites. The performance and validation of the FAIR4Health use case that was focused on the development of an early predictive model for 30-day readmission risk in patients with chronic obstructive pulmonary disease (COPD) is described in this paper.

Background

FAIR Data Principles

The aim of the FAIR data principles [1] is to ensure that data are shared in a way that enables and enhances reuse by humans and machines. Although FAIR data emerged from a workshop for the life science community, the FAIR principles are intended to be applied to data and metadata from all disciplines.

Since its formal release via the FORCE11 community [13], the FAIR data principles have been adopted by several funders and governments worldwide. The European Commission's data management guidelines were updated in 2017 to introduce the FAIR principles. Similarly, following the summit in June 2017, the European Open Science Cloud Declaration was launched [14]. In contrast, the recent staff working document proposed an implementation road map for the European Open Science Cloud [15]. These 2 relevant documents emphasize the central role of FAIR data.

FAIR principles are being adopted by a diverse range of research disciplines, such as economics, semantic web, and environment. Several groups have assessed the uptake to date and the challenges encountered. FAIR4Health [11] and other projects add to the state-of-the-art by documenting good practices and applying them to other domains, where possible, such as the medical domain.

FAIR4Health adds to the analysis and experience of the application of FAIR principles in the health research field, specifically in health research data sets on COPD.

COPD and Readmissions

COPD is a respiratory disease characterized by persistent symptoms and chronic limitation of airflow. This disease is known to be underdiagnosed even though it affects almost 10%

of adults worldwide [16] and its prevalence continues to increase with the aging of the population. The study by Mannino et al [17] showed that >50% of adults with impaired lung function were unaware that they were diagnosed with COPD [17]. COPD frequently presents itself with other comorbidities, such as cardiovascular disease, hypertension, and diabetes [18,19]. It has been shown that other comorbidities present in patients with COPD are observed at a younger age [20]. The cross-sectional studies conducted by Anecchino et al [21] and Holguin et al [22] showed that 68% of patients admitted for COPD had at least one comorbidity, 16% had 2 or more comorbidities, and 30% had 4 or more comorbidities. It is also the third leading cause of death in the world [23]. This implies a significant need for the use of health services [24,25]. Therefore, the need and importance of using a FAIR strategy would facilitate data sharing and, thus, scientific discovery, in line with the objectives addressed in FAIR4Health.

Previous studies have shown that there are several risk factors associated with readmission in patients with COPD, such as significant deterioration of lung function, low oxygen saturation in pulse oximetry, decreased activity levels, comorbidities, and the absence of medication reconciliation during hospitalization [26]. Hospital readmissions usually have a negative impact on the quality of life of patients and their families and present a considerable economic burden for health care systems. Furthermore, previous findings support the recognition of high readmission risks associated with patients who have been hospitalized frequently in the past, along with other assessments that may be useful in better predicting readmission risk over the course of a patient's stay [27].

Regarding the comorbidities, it is noted that several studies agree that the greater the number of comorbidities, the greater the risk of readmission for patients with COPD [28,29]. The rate of readmission within 30 days of discharge is used on many occasions to judge the quality of hospital care received. Using data from Medicare beneficiaries, it is estimated that approximately 1 in 5 patients discharged from the hospital because of COPD are readmitted within 30 days [30,31]. A recently published study by Gershon et al [24] analyzed 252,756 individuals hospitalized for COPD and showed that the risk factors for readmission during this period were the number of previous admissions, the modified Medical Research Council dyspnea score, age, and chronic heart failure (both right and left).

Therefore, COPD is a major health problem that must be addressed and analyzed [32]. Several studies have evaluated the risk of readmission rate for these patients [30,33,34], but few studies have considered this risk for a period of 30 days. In addition, few studies have succeeded in considering the comorbidities and functional and care data of these patients. For all these reasons, the FAIR4Health pathfinder case study included the development of an early predictive model for 30-day readmission risk in patients with COPD. This study was carried out to understand the impact of these data on the rate of readmission within 30 days of discharge. Addressing these aspects, which are of high risk during the planning of hospital discharge, could help prevent readmission and develop a model that helps predict which patients demonstrate greater frailty and, therefore, a higher risk of hospital readmission.

Goal of This Study

In this paper, the clinical validation of the FAIR4Health solution is described, including the development and selection of the most appropriate model for predicting 30-day readmission risk in patients with COPD and the assessment of such a model. This study builds upon the FAIRification of health research data sets of different health research performing organizations and a federated machine learning architecture on top of the FAIRified data sets of different organizations. The entire FAIR4Health solution was validated in real-world settings with the clinical use case described in this paper.

Methods

Study Design and Recruitment

The use case that was designed in this study to validate the FAIR4Health solution was composed of two phases: (1) a retrospective multicenter observational study, including the training of the predictive models in the FAIR4Health platform, and (2) an observational prospective study with a 30-day follow-up.

Retrospective Study

In the retrospective study, the population included patients aged >18 years diagnosed with COPD, considering that COPD-related comorbidities are observed at a younger age [20]. Patients with programmed admission in any hospital department within 30 days of discharge, patients with psychiatric disease, and patients with neurodegenerative diseases were excluded from the study. Following the clinical protocol defined in this study, this first phase covered retrospective data collection from the relevant data sources specified below.

In the first phase, which is to train the federated machine learning models, three different organizations participated with their health care (hospital, primary care, and nursing homes) and health research data sets: (1) Universite De Geneve from Switzerland provided health care data from the electronic health record (EHR) of the University Hospitals of Geneva; (2) Virgen del Rocío University Hospital as part of the Andalusian Health Service (Servicio Andaluz de Salud [SAS]) from Spain provided health care data from the EHR of the Virgen del Rocío University Hospital in Seville; and (3) Instituto Aragonés de Ciencias de la Salud and Instituto de Investigación Sanitaria Aragón from Spain provided a health research data set based on the EpiChron Cohort [20,35], a study carried out by the Instituto Aragonés de Ciencias de la Salud.

For organizations contributing with health research data sets from previous research projects, the sample size was defined by taking into account the original size of the data sets in the previous research, whereas for organizations contributing with health care data sets from the EHRs, it was defined from the number of patients fulfilling the inclusion and exclusion criteria.

The variables for the training and prediction processes were related to demographic, multimorbidity, comorbidities, polypharmacy, laboratory, and hospitalization data. The

principal dependent variable was readmission, defined as unplanned hospitalization for any cause related to COPD within 30 days of hospital discharge.

Prospective Study

Following the clinical protocol defined in the study, an observational prospective study with a 30 days follow-up was carried out after the retrospective study to assess the impact of the early predictive model by collecting data from a cohort of recruited patients. Patients aged ≥18 years with a diagnosis of COPD who were admitted to the hospital for this disease (unplanned hospitalization) and who signed the informed consent form (ICF) were included in the observational prospective study, complying with the same inclusion and exclusion criteria as described for the retrospective study.

Two health care organizations participated in the observational prospective study in which the trained predictive model was tested: (1) Internal Medicine Department of the Virgen del Rocío University Hospital in Seville as part of the Andalusian Health Service (SAS) from Spain and (2) Clinic for Obstructive Pulmonary Diseases and Acute Pneumopathies of the Institute for Pulmonary Diseases of Vojvodina (IPBV) from Serbia. In both cases, the sample size was defined by considering the number of patients admitted to the hospital during the prospective study period, thus fulfilling the inclusion and exclusion criteria.

Regarding the study variables, the same variables were collected at the time of inclusion of each patient during the prospective study as in the retrospective study. As a monitoring variable, aiming to assess the prediction performance of the model on the patient's risk of readmission, it was analyzed whether the patient with COPD had a readmission within the 30 days of discharge.

Ethics Approval

Ethical approval for this study was obtained from all participating health research organizations based on regional regulations before involving them in the execution of the case studies (Universite De Geneve and University Hospitals of Geneva from Switzerland: 2020-02683; Virgen del Rocío University Hospital as part of the Andalusian Health Service from Spain: 1269-M1-20; and Instituto Aragonés de Ciencias de la Salud and Instituto de Investigación Sanitaria Aragón from Spain , 1269-M1-20).

Technical and organizational measures were defined to safeguard the rights and freedoms of the data participants, including the data minimization principle. Informed consent procedures were defined, including informed consent and information sheets. A data protection officer was appointed at each data owner institution. To reinforce the appropriate coverage of these ethical aspects, at the beginning of the study, an external ethics advisory board was made up, which involved reviewing deliverables, generating reports, and performing presentations to support the FAIR4Health Consortium.

FAIRification Workflow and Tools

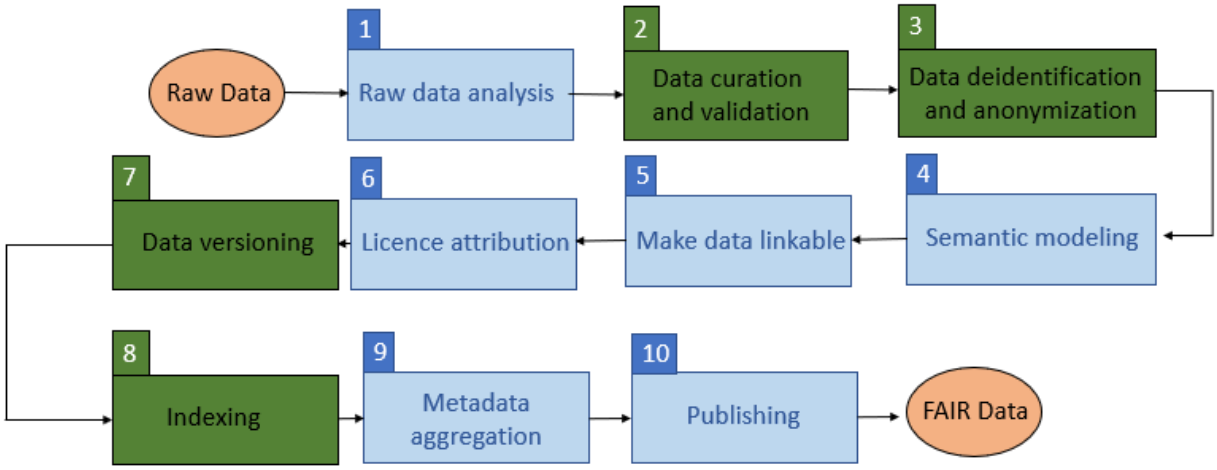
Making health data FAIR opens up new horizons, especially for the secondary use of health care and reuse of health research data sets. The FAIR4Health project proposed a FAIRification workflow [12] to be used for making existing health data sets FAIR. This workflow includes a series of actionable steps and a technological design and implementation guide for each step.

To address the challenges of the health domain, the proposed workflow adapted the generic FAIRification process defined by GO FAIR [36]. First, this workflow contextualizes the generic steps. Second, the FAIR4Health workflow introduced new steps with a strong consideration of the legal, technical, and ethical implications that reusing health data sets may have.

These steps were (1) raw data analysis, (2) data curation and validation, (3) data deidentification and anonymization, (4) semantic modeling, (5) making data linkable, (6) license attribution, (7) data versioning, (8) indexing, (9) metadata aggregation, and (10) publishing.

Steps 2, 3, 7, and 8 were newly introduced in the FAIR4Health FAIRification workflow. Figure 1 shows a visual representation of this workflow.

Figure 1. The FAIR4Health FAIRification workflow (redefined from the study by Sinaci et al [12]). FAIR: Findable, Accessible, Interoperable, and Reusable.



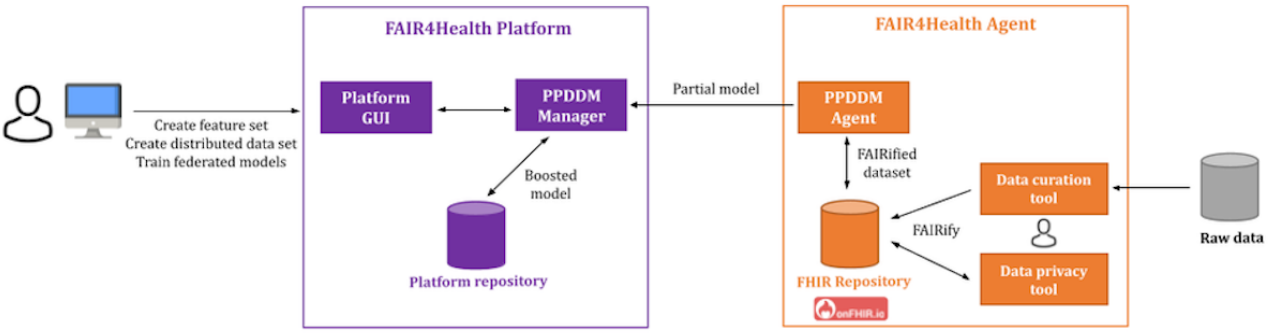
The FAIRification workflow was based on the HL7 Fast Healthcare Interoperability Resources (FHIR) [37]. Making data FAIR by using a well-established standard such as HL7 FHIR not only contributed to FAIRification but also helped the data owner organizations conform to a widely adopted standard. The FAIR4Health project developed a set of software tools around the HL7 FHIR as an implementation of the FAIRification workflow, the so-called FAIRification tools. In addition to the methodology and FHIR usage, these tools, namely, onFHIR.io repository [38], data curation tool (DCT) [39], and data privacy tool (DPT) [40], were deployed and used at each of the 3 data source organizations for the retrospective study. A set of FHIR profiles to serve as the common data model [41] of the FAIR4Health project was developed to cover the data requirements of the use cases. The onFHIR.io installations of the FAIR4Health project were shipped with the FAIR4Health profiles; hence, the FAIR4Health design led to uniform, interoperable, and reusable data sets once FAIRification was completed at each retrospective study organization.

Along with the onFHIR.io repositories, at each organization, a DCT and a DPT were installed, and these tools were used by the data managers and FAIR4Health researchers to FAIRify their existing data sets, collaborating to appropriately treat the databases. Following the FAIRification workflow, the raw data were first transformed into HL7 FHIR by creating the associated FHIR resources through the DCT. It was shown that the DCT is a valid software tool that meets the challenges of raw data analysis, curation, and validation steps [42]. Once the data were migrated into the onFHIR.io repository, the DPT was used to deidentify the resources with respect to the policy requirements of the organizations. The use of FHIR resources and the onFHIR.io repository helped us to successfully cover the other workflow steps such as versioning, indexing, and license attribution. At the end of the FAIRification process for each organization, the FAIR data were ready to be consumed by the federated machine learning algorithms so that predictive models could be built on top of the retrospective data.

Federated Machine Learning Models

The FAIR4Health project implemented the PPDDM philosophy by designing and implementing a federated machine learning architecture. The ultimate aim of this architecture is to address the challenging security and privacy concerns of health data owners. The PPDDM architecture does not allow data to leave their servers. Partial machine learning models were trained on each FAIRified data set at each organization, and then these partial models were used to develop a boosted machine learning model on the central FAIR4Health platform. The platform provides a web-based graphical user interface to the researchers so that they can define their features, create distributed data sets, and then train federated models. The PPDDM architecture was composed of the agent implementation. Then, the agents were deployed at each data source organization on top of their FAIRified data sets. These agents communicated with their associated onFHIR.io repositories at each deployment site. A manager was deployed as a backend to the FAIR4Health platform graphical user interface so that these agents can be orchestrated to build distributed data sets and federated predictive models on top of those distributed data sets. Figure 2 shows a graphical representation of the FAIR4Health federated architecture.

Figure 2. The FAIR4Health federated architecture. GUI: graphical user interface; PPDDM: privacy-preserving distributed data mining.



During the retrospective study, the researchers of the data owner organizations used the platform to train federated machine learning models on the retrospective data sets that were previously made FAIR using the FAIRification tools. The PPDDM implementation provided a set of machine learning algorithms to the researchers to be executed in a federated manner. These algorithms were grouped as (1) support vector machine, (2) logistic regression, (3) decision trees, (4) random forest, and (5) gradient-boosted trees.

Results

Model Generation and Adjustment

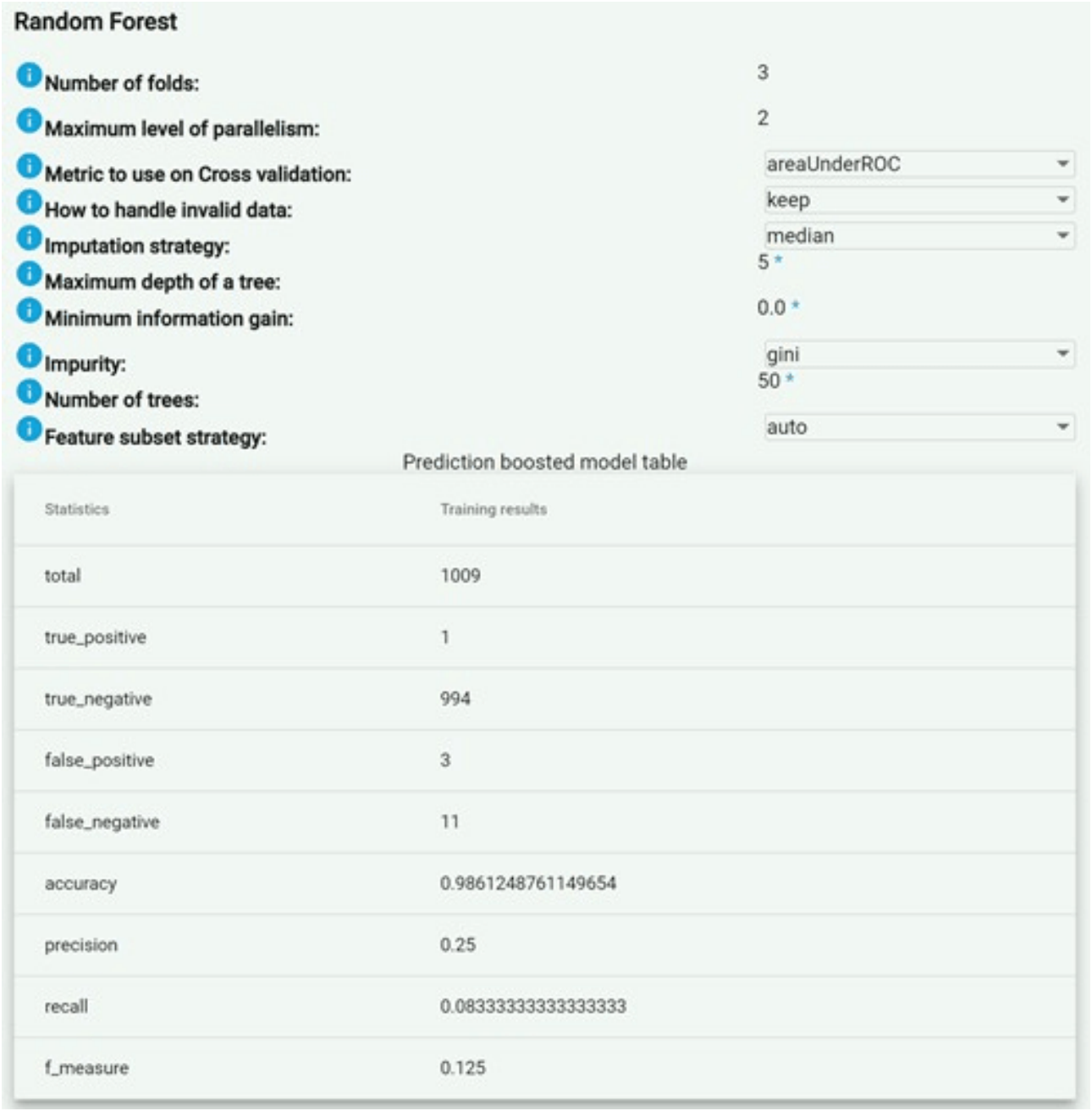
During the retrospective study, a number of machine learning models were generated by using the prediction algorithms listed above as well as trying out various values for different parameters (eg, imputation strategy, classification threshold, maximum depth of a tree, and feature subset strategy). More focus was given to the tree-based algorithms because the data in the agents were skewed in one direction, and tree-based methods produced better results than the others when the data were unbalanced. In addition, k-fold cross-validation was used to split the data into a set of nonoverlapping training and test sets to obtain more accurate results.

In the experiments, better results were obtained with the predictive models generated using the random forest algorithm. An example screenshot of the platform is shown in Figure 3. While creating the model, different values for the maximum depth of a tree (range 5-15), minimum information gain

(between 0.0 and 0.5), impurity (gini or entropy), and number of trees (range 25-100) were provided. The FAIR4Health platform tried all these values with the grid search functionality to determine the best combination. Therefore, considering the knowledge of the FAIR4Health researchers with an expert background in this kind of algorithm, the best model with an accuracy of 98.6% was generated and selected with the following values:

- 3-fold cross-validation with area under the curve of the receiver operating characteristic evaluation metric
- Imputation strategy: median—replaces the missing values using the approximate median value of the feature
- Maximum depth of a tree: 5
- Minimum information gain: 0.0
- Impurity: gini
- Number of trees: 50
- Feature subset strategy: auto-calculates the number of features at each tree node as the square root of the total number of features in the classification algorithm.

Figure 3. Result of a random forest model.



Clinical Validation

After the parameters of the algorithm were selected, the predictive model was generated using retrospective data sets of 4.944 patients with COPD. Subsequently, an observational prospective study was conducted to validate and evaluate an early predictive model for 30-day readmission risk in patients with COPD.

In total, 100 patients were recruited and included in the observational prospective study with a 30-day follow-up, from April 2021 to September 2021, including recruitment and

follow-up. During that period, the study participants were recruited by performing weekly prevalence cuts in which all patients hospitalized because of COPD conditions were systematically evaluated, offering inclusion in this study to all those who met the inclusion criteria and did not meet any exclusion criteria.

Clinicians and researchers performed functional and clinical validations of the FAIR4Health solution during the observational prospective study. As this was a multicenter observational study, the recruitment and inclusion of patients in the study were carried out as mentioned below.

For SAS, the clinical team reviewed 711 hospitalized patients during the study period, and 53 (7.5%) of them fulfilled the inclusion criteria and did not meet any exclusion criteria. Finally, 22 patients with COPD signed the ICF and were included in this observational prospective study. Out of the total recruited patients in SAS, 18% (4/22) were female and 82% (18/22) were male.

In the case of IPBV, out of 2070 hospitalized patients, 113 (5.46%) patients were hospitalized because of COPD exacerbation, and 83 (73.5%) patients met all inclusion criteria and did not meet any exclusion criteria and signed the ICF. A total of 78 patients were included in this observational prospective study.

Of the total patients recruited during the study period, 47% (37/78) were female and 53% (41/78) were male.

All data gathered from patients with COPD were entered into the FAIR4Health platform to obtain the prediction generated by the predictive model for 30-day readmission risk and to assess its performance.

Evaluation Outcomes

When the prediction was obtained using the FAIR4Health platform, a concordance analysis was performed to compare the real data with the predicted values. Concerning the reality of readmissions among the 100 patients recruited, in both cases, the patients were followed up during hospitalization, and the follow-up was performed during the following 30 days. Out of a total of 22 patients recruited from SAS, 3 (14%) were readmitted within 30 days of discharge (ie, during the follow-up period). Out of a total of 78 patients recruited from IPBV, 10 (15%) were readmitted during the follow-up period. Finally, from the 100 recruited patients, (1) the accuracy of predictions generated by the FAIR4Health platform was confirmed in 87% (87/100) of the cases; that is, either the patient was readmitted to the hospital because of COPD in real life and the algorithm predicted that there was early 30-day hospital readmission risk or the patient was not readmitted and the algorithm predicted that there was no early 30-day hospital readmission risk and (2) the prediction generated was not confirmed in 13% (13/100) of the cases; that is, in real life, the patient was readmitted within 30 days and the platform predicted that there was no early 30-day hospital readmission risk or the patient was not readmitted and the platform predicted that there was early 30-day hospital readmission.

Discussion

Principal Findings

The application of the FAIR principles in health research data sets of health research performing organizations from different countries allowed the federated data analysis to accelerate the discovery of scientific outputs. Therefore, the analysis of legal, technical, and ethical requirements of health research data were addressed during data FAIRification. Furthermore, a clinical decision support model for predicting 30-day readmission risk in patients with COPD at discharge based on the risk factors uncovered previously, using data mining approaches, was implemented, deployed, and validated. Finally, through a multicenter study in which the rate of readmission of patients with COPD within 30 days after hospital discharge was analyzed, clinical partners could reach use case objectives and obtain an early 30-day hospital readmission risk predictive model. Further details of the FAIR4Health pathfinder case studies can be found in the FAIR4Health public report on the demonstrators' performance [43].

It is important to highlight that the FAIR4Health solution was implemented following a practical extensibility capacity, so that other research questions can be covered using the solution without the need to perform adaptations. Furthermore, to improve the reusability capacity of the study, using both the open-source code and the generated metadata freely available in GitHub [44], the study can be reproduced.

Limitations

First, significant cross-cutting data-related challenges were addressed during data collection. Data extraction from EHRs and other types of health care sources aligning this extraction with a FAIR4Health common data model was not trivial and required a lot of conceptual and technical efforts because of (1) the complexity of the raw data (the sources of EHRs are commonly very complex including the information in several tables in the source databases), (2) free text used in some fields in the raw data sources, and (3) differences between the types of raw data sources. To address the complexity of the raw data, each health research organization from different countries that participated in the data extraction involved colleagues who were experts in each source data model. To handle the information in free text fields, natural language processing techniques were assessed. Finally, in some cases, manual natural language processing to extract structured information from unstructured information was performed. To manage the differences between the nature of the raw data sources, each raw data set was analyzed in depth in a collaborative effort between each clinical partner and the technical partners to reach the required configuration in the FAIR4Health solution, achieving the FAIRification of all raw data and finally achieving the PPDDM models' generation using all sources.

Second, concerning the predictive model generated in this study, it can be stated that it is possible to generate more efficient prediction parameters (with better accuracy, precision, and recall values) if the distribution of the readmission variable in the data sets is better adjusted. The readmission variable, which was the dependent variable, was not balanced in the data sets of the

retrospective studies (data sets used to generate the predictive model for this prospective study), which resulted in the generated results being good but not perfect as desired. For more effective models, in the future, a better adjustment of the distribution of the readmission variable using data sets with more patients will be addressed to boost the application of predictive models in clinical practice. Most studies of predictive models based on machine learning show poor methodological quality and are at a high risk of bias. The small study size, poor management of missing data, and failure to address overfitting are factors that contribute to the risk of bias [45].

In contrast, it is crucial to add that this study was carried out while these 2 health care organizations were experiencing the consequences of the COVID-19 pandemic, and clinical researchers had to make significant efforts to properly conclude the prospective study:

- IPBV as a health care institution was included in the national COVID-19 system of health care institutions caring for COVID-19 positive patients with severe clinical difficulties. Owing to this reorganization of the Serbian health care system, the likelihood of hospitalization of patients with COPD has been reduced since March 2020. Many of the researchers responsible for patient recruitment in the prospective study were engaged in COVID-19 departments, and the remaining researchers were overworked during the study period.
- On the side of SAS, this health care institution was involved in the care of patients with suspicion of COVID-19 and COVID-19–positive patients with severe clinical difficulties. All health professionals in SAS had a higher workload in health care. In fact, different clinical researchers participating in this observational study were transferred during the project to the COVID-19 Emergency Hospital in Seville (Spain), relieving each other, with an essential health care priority and looking after patients who did not meet the inclusion criteria of this study and could not be recruited. The clinical researchers identified a low use of health care services (both urgencies and consultancies) by patients with COPD; presumably, the patients waited for more severe symptoms to go to the health care centers because of the fear of having contact with COVID-19–positive patients. In addition, hospitalizations of patients with COPD were restricted, similar to what has happened in other pathologies, to avoid patient flow through health care centers.

Next Steps

Considering the final version of the FAIR4Health solution and the main outcomes of this study, some future advances can be taken into account:

- Both the FAIRification tools and the FAIR4Health platform were validated using the FAIR4Health common data model. The solution has been designed and developed by considering the extensive capacity of other data models, so it is appropriate to continue the validation and testing with other data models in future clinical validations.
- The whole FAIR4Health solution covers alignment with relevant standards: HL7 FHIR, International Classification of Diseases, SNOMED Clinical Terms, Logical Observation Identifiers Names and Codes, and the Anatomical Therapeutic Chemical classification system. Other standards such as other HL7 standards, epidemiological standards, and W3C standards could be considered to be integrated if viable.
- The FAIR4Health platform was validated using the following machine learning algorithms: frequent pattern growth, support vector machine, logistic regression, decision trees, random forest, and gradient-boosted trees. Deep learning algorithms such as neural networks can be considered in future studies to improve the capabilities of the FAIR4Health platform.

From a scientific point of view, some researchers of the FAIR4Health Consortium contribute to the application of the FAIR principles in the health research field, being involved in international working groups part of the European Open Science Cloud, the European Federation for Medical Informatics, the Research Data Alliance, the GO FAIR initiative, and HL7 International.

Conclusions

Despite the limitations mentioned above, the objective of this study was achieved: to validate the FAIR4Health solution through the assessment of a federated model that was generated by applying a federated machine learning architecture on top of the FAIRified data sets of different health research performing organizations for real-time prediction of 30-day readmission risk in patients with COPD.

The clinical, technical, and functional validation of the FAIR4Health solution was achieved through (1) the application of FAIR principles through the FAIR4Health FAIRification tools in health research data sets of different health research performing organizations and FAIRifying data from 4.944 patients with COPD; (2) development and use of federated machine learning architecture on top of the FAIRified data sets; and (3) clinical, technical, and functional development and assessment of a federated model for predicting 30-day readmission risk in patients with COPD, with an accuracy of 0.98, a precision of 0.25, and a confirmed prediction in 87% (87/100) of the cases.

In the retrospective study where 3 different organizations participated with their health care (hospital, primary care, and nursing homes) and health research data sets, the federated model was generated with an accuracy of 98.6% and a precision of 25%. In the observational prospective study in which 2 health care organizations participated, 100 patients were recruited for the federated model to predict their readmission risk to the hospital within 30 days because of COPD. Therefore, the accuracy of predictions generated by the model, and hence the FAIR4Health platform, was confirmed in 87% (87/100) of the cases.

Health research performing organizations are aware of the need to implement a FAIR data policy to facilitate data sharing and reuse following the discovery, access, integration, and analysis of health research data. One obvious example would be the COVID-19 pandemic, where international cooperation allowed

the rapid sequencing and epidemiological studies to be carried out, thus demonstrating the need and importance of data sharing to accelerate health research [46,47]. For this purpose, organizations are usually making efforts to align themselves with the FAIR principles. This is the real and practical consequence of the FAIR4Health project in terms of patient management and health planning: to improve health research in specific pathologies through the findability-, accessibility-, interoperability-, and reusability-enhanced features in the case of health data.

The FAIR4Health project proposes a technological solution in the health domain to facilitate the use of larger and more heterogeneous data sets, thus increasing the variability of the data and the size of the data sets. Therefore, an increase in the scope of the research will be obtained and a significant improvement in the ability to generate more accurate predictive models.

Acknowledgments

This work was supported by the FAIR4Health project [10], which received funding from the European Union's Horizon 2020 research and innovation program under grant 824666. This research has also been cosupported by the Carlos III National Institute of Health through the Programa de Ciencia de Datos de la Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología program (IMPaCT-Data, code IMP/00019) and through the Platform for Dynamization and Innovation of the Spanish National Health System industrial capacities and their effective transfer to the productive sector (code PT20/00088), both cofunded by the European Regional Development Fund Fondo Europeo de Desarrollo Regional "A way of making Europe." The authors would like to thank the clinical researchers of the project, coming from the organizations that are part of the FAIR4Health Consortium: Universite De Geneve (Switzerland), University Hospitals of Geneva (Switzerland), Università Cattolica Del Sacro Cuore (Italy), Universidade Do Porto (Portugal), Instituto Aragonés de Ciencias de la Salud (Spain), Institut Za Plucne Bolesti Vojvodine (Serbia), and Servicio Andaluz de Salud (Spain).

Conflicts of Interest

None declared.

References

1. Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018-160019 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
2. Parra-Calderón CL, Sanz F, McIntosh LD. The challenge of the effective implementation of FAIR principles in biomedical research. Methods Inf Med 2020 Aug 22;59(4-05):117-118. [doi: 10.1055/s-0040-1721726] [Medline: 33618419]
3. Delgado J, Llorente S. Security and privacy when applying FAIR principles to genomic information. Stud Health Technol Inform 2020 Nov 23;275:37-41. [doi: 10.3233/SHTI200690] [Medline: 33227736]
4. Dijkers MP. A beginner's guide to data stewardship and data sharing. Spinal Cord 2019 Mar 5;57(3):169-182. [doi: 10.1038/s41393-018-0232-6] [Medline: 30723254]
5. Couture JL, Blake RE, McDonald G, Ward CL. A funder-imposed data publication requirement seldom inspired data sharing. PLoS One 2018 Jul 6;13(7):e0199789 [FREE Full text] [doi: 10.1371/journal.pone.0199789] [Medline: 29979709]
6. Almada M, Midão L, Portela D, Dias I, Núñez-Benjumea FJ, Parra-Calderón CL, et al. [A new paradigm in health research: FAIR data (Findable, Accessible, Interoperable, Reusable)]. Acta Med Port 2020 Dec 02;33(12):828-834 [FREE Full text] [doi: 10.20344/amp.12910] [Medline: 33496252]
7. Holub P, Kohlmayer F, Prasser F, Mayrhofer MT, Schlünder I, Martin GM, et al. Enhancing reuse of data and biological material in medical research: from FAIR to FAIR-health. Biopreserv Biobank 2018 Apr;16(2):97-105 [FREE Full text] [doi: 10.1089/bio.2017.0110] [Medline: 29359962]
8. Mello MM, Lieou V, Goodman SN. Clinical trial participants' views of the risks and benefits of data sharing. N Engl J Med 2018 Jun 07;378(23):2202-2211 [FREE Full text] [doi: 10.1056/NEJMsa1713258] [Medline: 29874542]
9. Rios R, Zheng KI, Zheng MH. Data sharing during COVID-19 pandemic: what to take away. Expert Rev Gastroenterol Hepatol 2020 Dec;14(12):1125-1130. [doi: 10.1080/17474124.2020.1815533] [Medline: 32842793]
10. Inau E, Sack J, Waltemath D, Zeleke AA. Initiatives, concepts, and implementation practices of FAIR (findable, accessible, interoperable, and reusable) data principles in health data stewardship practice: protocol for a scoping review. JMIR Res Protoc 2021 Feb 02;10(2):e22505 [FREE Full text] [doi: 10.2196/22505] [Medline: 33528373]
11. FAIR4Health key outputs for the scientific community. FAIR4Health. URL: https://www.fair4health.eu/ [accessed 2022-05-11]
12. Sinaci A, Núñez-Benjumea FJ, Gencturk M, Jauer ML, Deserno T, Chronaki C, et al. From raw data to FAIR data: the FAIRification workflow for health research. Methods Inf Med 2020 Jun;59(S 01):e21-e32 [FREE Full text] [doi: 10.1055/s-0040-1713684] [Medline: 32620019]
13. The FAIR data principles. FORCE11. URL: https://www.force11.org/group/fairgroup/fairprinciples [accessed 2022-05-11]

14. EOSC Declaration. URL: https://eosc-portal.eu/sites/default/files/eosc_declaration.pdf [accessed 2022-05-11]
15. European Open Science Cloud (EOSC) Strategic Implementation Plan. European Commission. URL: https://ec.europa.eu/info/publications/european-open-science-cloud-eosc-strategic-implementation-plan_en [accessed 2022-05-11]
16. Adeloye D, Chua S, Lee C, Basquill C, Papana A, Theodoratou E, Global Health Epidemiology Reference Group (GHERG). Global and regional estimates of COPD prevalence: systematic review and meta-analysis. J Glob Health 2015 Dec;5(2):020415 [FREE Full text] [doi: 10.7189/jogh.05-020415] [Medline: 26755942]
17. Mannino D, Gagnon R, Petty T, Lydick E. Obstructive lung disease and low lung function in adults in the United States: data from the National Health and Nutrition Examination Survey, 1988-1994. Arch Intern Med 2000 Jun 12;160(11):1683-1689. [doi: 10.1001/archinte.160.11.1683] [Medline: 10847262]
18. Baty F, Putora P, 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]
19. Divo M, Cote C, de Torres JP, Casanova C, Marin JM, Pinto-Plata V, et al. Comorbidities and risk of mortality in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2012 Jul 15;186(2):155-161. [doi: 10.1164/rccm.201201-0034oc]
20. Divo MJ, Celli BR, Poblador-Plou B, Calderón-Larrañaga A, de-Torres JP, Gimeno-Feliu LA, EpiChron—BODE Collaborative Group. Chronic Obstructive Pulmonary Disease (COPD) as a disease of early aging: evidence from the EpiChron Cohort. PLoS One 2018 Feb 22;13(2):e0193143 [FREE Full text] [doi: 10.1371/journal.pone.0193143] [Medline: 29470502]
21. Anecchino C, Rossi E, Fanizza C, De Rosa M, Tognoni G, Romero M, working group ARNO project. Prevalence of chronic obstructive pulmonary disease and pattern of comorbidities in a general population. Int J Chron Obstruct Pulmon Dis 2007;2(4):567-574 [FREE Full text] [Medline: 18268930]
22. Holguin F, Folch E, Redd SC, Mannino DM. Comorbidity and mortality in COPD-related hospitalizations in the United States, 1979 to 2001. Chest 2005 Oct;128(4):2005-2011. [doi: 10.1378/chest.128.4.2005] [Medline: 16236848]
23. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012 Dec 15;380(9859):2095-2128. [doi: 10.1016/S0140-6736(12)61728-0] [Medline: 23245604]
24. Gershon A, Thiruchelvam D, Aaron S, Stanbrook M, Vozoris N, Tan W, et al. Socioeconomic status (SES) and 30-day hospital readmissions for chronic obstructive pulmonary (COPD) disease: a population-based cohort study. PLoS One 2019;14(5):e0216741 [FREE Full text] [doi: 10.1371/journal.pone.0216741] [Medline: 31112573]
25. About gold. Global Initiative for Chronic Obstructive Lung Disease. URL: http://goldcopd.org/ [accessed 2022-05-11]
26. Coventry P, Gemmell I, Todd C. Psychosocial risk factors for hospital readmission in COPD patients on early discharge services: a cohort study. BMC Pulm Med 2011 Nov 04;11:49 [FREE Full text] [doi: 10.1186/1471-2466-11-49] [Medline: 22054636]
27. Jiang W, Siddiqui S, Barnes S, Barouch LA, Korley F, Martinez DA, et al. Readmission risk trajectories for patients with heart failure using a dynamic prediction approach: retrospective study. JMIR Med Inform 2019 Sep 16;7(4):e14756 [FREE Full text] [doi: 10.2196/14756] [Medline: 31579025]
28. Brand C, Sundararajan V, Jones C, Hutchinson A, Campbell D. Readmission patterns in patients with chronic obstructive pulmonary disease, chronic heart failure and diabetes mellitus: an administrative dataset analysis. Intern Med J 2005 May;35(5):296-299. [doi: 10.1111/j.1445-5994.2005.00816.x] [Medline: 15845113]
29. Kelly M. Self-management of chronic disease and hospital readmission: a care transition strategy. J Nursing Healthcare Chronic Illness 2011;3(1):4-11. [doi: 10.1111/j.1752-9824.2010.01075.x]
30. The Lancet Respiratory Medicine. Reducing COPD readmissions—a personal and political priority. Lancet Respiratory Med 2013 Jul;1(5):347. [doi: 10.1016/s2213-2600(13)70153-x]
31. Jencks SF, Williams MV, Coleman EA. Rehospitalizations among patients in the medicare fee-for-service program. N Engl J Med 2009 Apr 02;360(14):1418-1428. [doi: 10.1056/nejmsa0803563]
32. Vos T, Allen C, Arora M, Barber R, Bhutta Z, Brown A, et al. . [doi: 10.1016/S0140-6736(16)31678-6]
33. Chan F, Wong F, Yam C, Cheung W, Wong E, Leung M, et al. Risk factors of hospitalization and readmission of patients with COPD in Hong Kong population: analysis of hospital admission records. BMC Health Serv Res 2011 Aug 10;11:186 [FREE Full text] [doi: 10.1186/1472-6963-11-186] [Medline: 21831287]
34. Jacobs DM, Noyes K, Zhao J, Gibson W, Murphy TF, Sethi S, et al. Early hospital readmissions after an acute exacerbation of chronic obstructive pulmonary disease in the nationwide readmissions database. Annals ATS 2018 Jul;15(7):837-845. [doi: 10.1513/annalsats.201712-913oc]
35. Prados-Torres A, Poblador-Plou B, Gimeno-Miguel A, Calderón-Larrañaga A, Poncel-Falcó A, Gimeno-Feliú LA, et al. Cohort profile: the epidemiology of chronic diseases and multimorbidity. The Epichron cohort study. Int J Epidemiol 2018 Apr 01;47(2):382-34f [FREE Full text] [doi: 10.1093/ije/dyx259] [Medline: 29346556]
36. GO FAIR initiative. GO FAIR. URL: https://www.go-fair.org [accessed 2022-05-11]
37. Welcome to FHIR. HL7 FHIR. URL: http://hl7.org/fhir/ [accessed 2022-05-11]
38. HL7 FHIR® Based Secure Data Repository. onFHIR.io. URL: https://onfhir.io [accessed 2022-05-11]

39. FAIR4Health data curation and validation tool. GitHub. URL: https://github.com/fair4health/data-curation-tool [accessed 2022-05-11]
40. FAIR4Health data privacy tool. GitHub. URL: https://github.com/fair4health/data-privacy-tool [accessed 2022-05-11]
41. fair4health / common-data-model. GitHub. URL: https://github.com/fair4health/common-data-model [accessed 2022-05-11]
42. Gencturk M, Teoman A, Alvarez-Romero C, Martinez-Garcia A, Parra-Calderon CL, Poblador-Plou B, et al. End user evaluation of the FAIR4Health data curation tool. Stud Health Technol Inform 2021 May 27;281:8-12. [doi: 10.3233/SHTI210110] [Medline: 34042695]
43. D5.5. Report on the demonstrators performance_v2_vf.pdf. OSF Home. URL: https://osf.io/tfnqa/ [accessed 2022-05-11]
44. FAIR4Health. GitHub. URL: https://github.com/fair4health/ [accessed 2022-05-11]
45. Andaur Navarro CL, Damen JA, Takada T, Nijman SW, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ 2021 Oct 20;375:n2281 [FREE Full text] [doi: 10.1136/bmj.n2281] [Medline: 34670780]
46. Kinsella C, Santos PD, Postigo-Hidalgo I, Folgueiras-González A, Passchier TC, Szillat KP, et al. Preparedness needs research: how fundamental science and international collaboration accelerated the response to COVID-19. PLoS Pathog 2020 Oct;16(10):e1008902 [FREE Full text] [doi: 10.1371/journal.ppat.1008902] [Medline: 33035262]
47. Besançon L, Peiffer-Smadja N, Segalas C, Jiang H, Masuzzo P, Smout C, et al. Open science saves lives: lessons from the COVID-19 pandemic. BMC Med Res Methodol 2021 Jun 05;21(1):117-118 [FREE Full text] [doi: 10.1186/s12874-021-01304-y] [Medline: 34090351]

Abbreviations

COPD: chronic obstructive pulmonary disease
DCT: data curation tool
DPT: data privacy tool
EHR: electronic health record
EU: European Union
FAIR: Findable, Accessible, Interoperable, and Reusable
FHIR: Fast Healthcare Interoperability Resources
ICF: informed consent form
IPBV: Institute for Pulmonary Diseases of Vojvodina
PPDDM: privacy-preserving distributed data mining
SAS: Servicio Andaluz de Salud

Edited by C Lovis; submitted 30.11.21; peer-reviewed by H Abaza, JJ Mira; comments to author 26.12.21; revised version received 16.03.22; accepted 21.04.22; published 02.06.22.

Please cite as:
Alvarez-Romero C, Martinez-Garcia A, Ternero Vega J, Díaz-Jimènez P, Jimènez-Juan C, Nieto-Martín MD, Román Villarán E, Kovacevic T, Bokan D, Hromis S, Djekic Malbasa J, Beslać S, Zaric B, Gencturk M, Sinaci AA, Ollero Baturone M, Parra Calderón CL
Predicting 30-Day Readmission Risk for Patients With Chronic Obstructive Pulmonary Disease Through a Federated Machine Learning Architecture on Findable, Accessible, Interoperable, and Reusable (FAIR) Data: Development and Validation Study
JMIR Med Inform 2022;10(6):e35307
URL: https://medinform.jmir.org/2022/6/e35307
doi:10.2196/35307
PMID:35653170

Original Paper

Virtual Specialist Care During the COVID-19 Pandemic: Multimethod Patient Experience Study

Katie N Dainty[1], BA, MSc, PhD; M Bianca Seaton[1], MA; Antonio Estacio[2]; Lisa K Hicks[2], MD; Trevor Jamieson[2], MD; Sarah Ward[2], MD; Catherine H Yu[2], MD; Jeffrey D Mosko[2*], MD; Charles D Kassardjian[2*], MD

[1]North York General Hospital, Toronto, ON, Canada

[2]St Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada

[*]these authors contributed equally

Corresponding Author:
Katie N Dainty, BA, MSc, PhD
North York General Hospital
4001 Leslie Street
LE-140
Toronto, ON, M2K 1E1
Canada
Phone: 1 16474482485
Email: katie.dainty@utoronto.ca

Abstract

Background: Transitioning nonemergency, ambulatory medical care to virtual visits in light of the COVID-19 global pandemic has been a massive shift in philosophy and practice that naturally came with a steep learning curve for patients, physicians, and clinic administrators.

Objective: We undertook a multimethod study to understand the key factors associated with successful and less successful experiences of virtual specialist care, particularly as they relate to the patient experience of care.

Methods: This study was designed as a multimethod patient experience study using survey methods, descriptive qualitative interview methodology, and administrative virtual care data collected by the hospital decision support team. Six specialty departments participated in the study (endoscopy, orthopedics, neurology, hematology, rheumatology, and gastroenterology). All patients who could speak and read English and attended a virtual specialist appointment in a participating clinic at St. Michael's Hospital (Toronto, Ontario, Canada) between October 1, 2020, and January 30, 2021, were eligible to participate.

Results: During the study period, 51,702 virtual specialist visits were conducted in the departments that participated in the study. Of those, 96% were conducted by telephone and 4% by video. In both the survey and interview data, there was an overall consensus that virtual care is a satisfying alternative to in-person care, with benefits such as reduced travel, cost, time, and SARS-CoV-2 exposure, and increased convenience. Our analysis further revealed that the specific reason for the visit and the nature and status of the medical condition are important considerations in terms of guidance on where virtual care is most effective. Technology issues were not reported as a major challenge in our data, given that the majority of "virtual" visits reported by our participants were conducted by telephone, which is an important distinction. Despite the positive value of virtual care discussed by the majority of interview participants, 50% of the survey respondents still indicated they would prefer to see their physician in person.

Conclusions: Patient experience data collected in this study indicate a high level of satisfaction with virtual specialty care, but also signal that there are nuances to be considered to ensure it is an appropriate and sustainable part of the standard of care.

(JMIR Med Inform 2022;10(6):e37196) doi:10.2196/37196

KEYWORDS

virtual care; specialist care; patient experience; COVID-19; medical care; virtual health; care data; decision support; telehealth; video consultation

Introduction

At its broadest definition, virtual care is "the use of any technology (e.g., telephone, private messaging, videoconferencing) that supports health providers to collaborate with one another and to deliver remote care to patients" [1]. In response to the COVID-19 pandemic, the Ontario Ministry of Health released guidelines advising against direct patient care in nonurgent situations, and directed clinicians to transition the delivery of care to telephone- or video-enabled virtual visits online in early 2020 [2]. They also released unique billing codes to allow appropriate remuneration for these types of virtual visits [3]. Many professional colleges similarly encouraged their members to use virtual care wherever possible to minimize the risk of infection among their patients, especially those at higher risk of harm from COVID-19 infection [4,5]. In response to these guidelines and to ensure continuity of care, many hospitals and clinics shifted a majority of ambulatory visits to a virtual model. Increasing the ability to provide health care virtually not only supports social distancing and minimizes further potential spread of the virus but is also essential to ensure that patients continue to have access to medical guidance for nonemergent, but potentially serious, medical conditions. This is particularly true for urgent situations when the costs outweigh the benefits of bringing the person in for a physical visit (eg, immunosuppressed patients, those at high risk of infection, those with transportation issues).

Transitioning a large proportion of nonemergency, ambulatory medical care to virtual visits has been a massive shift in philosophy and practice that naturally came with a steep learning curve for patients, physicians, and clinic administrators. It has been widely reported that virtual care adoption has accelerated during the COVID-19 pandemic; however, published data remain limited. Previous studies have largely been limited to primary care, rural medicine, and nongeneralizable patient subgroups [6-8]. To provide optimal, safe care during the COVID-19 pandemic and beyond, we need to understand the key factors associated with successful and less successful experiences of virtual care. This is particularly important for areas newer to virtual care, such as specialist care delivery.

Increasingly, patient experience is recognized as an independent dimension of health care quality, along with clinical effectiveness and patient safety [9,10]. Accordingly, the aim of our study was to understand how patient experience data inform the development of guidance regarding characteristics associated with high satisfaction of virtual specialist care during the COVID-19 pandemic. This article specifically reflects our findings regarding the patient and family experience.

Methods

Study Setting

This study was conducted at St. Michael's Hospital, a quaternary-care teaching and research hospital in downtown Toronto (Ontario, Canada), and part of the Unity Health system. As downtown Toronto's adult trauma center, the hospital is a hub for neurosurgery, complex cardiac and cardiovascular care, complex medical specialty care, and care of the homeless and disadvantaged. Fully affiliated with the University of Toronto, St. Michael's Hospital provides medical education to health care professionals in 29 academic disciplines.

Study Design

This study was designed as a multimethod patient experience study using survey methods, descriptive qualitative interview methodology, and administrative data collected by the hospital decision support team.

Ethics Approval

Institutional review board approval for this study was obtained from the Human Research Ethics Board of St. Michael's Hospital (REB #20-198). All participants were given time to review the project information letter and provided written or verbal consent prior to the start of data collection. This report was compiled in compliance with the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines for qualitative research reporting [11].

Administrative Virtual Care Data

Deidentified administrative data on ambulatory visits (in-person, phone, and video) were retrospectively accessed via the admission-discharge-transfer systems at the hospital and categorized by clinical discipline (eg, neurology, orthopedic surgery) according to our internal coding structure. Visits to the emergency department, for day surgeries, and for diagnostic investigations other than those performed directly by a physician (eg, endoscopy) were excluded from the analysis.

Sampling and Recruitment for Surveys and Interviews

A wide variety of ambulatory patients from both medical and surgical specialties were included in this study. Participating specialty departments included endocrinology, orthopedic surgery, neurology, hematology, rheumatology, and gastroenterology. This ensured that we were able to capture patients seen for a range of presenting complaints as well as appointment types. All patients who could speak and read English (or understand with help) and attended a virtual appointment in a participating clinic between October 1, 2020, and January 30, 2021, were eligible to participate.

Participants were recruited during their clinical visit with a participating physician. Following the appointment, the physician asked the patient if they would be willing to be contacted regarding a research survey and interviews about their experience with the virtual visit.

Survey Data Collection and Analysis

During phase I, patient satisfaction with virtual visits was assessed through completion of an online survey made available through the SurveyMonkey platform (see Multimedia Appendix 1). At study initiation, there were no validated survey instruments to evaluate virtual medical care; thus, we adapted an instrument previously used to evaluate patient care visits provided through the Ontario Telemedicine Network [12]. The survey was only available in English due to time and resource constraints. The final question of the survey invited participants to provide their contact information if they would be willing to participate in a more in-depth telephone interview.

Survey data results are reported as simple counts or percentages. Logistic regression of variables was not possible owing to the low sample size.

Interview Data Collection and Analysis

Patients and families who consented to participate in an interview via the online patient survey (described above) were contacted by the research coordinator. Interviews were conducted by telephone to support social distancing restrictions and avoid geographic bias in recruitment.

The interviews followed a semistructured format using an interview guide informed by the study objective and addressing key topics, including understanding of the patient's functional status, health concerns, and experience of the virtual visit with their specialist (see Multimedia Appendix 2). The selection of follow-up questions, question order, and phrasing varied according to each participant's narrative. This approach enabled the emergence of participant-led accounts, reflecting their varied histories, modes of expression, and foci of experience. All interviews were conducted by an experienced qualitative interviewer (MBS), audio-taped, and transcribed verbatim by an external transcription service. The qualitative data were managed using NVivo (NVivo 12, QSR International Pty Ltd) qualitative software. The interview transcripts were supplemented with field notes to collect data that were not captured on audiotape (eg, dynamics, emotional aspects, contextual factors).

In keeping with the iterative process of qualitative methodology, data analysis occurred in conjunction with data collection to continuously monitor emerging themes and general areas for further exploration. We used a thematic analysis approach, as described by Braun and Clarke [13], to enable the identification of patterns and meaning across the sample. The analysis was led by two members of the research team with extensive qualitative expertise (KND and MBS) with regular collaboration with the rest of the study team. We extracted and collated the interview sections that reflected the key areas of interest and carried out the initial coding process. We then used the emergent codes to guide a de novo analysis of the entire corpus for overarching subthemes and used NVivo to record which subthemes occurred in each interview, ensuring their accurate representation in the analysis. Subthemes that expressed similar experiential patterns were brought together to develop the themes that we felt best represented the participants' perspectives. Versions of the analysis were reviewed with the research team at regular intervals, and the final analytic framework was discussed among all authors until we reached consensus on its validity and applicability. We employed the following techniques to support the analytic rigor and trustworthiness of our analysis: comparison of coding between analysts, seeking alternative explanations for the data during development of the final analytic framework, and interrogating the coherence of interpretations through discussion with the research team [14].

Results

Virtual Care at St. Michael's Hospital During the Study Period

The quantitative analysis results of virtual visit volumes at the study center are first described to provide context for the survey and interview findings. During the study period from April to December 2020, there were 593,172 total ambulatory visits at St. Michael's Hospital, 50.76% (n=301,105) of which were conducted virtually. Of note, only about 5176 visits (approximately 1.72%) were conducted virtually in the period of 2018-2019, before the pandemic, at the study hospital. In the specialty departments that participated in this study, 93,920 total ambulatory visits were conducted, over 50% of which were performed virtually. Among the virtual visits, over 96% were by phone. Some disciplines, notably orthopedic surgery and rheumatology, did not adopt virtual care as robustly as other specialties such as neurology and endoscopy. Neurology also had a much higher adoption of video visits than the other specialties, which may be attributed to their need for visual examination. Ambulatory volumes across disciplines were not linked to the uptake of virtual visits overall or by video specifically. A summary of the visit types for each specialty that participated in this study is provided in Table 1.

Table 1. Visit data for participating specialties (April 1, 2020, to January 30, 2021).

Specialty	Total visits, n	In-person visits, n (%)	Virtual visits		
			Total, n (%)	Phone visits, n (% of total, % of virtual)	Video visits, n (% of total, % of virtual)
Endoscopy	25,115	6992 (27.84)	18,163 (72.32)	18,012 (71.72, 99.39)	111 (0.44, 0.62)
Orthopedics	17,033	13,568 (79.66)	3465 (20.34)	3462 (20.33, 99.91)	3 (0.02, 0.09)
Neurology	18,774	6309 (33.60)	12,435 (66.34)	10,802 (57.54, 86.66)	1663 (8.87, 13.37)
Hematology	10,303	4413 (42.83)	5890 (57.17)	5889 (57.16, 99.83)	1 (0.01, 0.02)
Rheumatology	7224	4419 (61.17)	2805 (38.83)	2702 (37.40, 96.33)	103 (1.43, 3.67)
Gastroenterology	15,471	6517 (42.12)	8954 (57.88)	8815 (56.98, 98.45)	139 (0.90, 1.55)
Total	93,920	42,218 (44.95)	51,702 (55.05)	49,682 (52.90, 96.09)	2020 (2.15, 3.91)

Survey and Interview Sample

Between October 2020 and January 2021, 216 patients from the seven participating clinics at St. Michael's Hospital completed the virtual care experience survey. The large majority of the sample had attended their virtual visit for a follow-up appointment (vs an initial visit or emergency situation) and had previously attended a virtual medical appointment. Almost half of the sample was over 65 years of age and almost 60% identified as female. In the same time period, 125 patients agreed to participate in an interview and 18 patients with diverse characteristics, health conditions, and virtual visit types were selected for interviews. Detailed demographics of the participants are provided in Table 2.

Table 2. Survey and interview participant demographics.

Demographic characteristics	Survey (n=216), n (%)	Interviews (n=18), n (%)
Gender		
Female	125 (58.1)	11 (61)
Nonbinary	1 (0.5)	0 (0)
Education		
High school	27 (12.7)	4 (22)
Postsecondary diploma/degree	81 (43.6)	9 (50)
Graduate degree	39 (18.3)	3 (17)
Health care professional	8 (3.7)	0 (0)
Professional school	23 (10.6)	2 (11)
Age (years)		
18-34	14 (6.5)	1 (6)
35-54	53 (24.8)	7 (39)
55-80	125 (60.8)	10 (56)
80+	17 (7.9)	0 (0)
Location of birth		
In Canada	131 (61.2)	12 (67)
Outside Canada	83 (38.8)	6 (33)
First virtual care visit	68 (31.9)	11 (61)
Type of appointment		
First visit with specialist	19 (8.9)	4 (22)
Follow-up visit	181 (85.4)	14 (78)
Other	12 (5.7)	
Phone call	186 (86.9)	13 (72)
Video call	25 (11.6)	3 (17)
Telephone and video call	0 (0)	2 (11)

Survey Results

Survey respondents overwhelmingly had a very positive experience with virtual care. Almost 87% of people surveyed indicated that their virtual visit had been conducted by telephone (rather than video conference). They reported feeling comfortable connecting with their physician virtually, felt the physician spent sufficient time with them, and that their privacy was respected during the virtual call. Very few (3.8%) needed help with their virtual appointment or experienced technical difficulties during the visit (6.6%). However, despite 93% of respondents being satisfied with their virtual care experience, 50% still reported that they would prefer to see their physician in person if it were safe to do so. In addition, only 68% felt that the physician-patient relationship was the same as during an in-person visit.

When asked more generally about their opinion of virtual care, 25% were still unsure if virtual care is an acceptable way to provide health care for an initial consultation, but the majority agreed it was acceptable for follow-up visits (86%) and to discuss test results (85%). Full survey results are provided in Table 3 and Table 4.

Table 3. Survey results for questions scored on a 5-point scale (N=216).

Survey questions	Strongly disagree (1), n (%)	Disagree (2), n (%)	Neither agree nor disagree (3), n (%)	Agree (4), n (%)	Strongly agree (5), n (%)
I was comfortable connecting with my physician virtually (phone/video)	2 (0.9)	0 (0)	5 (2.3)	78 (36.6)	128 (60.1)
My privacy was respected	2 (0.9)	0 (0)	5 (2.4)	58 (28.0)	142 (68.6)
I felt that my physician spent sufficient time with me	3 (1.4)	3 (1.4)	6 (2.8)	65 (30.4)	137 (64.0)
My telephone/video assessment was thorough	2 (0.9)	3 (1.4)	17 (8.1)	81 (38.4)	108 (51.2)
I left the virtual visit with a clear understanding of the next steps	2 (0.9)	4 (1.9)	8 (3.8)	74 (34.9)	124 (58.5)
Compared to an in-person visit, the physician-patient relationship was the same	4 (1.9)	29 (13.7)	35 (16.5)	70 (33.0)	74 (34.9)
Having a virtual visit saved me time	2 (0.9)	3 (1.4)	13 (6.1)	55 (25.9)	139 (65.6)
I experienced technical difficulties during my appointment	132 (62.3)	59 (27.8)	7 (3.3)	10 (4.7)	4 (1.9)
I needed help with my virtual visit from a family member or friend	154 (72.6)	40 (18.9)	10 (4.7)	4 (1.9)	4 (1.9)
If it were safe to do so, I would prefer to meet with my care provider in person	14 (6.6)	19 (9.0)	72 (34.0)	62 (29.3)	45 (21.2)
I was satisfied with my virtual visit	2 (0.9)	2 (0.9)	10 (4.7)	85 (40.3)	112 (53.1)

Table 4. Survey results for questions scored on a 3-point scale.

Survey questions	Agree, n (%)	Disagree, n (%)	Not sure, n (%)
A virtual visit is an acceptable way to provide care for an initial consultation (N=211)	77 (36.5)	83 (39.3)	51 (24.2)
A virtual visit is an acceptable way to provide care for a routine follow-up appointment (N=214)	185 (86.5)	13 (6.1)	16 (7.5)
A virtual visit is an acceptable way to discuss test results (N=214)	182 (85.1)	5 (2.3)	27 (12.6)
A virtual visit is an acceptable way to provide an urgent follow-up assessment (N=216)	118 (55.4)	48 (22.5)	47 (22.1)

Interview Results

Overview

We interviewed 18 patients who had a minimum of one virtual specialist appointment at St. Michael's Hospital. The majority of interview participants were female, ranged in age from 45 to 64 years, and over one third reported having seen more than one type of specialist virtually during the COVID-19 pandemic (Table 2). Overall, participants were extremely happy with the opportunity to connect virtually with their physicians and were generally very satisfied with the appointments conducted. Our qualitative data analysis revealed three key themes that provide a deeper understanding of participants' experience of virtual care: (1) the impact of improved access, (2) influence of the nature of the visit, and (3) consideration of the nature of the medical condition.

Impact of Improved Access

Due to the nature of St. Michael's Hospital as a large quaternary-care academic health science center, people frequently travel from outside Toronto to see the specialty physicians affiliated with the hospital. Many of the patients we spoke with mentioned that their high satisfaction with virtual care was driven by being able to access their specialists without the nuisance of the potentially long trip to the city.

> *Yeah, it's great. Because just for us to go to Toronto, you know, there's always an overnight. Because I can't go there and back in one day. It becomes an expensive journey just to go to the hospital for a follow-up.* [Participant 6]

The COVID-19 pandemic appeared to amplify preexisting travel challenges for patients who were very wary of travelling to a large urban center for fear of exposure to the virus. Participants told us that it was important to be able to access their specialists despite the pandemic and the option of virtual appointments met that need. They expressed significant concerns about the risk to themselves and their caregivers/family members of exposure to SARS-CoV-2 from coming into the city in person, and were very grateful for the opportunity to keep their appointments using the virtual medium.

> *I don't see how the consult I had on the phone was going to be any different than going to the place. In*

fact, it felt safer right now during a pandemic to not have to leave the house and not have to go into the hospital. [Participant 17]

Influence of Reason for the Visit

Many of the participants had seen more than one type of specialist during the pandemic, and thus were able to draw on significant experience of virtual care during the interviews. Participants told us they felt that virtual care is acceptable for certain types of appointments such as initial consults that are conversation-based (to discuss and explain requirements for further tests, examinations, and procedures), routine follow-up for stable conditions, to review tests results, and whenever the interaction between the patient and specialist would mostly be question and answer–based to support the provision of information and advice. However, for appointments that would typically involve a physical examination, when they are experiencing pain, or if their health condition has progressed or changed, they explained that they would prefer to be assessed and have an opportunity to speak with their specialist in person. For these types of appointments, participants perceived virtual care to come with higher potential for misinterpretation or a misdiagnosis than in-person care, and that this could ultimately impact the trajectory of their treatment.

So, I think it depends on the appointment. If it's just a routine follow-up to go through test results, that's fine. But if it's an actual, "Hey, you know I'm not feeling well. This is what is going on," I'd prefer in person so they can actually touch it or see it. [Participant 5]

It's good but I'm really not sure if what she's seeing is right, because it's different than what I'm feeling. And I think that part was a little bit frustrating or worrisome, because I don't want something to be inaccurately marked and somehow – like that has the potential to affect my care. [Participant 10]

Conversely, patients spoke about virtual appointments being more efficient for themselves and their physicians. Some mentioned the time savings of not having to take time off work or other responsibilities for routine follow-up appointments, and noted that their physicians seemed to be "more prepared" for the virtual appointments than they typically were for previous in-person visits. It was felt that this greater familiarity with the patient's chart and recent test results made for a better discussion about their state of health and current care needs.

It's not only that [referring to time savings], but I'm actually more confident. Because when XX phones me, she has reviewed the file, knows what she's going to say, and off I go. Previously, I would sit there while she reviewed the file on the screen and got up to date on it. This way, she gets up to date on her own speed, and when I talk to her, it's usually a very brief interview. She tells me what she sees. I ask her questions and it's over. No, no. This is much, much better. [Participant 1]

Considering the Nature and Status of the Medical Condition

Several interview participants had long, complicated medical histories and chronic condition(s). For the most part, these participants still felt that the virtual appointments had met their care needs, and told us they were satisfied with both the quality of care received and the interaction they had with their physicians. However, these were the same patients who most frequently expressed some hesitation around virtual care, explaining that they did not believe that the specific health issues they have, including rashes, vision-related problems, and tremors, could be assessed clearly through video or adequately described by phone.

Because rheumatology is such a hands-on profession, I think they can't assess your joints at all over Zoom or OTN [Ontario Telemedicine Network]. I find that their rheumatologists are probably missing quite a bit because they can't get the information that they would normally get. The neurologist…it's probably also missing that physical component because they can't assess your tremors, or your eye tracking, or your reflexes. They can't do any of that. The conversations have been kind of restricted to things like which drugs we're on…like which drugs I'm on and kind of skipping over that physical component. [Participant 13]

For the participants managing comorbidities and complex chronic health problems, an annual check-up appointment by telephone or online was perceived to be acceptable so long as their conditions remain stable. To discuss changes in their health status and treatment options, these patients definitely preferred the option to see their physician in person.

I know my condition well. If everything is going well and I am stable then the phone appointments are fine. When there are specific flare-ups or…my blood work is off for too long – then he needs to see me. [Participant 15]

Discussion

Principal Findings

Using survey and qualitative interview methods, we examined the experiences of patients accessing virtual specialist care at a quaternary-care center in Toronto, Ontario, Canada. Our results indicate that, overall, patients were very satisfied with the quality and efficiency of virtual care, and value it as an option for safe and equitable access to specialist care. However, our data also revealed nuances to that value, which are important to take into account as we consider virtual care as a permanent care delivery option.

There is an increasing number of publications on the virtual experiences of patients, most of which have been conducted in singular fields such as oncology, pre- and postnatal care, and psychiatric care [15-20]. Similar to our findings, in the existing literature, there appears to be a consensus that virtual care is a satisfying alternative to in-person care, with benefits such as reduced travel, cost, time, and infectious exposure, and increased

convenience [15-20]. Our analysis further revealed that the specific reason for the visit, and the nature and status of the medical condition are important considerations in terms of guidance on where virtual care is most effective. Previous barriers or challenges identified also included technological issues, a potential lack of personal connection, inability to perform physical and visual assessments, and inequities in access [20,21]. Although some of these align with our data, we did not hear specifically about technological issues or a lack of personal connection. This likely can be attributed to the fact that the majority of the visits in our study sample were conducted by telephone (vs video or a new virtual platform), and that many of our participants had longstanding relationships with their specialists and therefore the personal connection may be stronger. The remuneration of both telephone and video care at equivalent rates in Ontario, Canada, may play a factor in the lack of uptake of video conferencing given reports from other jurisdictions (ie, telephone may be perceived as more cost-efficient) [22,23].

A few findings were notable in our data. First, as mentioned above, a large majority of "virtual" visits reported by participants were conducted by telephone, which is an important distinction from previous reports. Virtual care has been defined as:

> *any interaction between patients and/or members of their circle of care, occurring remotely, using any forms of communication or information technologies, with the aim of facilitating or maximizing the quality and effectiveness of patient care [24].*

However, it seems that discussion of virtual care often assumes a technology- or video-based component. Other terms that have been used in this literature include "telehealth" or "telemedicine," which may be more representative of the preferred medium. This finding is noteworthy in terms of unpacking held assumptions about what is possible with virtual care and understanding existing virtual care infrastructure. The availability, familiarity, and technical ease of telephone-based care (for both the provider and recipient) likely contributed to the overall predominance of telephone visits. Further investigation is required to determine the relative advantages and drawbacks of the different modalities for delivering virtual care.

Second, despite the positive value of virtual care discussed by the majority of interview participants, 50% of the survey respondents still indicated they would prefer to see their physician in person. This could reflect the fact that at the time of the survey, virtual care may have still been seen as a COVID-19–related intervention rather than a potentially permanent option for care delivery. In addition, only 68% of respondents felt that the physician-patient relationship was the same as during an in-person visit. These findings highlight that there still may need to be a cultural shift before there is complete comfort with virtual options as part of the standard of care in Ontario.

Our study adds to the growing literature on virtual care in the era of the COVID-19 pandemic and helps to inform virtual care implementation beyond the pandemic as well. As the pandemic has evolved, it has become clear that enhanced infection prevention protocols are likely to remain at some level in ambulatory care [25]. Moreover, now that the benefits of virtual care have been experienced by patients and physicians, it is likely that demand for virtual care will continue beyond the pandemic. As such, we need to use experience data such as those presented here to understand how virtual care performs in real time from both the perspective of those who access care and those who deliver it. Based on our data, we recommend a flexible, blended care model utilizing virtual care and in-person visits based on the type of appointment (eg, new patient, routine follow-up, assessment of new problem), patient preference, travel burden, and infection considerations. One size will not fit all, and a blended model combining in-person and virtual visits when tailored to each patient and visit is consistent with a patient-centered approach to care delivery. Virtual visits (including telephone visits) appear to be particularly well-suited for routine follow-up appointments focused on nonurgent matters. Virtual visits are also valued by patients when travel is either too costly or burdensome. However, further work is required to delineate the balance between telephone- versus video-based virtual care and to determine which types of care visits are most effective virtually. As new virtual technologies and systems emerge, such as secure texting and remote monitoring, it will also be important to reevaluate the benefits and drawbacks of each approach, and to ensure that privacy, confidentiality, and data quality are maintained. Prior to COVID-19, virtual care was largely managed as a distinct service model from in-person services; however, it is clear from learnings throughout the pandemic, including our work, that virtual and traditional care are complementary and that patients need the flexibility to seamlessly transition between both modalities. Considerable further work around quality, safety, convenience, preference, and appropriateness needs to occur so that the decisions on what modalities to offer and in what circumstances are evidence-informed. We must also evaluate the impact of virtual care on health care equity, as not all patients will have access to the same technology, and we must ensure that socioeconomic disparities are not widened or exacerbated by the adoption of virtual health care options.

Strengths and Limitations

This study represents a robust and diverse sample of patient responses, including diversity in gender, immigrant status, and specialty care provided. We also leveraged the strength of multiple data collection methods to be able to both capture the experience of a large number of patients while at the same time being able to gather a deeper understanding through the individual interviews.

Despite its strengths, the study does have some limitations. We only performed the data collection within a single hospital site located in downtown Toronto, Canada. While our results may not be generalizable to more rural or remote locations, we do feel that our site represents a fairly typical tertiary health science center, and therefore the insights learned here should be useful to other centers.

Use of a web-based survey prevented us from recruiting patients without an email address, which may have biased our sample

toward respondents with higher digital literacy and educational attainment. The survey and interviews were only conducted in English, which also may have introduced a language bias to the sample. Both of these methodologic choices may also explain a lower survey sample size than we expected. In addition, as with all patient-reported and qualitative data, there is some level of volunteer bias. That said, the survey was offered to all patients who participated in a virtual visit (physicians did not hand select who would receive the survey). Volunteer bias in surveys and interviewing is almost impossible to avoid; however, we used rigorous qualitative methodology to ensure we recruited a balanced and saturated sample for the interviews. Lastly, the predominance of telephone and follow-up visits in our data, while reflective of the "real-life" use of virtual care during the pandemic, limits the ability to draw conclusions about video visits or other care settings (eg, urgent care).

Conclusion

Providing alternative ways for providers and patients to deliver and access high-quality specialist care has become a necessity during the COVID-19 pandemic; however, the need and preference for virtual care are likely to only increase in the future. The patient experience data captured in this study indicate a high level of satisfaction with virtual specialist care, but also signal that there are nuances to be considered to ensure it is an appropriate and sustainable part of the standard of care. This type of multimethod, patient-oriented research combined with provider experience insight will be crucial in informing realistic guidance for health care systems across Canada.

Acknowledgments

This study was funded by the St. Michael's Alternative Funding Plan Research Fund. The authors would like to acknowledge the patients and family members who gave their time to complete the survey and participate in the interviews. We are very grateful for their willingness to help us understand this evolving model of care.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study survey.
[PDF File (Adobe PDF File), 129 KB - medinform_v10i6e37196_app1.pdf]

Multimedia Appendix 2
Patient interview guide.
[DOCX File , 17 KB - medinform_v10i6e37196_app2.docx]

References

1. Waddell K, Scallan E, Wilson MG. Understanding the use of and compensation for virtual-care services in primary care. McMaster Health Forum. 2018 Jul 27. URL: https://www.mcmasterforum.org/find-evidence/products/project/understanding-the-use-of-and-compensation-for-virtual-care-services-in-primary-care [accessed 2022-02-01]
2. COVID-19 Guidance for the Health Sector. Government of Ontario Ministry of Health and Ministry of Long-Term Care. URL: https://www.health.gov.on.ca/en/pro/programs/publichealth/coronavirus/2019_guidance.aspx [accessed 2022-02-01]
3. A comprehensive guide to OHIP billing codes for virtual care and COVID-19. DoctorCare. URL: https://www.doctorcare.ca/a-comprehensive-guide-to-ohip-billing-codes-for-virtual-care-and-covid-19/#:~:text=download%20it%20here.-,2.,%2C%20by%20telephone%2C%20or%20video [accessed 2022-02-01]
4. COVID-19 FAQs for physicians. College of Physicians and Surgeons of Ontario (CPSO). URL: https://www.cpso.on.ca/Physicians/Your-Practice/Physician-Advisory-Services/COVID-19-FAQs-for-Physicians [accessed 2022-02-01]
5. COVID-19 basic resources and guidance. Registered Nurses' Association of Ontario (RNAO). URL: https://rnao.ca/covid19/covid-19-basic-resources-and-guidance [accessed 2022-01-31]
6. Bhatia RS, Chu C, Pang A, Tadrous M, Stamenova V, Cram P. Virtual care use before and during the COVID-19 pandemic: a repeated cross-sectional study. CMAJ Open 2021;9(1):E107-E114 [FREE Full text] [doi: 10.9778/cmajo.20200311] [Medline: 33597307]
7. Baum A, Kaboli PJ, Schwartz MD. Reduced in-person and increased telehealth outpatient visits during the COVID-19 pandemic. Ann Intern Med 2021 Jan;174(1):129-131 [FREE Full text] [doi: 10.7326/M20-3026] [Medline: 32776780]
8. Alexander GC, Tajanlangit M, Heyward J, Mansour O, Qato DM, Stafford RS. Use and content of primary care office-based vs telemedicine care visits during the COVID-19 pandemic in the US. JAMA Netw Open 2020 Oct 01;3(10):e2021476 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.21476] [Medline: 33006622]
9. Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. BMJ Open 2013 Jan 03;3(1):e001570 [FREE Full text] [doi: 10.1136/bmjopen-2012-001570] [Medline: 23293244]

10. Committee on Quality of Health Care in America, Institute of Medicine. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: National Academies Press; 2001.
11. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
12. Morgan DG, Kosteniuk J, Stewart N, O'Connell ME, Karunanayake C, Beever R. The telehealth satisfaction scale: reliability, validity, and satisfaction with telehealth in a rural memory clinic population. Telemed J E Health 2014 Nov;20(11):997-1003 [FREE Full text] [doi: 10.1089/tmj.2014.0002] [Medline: 25272141]
13. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan 23;3(2):77-101 [FREE Full text] [doi: 10.1191/1478088706qp063oa] [Medline: 34554280]
14. Guba EG. Criteria for assessing the trustworthiness of naturalistic inquiries. ECTJ 1981 Jun;29(2):103251 [FREE Full text] [doi: 10.1007/BF02766777] [Medline: 35059193]
15. Natafgi N, Childers C, Pollak A, Blackwell S, Hardeman S, Cooner S, et al. Beam me out: review of emergency department telepsychiatry and lessons learned during COVID-19. Curr Psychiatry Rep 2021 Oct 06;23(11):72 [FREE Full text] [doi: 10.1007/s11920-021-01282-4] [Medline: 34613436]
16. Liu CH, Goyal D, Mittal L, Erdei C. Patient satisfaction with virtual-based prenatal care: implications after the COVID-19 pandemic. Matern Child Health J 2021 Nov;25(11):1735-1743 [FREE Full text] [doi: 10.1007/s10995-021-03211-6] [Medline: 34410565]
17. Brady G, Ashforth K, Cowan-Dickie S, Dewhurst S, Harris N, Monteiro A, et al. An evaluation of the provision of oncology rehabilitation services via telemedicine using a participatory design approach. Support Care Cancer 2022 Mar;30(2):1655-1662 [FREE Full text] [doi: 10.1007/s00520-021-06552-8] [Medline: 34554280]
18. Chadha RM, Paulson MR, Avila FR, Torres-Guzman RA, Maita K, Garcia JP, et al. Surgical patient satisfaction with a virtual hybrid care hotel model: a retrospective cohort study. Ann Med Surg 2022 Mar;74:103251 [FREE Full text] [doi: 10.1016/j.amsu.2022.103251] [Medline: 35059193]
19. Han L, Hazlewood GS, Barnabe C, Barber CEH. Systematic review of outcomes and patient experience with virtual care in rheumatoid arthritis. Arthritis Care Res 2021 Mar 01;27(1):21-26 [FREE Full text] [doi: 10.1002/acr.24586] [Medline: 33650316]
20. Nanda M, Sharma R. A review of patient satisfaction and experience with telemedicine: a virtual solution during and beyond COVID-19 pandemic. Telemed J E Health 2021 Dec;27(12):1325-1331. [doi: 10.1089/tmj.2020.0570] [Medline: 33719577]
21. Ball E, Rivas C, Khan R. If virtual gynecology clinics are here to stay, we need to include everyone. AJOG Glob Rep 2022 Mar;2(1):100043 [FREE Full text] [doi: 10.1016/j.xagr.2021.100043] [Medline: 34909705]
22. Mehrotra A, Bhatia RS, Snoswell CL. Paying for telemedicine after the pandemic. JAMA 2021 Feb 02;325(5):431-432. [doi: 10.1001/jama.2020.25706] [Medline: 33528545]
23. Rodriguez JA, Betancourt JR, Sequist TD, Ganguli I. Differences in the use of telephone and video telemedicine visits during the COVID-19 pandemic. Am J Manag Care 2021 Jan;27(1):21-26 [FREE Full text] [doi: 10.37765/ajmc.2021.88573] [Medline: 33471458]
24. Shaw J, Jamieson T, Agarwal P, Griffin B, Wong I, Bhatia RS. Virtual care policy recommendations for patient-centred primary care: findings of a consensus policy dialogue using a nominal group technique. J Telemed Telecare 2018 Oct;24(9):608-615. [doi: 10.1177/1357633X17730444] [Medline: 28945161]
25. Wei EK, Long T, Katz MH. Nine lessons learned from the COVID-19 pandemic for improving hospital care and health care delivery. JAMA Intern Med 2021 Jul 23:online ahead of print. [doi: 10.1001/jamainternmed.2021.4237] [Medline: 34297056]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research

Edited by C Lovis; submitted 15.02.22; peer-reviewed by M Levine, R Boumans; comments to author 11.04.22; revised version received 25.04.22; accepted 27.04.22; published 28.06.22.

Please cite as:
Dainty KN, Seaton MB, Estacio A, Hicks LK, Jamieson T, Ward S, Yu CH, Mosko JD, Kassardjian CD
Virtual Specialist Care During the COVID-19 Pandemic: Multimethod Patient Experience Study
JMIR Med Inform 2022;10(6):e37196
URL: https://medinform.jmir.org/2022/6/e37196
doi:10.2196/37196
PMID:35482950

Original Paper

Conditional Probability Joint Extraction of Nested Biomedical Events: Design of a Unified Extraction Framework Based on Neural Networks

Yan Wang[1], PhD; Jian Wang[1], PhD; Huiyi Lu[2], PhD; Bing Xu[2], PhD; Yijia Zhang[3], PhD; Santosh Kumar Banbhrani[1], PhD; Hongfei Lin[1], PhD

[1]School of Computer Science and Technology, Dalian University of Technology, Dalian, China

[2]Department of Pharmacy, The Second Affiliated Hospital of Dalian Medical University, Dalian, China

[3]School of Information Science and Technology, Dalian Maritime University, Dalian, China

Corresponding Author:
Jian Wang, PhD
School of Computer Science and Technology
Dalian University of Technology
No 2 Linggong Road
Dalian, 116024
China
Phone: 86 13604119266
Email: wangjian@dlut.edu.cn

Abstract

Background: Event extraction is essential for natural language processing. In the biomedical field, the nested event phenomenon (event A as a participating role of event B) makes extracting this event more difficult than extracting a single event. Therefore, the performance of nested biomedical events is always underwhelming. In addition, previous works relied on a pipeline to build an event extraction model, which ignored the dependence between trigger recognition and event argument detection tasks and produced significant cascading errors.

Objective: This study aims to design a unified framework to jointly train biomedical event triggers and arguments and improve the performance of extracting nested biomedical events.

Methods: We proposed an end-to-end joint extraction model that considers the probability distribution of triggers to alleviate cascading errors. Moreover, we integrated the syntactic structure into an attention-based gate graph convolutional network to capture potential interrelations between triggers and related entities, which improved the performance of extracting nested biomedical events.

Results: The experimental results demonstrated that our proposed method achieved the best F1 score on the multilevel event extraction biomedical event extraction corpus and achieved a favorable performance on the biomedical natural language processing shared task 2011 Genia event corpus.

Conclusions: Our conditional probability joint extraction model is good at extracting nested biomedical events because of the joint extraction mechanism and the syntax graph structure. Moreover, as our model did not rely on external knowledge and specific feature engineering, it had a particular generalization performance.

(JMIR Med Inform 2022;10(6):e37804) doi:10.2196/37804

KEYWORDS

nested biomedical event; joint extraction; graph convolutional network; GCN; Dice loss; syntactic structure

Introduction

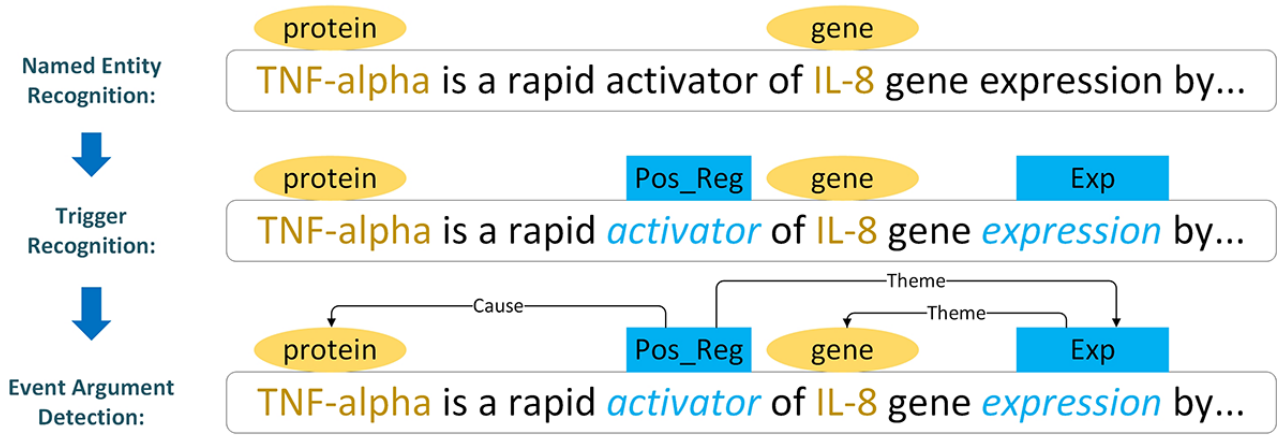
Background

In recent years, event extraction research has attracted wide attention, especially in biomedical event extraction, which is critical for understanding the biomolecular interactions described in the scientific corpus. Events are important concepts in the field of information extraction. However, researchers have different definitions of events, based on different research purposes and perspectives. In the general domain, an event is

a specific thing that describes a state change involving different participants, such as the evaluation of automatic content extraction, in which 8 categories and 33 subcategories of events are defined in a hierarchical structure, and each type of event contains a different semantic role. In the biomedical field, McDonald et al [1] defined event extraction as multirelationship extraction, the purpose of which was to extract semantic role information between different entities in an event. For example, the biomedical natural language processing (BioNLP) evaluation task defined 9 different categories of biochemical events. Each event included an event trigger and at least one event argument, and the different event types had different semantic roles. Unlike the events in automatic content extraction, biomedical events may have nested event phenomena.

To clearly describe the progress of biomedical event extraction, we defined 4 concepts for biomedical events, as shown in Figure 1 and Textbox 1.

Figure 1. Basic progress of biomedical event extraction, where yellow boxes represent the type of entity and the blue boxes represent the type of trigger. Theme and cause represent the relationship between participant and event, namely, argument detection. IL-8: interleukin 8; TNF-alpha: tumor necrosis factor.



Textbox 1. Concepts for biomedical events.

Event type

The semantic type of different events

Event description

A complete sentence or clause in the text that specifically describes at least one event

Event trigger

A word or phrase representing the occurrence of an event in the event description; usually of a *verb* or *nonverb* nature, and its category is event type; it should be noted that each event has only 1 event trigger.

Event argument

The event participants describe the different semantic roles in the event, whose type represents the relationship between the event and related participants; in the biomedical event system, there are 6 different semantic roles, where "theme" and "cause" are core arguments.

The task of event extraction comprises 3 subtasks: named entity recognition, trigger recognition, and event argument detection. Previous studies have relied on pipeline methods [2-5] to extract biomedical events. For example, given the event description (a sentence) shown in Figure 1, the event extraction system can find 2 entities ("TNF-alpha" and "IL-8") in this sentence at the named entity recognition step. After recognizing triggers, it can identify a *positive regulation* ("Pos_Reg") event mention triggered by a word *activator* and an *expression* ("Exp") event mention triggered by a word *expression*. On the basis of the recognized entities and triggers, the system detects arguments and associates them with the related event triggers. Thus, the entity "TNF-alpha" is a participant in the *positive regulation* event, and the entity "IL-8" is a participant in the *expression* event. As the result of the previous step is the input of the subsequent step, the pipeline methods probably introduce cascading errors if the precision of the previous step is biased.

As the syntactic dependency tree enriches the feature representation, previous studies tended to use syntactic relations to improve the performance of event extraction. For example, Kilicoglu et al [2] leveraged external tools to segment sentences, annotate parts of speech (POS), and parse syntactic dependency. Then, they joined these features to extract biomedical events using a dictionary and rules. Björne et al [4] transferred the syntactic relations to the path embeddings, then combined them with word embeddings, POS embeddings, entity embeddings, distance embeddings, and relative position embeddings to feed into the convolutional neural network (CNN) model to extract biomedical events. However, the previous studies only adopted syntactic relations as the external features and ignored the interrelations between triggers and related entities obtained from the syntactic dependency tree, which improved the performance of extracting simple events but not nested events.

In this study, we mainly used the multilevel event extraction (MLEE) corpus [6] and the BioNLP shared task (BioNLP-ST) 2011 Genia event (GE) corpus [7] to evaluate our method. There is some explanation regarding the MLEE extending event extraction methods to the biomedical information field and covering all levels of biological tissue from molecules to entire organisms. The MLEE label scheme is the same as the BioNLP event system but has more abundant event types: 4 major categories (anatomical, molecular, general, and planned) and 19 subcategories. The specific information is shown in Table 1.

Table 1. Primary event types and argument roles in the multilevel event extraction corpus (N=6827).

Event and subevent types	Core arguments	Values, n (%)
Anatomical		
Cell proliferation	Theme (entity)	133 (2.42)
Development	Theme (entity)	316 (4.81)
Blood vessel development	Theme (entity)	855 (12.91)
Growth	Theme (entity)	469 (2.65)
Death	Theme (entity)	97 (1.53)
Breakdown	Theme (entity)	69 (1.1)
Remodeling	Theme (entity)	33 (0.45)
Molecular		
Synthesis	Theme (entity)	17 (0.3)
Gene expression	Theme (entity)	435 (6.66)
Transcription	Theme (entity)	37 (0.61)
Catabolism	Theme (entity)	26 (0.39)
Phosphorylation	Theme (entity)	33 (0.5)
Dephosphorylation	Theme (entity)	6 (0.09)
General		
Localization	Theme (entity)	450 (6.87)
Binding	Theme (entity)	187 (2.92)
Regulation	Theme (entity or event) and cause (entity or event)	773 (11.81)
Positive regulation	Theme (entity or event) and cause (entity or event)	1327 (20.33)
Negative regulation	Theme (entity or event) and cause (entity or event)	921 (14.08)
Planned		
Planned process	Theme (entity or event)	643 (9.9)

To abate the impact of cascading errors, we propose an end-to-end conditional probability joint extraction (CPJE) method that can effectively transmit trigger distribution information to the event argument detection task. To capture the interrelations between triggers and related entities and improve the performance of extracting nested biomedical events, we integrated the syntactic dependency tree into an attention-based gate graph convolutional network (GCN), which can capture the flow direction of the key information. The contributions of this study are as follows:

1. We propose an end-to-end CPJE framework, CPJE, which effectively leverages trigger distribution information to enhance the performance of event argument detection and weakens cascading errors in the overall event extraction process.
2. We used the syntactic dependency tree to capture the interrelations between triggers and related entities and integrated the tree into an attention-based gate GCN to extract nested biomedical events.
3. We obtained state-of-the-art performance on the MLEE and BioNLP-ST 2011 GE corpora for extracting nested biomedical events.

We summarize the current frameworks for event extraction tasks in the *Related Works* section. We introduce our framework in the *Methods* section. We display the overall performance in the *Results* section. We present the ablation study, visualization, and case study in the *Discussion* section. We summarize this work and discuss future research directions in the *Conclusions* section.

Related Works

The biomedical event extraction problem is similar to general domain event extraction and entity relationship extraction; therefore, we have many theoretical foundations and experimental methods that can be used for reference.

Entity Relationship Extraction

Biomedical events can be regarded as complex relationship extraction tasks, and relationship extraction methods have achieved excellent results in various fields. Therefore, we studied some relationship extraction methods to help conceive the construction of event extraction models. With the development of deep learning, an increasing number of researchers have used deep learning algorithms to achieve the joint extraction of entity relationships [8]. To solve the problem of a sparse number of labeled samples, distant supervision methods have been applied to the relationship extraction task [9]. Deep reinforcement learning (RL) algorithms have also been applied to the relationship extraction task to solve noisy data samples [10]. In addition, with the widespread application of graph neural networks (GNNs), GCNs have been used in certain relation-extraction tasks [11,12].

General Domain Event Extraction

In general, news event extraction is a research hot spot. Some methods have improved the performance of event extraction by studying feature engineering. Sentence-level feature extraction included combinational features of triggers and event arguments [13] or combinational features of triggers and entity relationships [14]. Document-level feature extraction included common information event extraction from multiple documents [15] and joint event argument extraction based on latent-variable semi-Markov conditional random fields [16]. Others have also used deep learning to reduce feature engineering, which improves a model's generalization ability and extraction performance; for example, learning context-dependency information with recurrent neural networks [17], detecting events with nonconsecutive CNNs [18], and obtaining syntactic structure information with GCNs [19]. All these methods have laid a better foundation for the extraction of biomedical events.

Biomedical Event Extraction

Extracting biomedical events is one of the BioNLP-STs [7,20,21]. Previous studies mainly explored human-engineered features based on a support vector machine model [22-25]. Owing to error transmission in the pipeline approach, Riedel et al [26] developed a joint model with dual decomposition, and Venugopal et al [27] leveraged Markov logic networks for joint inference. Recently, most studies have observed remarkable benefits of neural models. For example, some have started to add POS tags and syntactic parsing with different neural models [28], improved the biomedical event extraction model using semisupervised frameworks [29], attempted to use attention mechanisms to obtain the semantic relationship of biomedical texts [5], and used distributed representations to obtain context embedding [3,4,30,31]. To incorporate more information from the biomedical knowledge base (KB), Zhao et al [32] leveraged a RL framework to extract biomedical events with representations from external biomedical KBs. Li et al [33] fused gene ontology into tree long short-term memory (LSTM) models with distributional representations. Huang et al [34] used a GNN to hierarchically emulate 2 knowledge-based views from the Unified Medical Language System with conceptual and semantic inference paths. Trieu et al [35] used multiple overlapping, directed, acyclic graph structures to jointly extract biomedical entities, triggers, roles, and events. Zhao et al [36] combined a dependency-based GCN with a hypergraph to jointly extract biomedical events. Ramponi et al [37] proposed a joint end-to-end framework that regards biomedical event extraction as sequence labeling with a multilabel aware encoding strategy.

Compared with these methods, our approach joint extracts the biomedical events with a probability distribution of triggers, which alleviates the cascading errors introduced by the pipeline methods. Moreover, considering the potential interrelations between triggers and related entities, our approach integrates the syntactic structure into an attention-based gate GCN to capture the flow direction of key information, which greatly improves the extraction performance for nested biomedical events. It is important to mention that our approach does not require any external resources to assist the biomedical event extraction task.

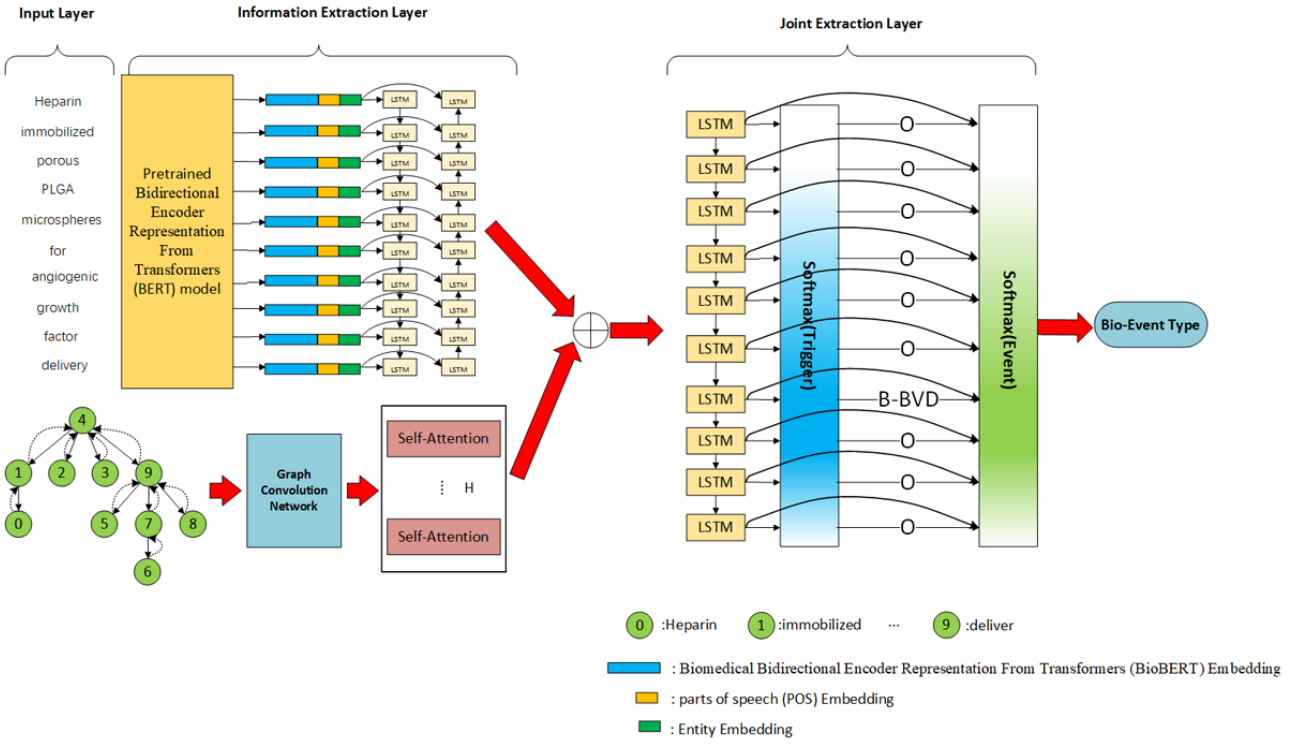
Methods

Overview

This section illustrates the proposed CPJE model. Let $W=\{w_1,w_2,...,w_n\}$ be a sentence of length n, where w_i is the ith word in a sentence. Similarly, $E=\{e_1,e_2,...,e_k\}$ is a set of entities mentioned in a sentence, where k is the number of entities. As the trigger may comprise multiple tokens, we used the BIO tag scheme to annotate the trigger type of each token in the sentence. When we obtained the corresponding event trigger in the sentence, we used this information to predict the corresponding event arguments.

As shown in Figure 2, our CPJE model mainly includes 3 layers: an input layer, an information extraction layer, and a joint extraction layer. The input layer converts unstructured text information (such as word sequences, syntactic structure trees, POS label representations, and entity label information) into a structured discrete representation and inputs it into the next layer. The information extraction layer converts discrete information into continuous feature representations, which deeply extracts the semantic and dependence information in a sentence. The joint extraction layer parses the previous fusion information and sends the parsed information into the trigger softmax classifier and event softmax classifier to jointly extract biomedical events.

Figure 2. The architecture of the conditional probability joint extraction framework, where numbers 0 to 9 represent each word in the sentence, the blue bar represents BioBERT embedding, the yellow bar represents POS-tagging embedding, and the green bar represents entity embedding. BERT: Bidirectional Encoder Representation From Transformers; BioBERT: Biomedical Bidirectional Encoder Representation From Transformers; B-BVD: B-blood vessel development; LSTM: long short-term memory; POS: parts of speech.



Information Extraction Layer

This is not explained in detail as the input layer was too superficial (only converting the text into a sequence of numbers). Each module of the information extraction layer is presented in the following sections.

Word Representation

In the word representation module, to improve the representation capability of the initial features, each word w_i in the sentence is transformed to a real-valued vector x_i by concatenating the embeddings described in the following sections.

Biomedical Bidirectional Encoder Representation From Transformers Embedding

We used the Biomedical Bidirectional Encoder Representation from Transformers (BioBERT) pretraining model [38] to obtain the dynamic semantic representation of the word w_i. BioBERT embedding comprises token embedding, segment embedding, and position embedding, which is encoded as a consequence by a multilayer bidirectional transformer. Thus, it includes rich semantic and positional information. Furthermore, it can solve the polysemy problem of words. We define a_i as the word vector representation of the word w_i.

POS-Tagging Embedding

We used a randomly initialized POS-tagging embedding table to obtain each POS-tagging vector. We defined b_i as the POS-tagging vector representation of the word w_i.

Entity Label Embedding

Similar to the POS-tagging embedding, we used the BIO label scheme to annotate the entities mentioned in the sentence and convert the entity type label into a real-value vector by consulting the embedding table. We defined c_i as the entity vector representation of the word w_i.

The transformation from the token w_i to the vector x_i converts the input sentence W into a sequence of real-valued vectors $X=\{x_1,x_2,...,x_n\}$, , where is the concatenation operation, x_i is the μ dimension (ie, the sum of the dimensions of a_i, b_i, and c_i), and . X is fed into the subsequent blocks to obtain more valuable information for extracting biomedical events.

Bidirectional LSTM

To obtain the context information of the input text and avoid the gradient explosion problem caused by long texts, we chose the classic bidirectional LSTM (BiLSTM) structure to extract the context features of the word representations.

We fed the word representation sequence $X=\{x_1,x_2,...,x_n\}$ into BiLSTM to obtain the forward hidden unit h_t^f and the backward hidden unit h_t^b with ϕ dimension in time t according to equation 1. We represented all the hidden states of the forward LSTM and backward LSTM as and , respectively, where n is the number of LSTM hidden units:

Finally, we concatenated these 2 matrices to obtain the context representation  of BiLSTM:

Gate GCN

To obtain the syntactic dependence in a sentence, we reference the method proposed by Liu et al [19] to apply a gate GCN model to analyze the sentence-dependent features. We considered an undirected graph G=(V, ε) as a syntactic dependency tree for the sentence W, where V is the set of nodes and ε is the set of edges. Defining  , v_i represents each word w_i of sentence W, and each edge  represents a directed syntactic arc from word w_i to word w_j, with dependency type Re. In addition, for the sake of moving information along the direction, we add the corresponding reversed edge (v_w, v_i) with dependency type Re' and self-loops (v_i, v_i) for any node v_i. According to statistics, we used the Stanford Parser [39] to obtain approximately 50 different kinds of syntactic dependency. To facilitate the GCN internal calculation, we only considered the direction of information flow and simplified the original dependency into 3 forms, as shown in equation 4:

For node  , we can use the hidden vector $h_v^{(j)}$ in the jth gate GCN layer to compute the hidden vector $h_v^{(j+1)}$ of the next layer:

where $Re(u,v)$ is the dependency type between nodes u and v, $W_{Re(u,v)}^{(j)}$ and $b_{Re(u,v)}^{(j)}$ are the weight matrix and bias, respectively. N (v) is the set of neighbors of node v, including V. The weight of edge (u, v) is $g_{u,v}^{(j)}$, which applies the gate to the edge to indicate the importance of the edge, as shown in equation 6:

Here, $V_{Re(u,v)}^{j}$ and $d_{Re(u,v)}^{j}$ are the gate weight matrix and bias, respectively. We used BioBERT embedding $A=\{a_1,a_2,...,a_n\}$ to initialize the input of the first GCN layer. Stacking k GCN layers can obtain a syntactic information matrix  , where m is the dimension of node v_i with the same dimension of a_i.

Multi-Head Attention

As shown in Figure 2, multi-head attention [40] comprises H self-attentions, which can thoroughly learn the similarity between nodes and calculate the importance of each node so that the model can focus on more critical node features. Let W_i^Q, W_i^K, and W_i^V be the ith initialized weight matrix of Q, K, and V, known by equation 7:

Here,  ,  ,  , and $d_k=d_v=m/H$.

We calculated the scoring matrix of the ith head according to equation 8. After concatenating H heads, we used equation 9 to obtain the attention output matrix M.  is the linear transformation matrix:

Joint Extraction Layer

Tagger

The tagger comprises a unidirectional LSTM that takes the context representation given by BiLSTM as the input and the syntactic dependency representation generated by the attention GCN module to parse the information of the previous layer. Let  . After the tagger module, we obtained the output matrix O, which was sent to the conditional probability extraction module.

Conditional Probability Extraction

Most joint extraction models input the same source information into different subtask classifiers simultaneously to achieve information sharing, as shown in equation 10, where  is the output of the trigger in time step i and  is the output of the argument in step j.

However, when the occurrence frequency of 2 subtasks in the same data set varies significantly, the model easily focuses on high-frequency subtasks and ignores low-frequency subtasks. Similar to the biomedical event extraction task, for the trigger recognition and event argument detection subtasks, each event trigger (ie, biomedical event) may contain 0, 1, or 2 participating elements, and the participating element may also be another event; therefore, the contribution of the trigger recognition task will be greater than that of the event argument detection task. To alleviate the abovementioned problems and reduce the cascading errors between these 2 subtasks, we combined the softmax output after trigger recognition and the source information to extract the trigger vector Tr_i and event argument vector Can_j according to the location of triggers and candidate arguments. Finally, by aggregating and inputting them into the event extraction classifier and learning the distribution features of the trigger label, our model directly achieved biomedical event extraction without postprocessing.

Here, W^{tri} and b^{tri} are the weight matrix and bias for trigger recognition, separately. The probability output of the trigger softmax of the kth word is $soft_k$. W^{event} and b^{event} are the weight matrix and bias for event extraction, separately. The number of

words of the *i*th trigger and the *j*th candidate argument are i_m and j_n, separately. O_k is the source information vector of the *k*th word.

Comparing equation 10 with equation 11, we found that it only realizes the joint extraction of triggers and event arguments using equation 10; therefore, it needs postprocessing to seek out the tuple of events. However, owing to the aggregation of trigger distribution information, we can discover which event argument belongs to the trigger of step *t* using equation 11.

Joint Dice Loss

Owing to the sparse data of the biomedical event corpus and the imbalance between positive and negative examples, the cross-entropy or negative log-likelihood loss function causes a large discrepancy between precision and recall. To alleviate this problem, we propose using a joint weight self-adjusting Dice loss function [41], as follows:

Here, *N* is the number of sentences in the corpus; n_p, t_p, and e_p are the number of tokens, extracted trigger candidates, and arguments of the *l*th sentence, λ is for smoothing purposes, β is a hyperparameter to adjust the loss, and θ is the model's parameters that should be trained.

Training

The CPJE model was trained using several epochs. In each epoch, we divided the training set into batches, each containing a list of sentences and each sentence containing a set of tokens of variable lengths. One batch was in progress at a time step.

For each batch, we first ran the information extraction layer to generate the context representation  and the attention representation with syntactic information  . Then, we combined *L* and *M* as the input of LSTM to generate source information *O*. In the end, we ran the joint extraction layer to compute gradients for overall network output (triggers and events). After that, we back propagated the errors from the output to the input through CPJE and updated all the network parameters. The overall procedure of the CPJE model is summarized in Textbox 2.

Textbox 2. The training procedure of the conditional probability joint extraction model.

Input

1. Sequence of tokens $\{w_1,...,w_n\}$ along with corresponding event labels
2. Set of edges $\{e_{12},...,e_{ij},...,e_{mn}\}$ for each corresponding token

Output

All parameters in the conditional probability joint extraction model

1. For each epoch do
2. For each epoch do
3. Generate *L* and *M* by information extraction layer via equations 3 and 9
4. Concatenate *L* and *M* as *T*
5. Generate the source information $O=\{o_1,...,o_n\}$ by long short-term memory
6. Compute the trigger scores y_t and the trigger softmax probability *soft* by the "SoftMax Trigger" block in the joint extraction layer via the first equation in equation 11
7. Fuse *O* and *soft* via the second and third equations in equation 11
8. Compute the event scores y_t. by the "SoftMax Event" block in the joint extraction layer via the fourth equation in equation 11
9. Update the parameters by the back propagation algorithm
10. End for
11. End for

Data

Our experiments were conducted mainly on the MLEE corpus [6], as shown in Table 2, which has 4 categories containing 19 predefined trigger subcategories. There are 262 documents with 56,588 words in total, with 8291 entities and 6677 events. From Table 2, we note that the number of anatomical-level events is higher than the number of molecular-level and planned-level events, although general biomedical events dominate overall. Overall, 18% (1202/6677) of the total events involved either direct or indirect arguments at both the molecular and anatomical levels. From Table 1, we find that the arguments of regulation, positive regulation, negative regulation, and planned process events may not be only entities but also other events; therefore, these events are nested events, which account for approximately 54.87% (3664/6677) of all events.

Table 2. The multilevel event extraction statistical information.

Item	Training, n (%)	Development, n (%)	Test, n (%)	Total, N
Document	131 (50)	44 (16.8)	87 (33.2)	262
Sentence	1271 (48.73)	457 (17.52)	880 (33.74)	2608
Word	27,875 (49.26)	9610 (16.98)	19,103 (33.76)	56,588
Entity	4147 (50.02)	1431 (17.26)	2713 (32.72)	8291
Event	3296 (49.36)	1175 (17.6)	2206 (33.04)	6677
Anatomical	810 (48.36)	269 (16.06)	596 (35.58)	1675
Molecular	340 (48.2)	125 (17.7)	240 (34.0)	705
General	1851 (50.66)	627 (17.16)	1176 (32.18)	3654
Planned	295 (45.9)	154 (24.0)	194 (30.2)	643

In addition, we verified our experiment using the BioNLP-ST 2011 GE corpus [7]. As shown in Table 3, the BioNLP-ST 2011 GE corpus defines 9 biomedical event types. It is noted that a *binding* event probably requires >1 protein entity as its theme argument, and a *regulation* event is likely to require a protein or an event as its theme argument and needs a protein or an event as its cause argument. There were 37.20% (9288/24,967) of events (regulation, positive regulation, and negative regulation) that led to a nested structure.

Table 3. The primary event types and core argument roles in the BioNLP-STa 2011 GEb corpus and the important statistical information of the GE corpus.

Event types and BioNLP-ST 2011 GE items	Core arguments	Values, N
Event type		
Gene expression	Theme (protein)	N/A[c]
Transcription	Theme (protein)	N/A
Protein catabolism	Theme (protein)	N/A
Phosphorylation	Theme (protein)	N/A
Localization	Theme (protein)	N/A
Binding	Theme (protein)[d]	N/A
Regulation	Theme (protein or event) and cause (protein or event)	N/A
Positive regulation	Theme (protein or event) and cause (protein or event)	N/A
Negative regulation	Theme (protein or event) and cause (protein or event)	N/A
BioNLP-ST 2011 GE corpus statistics		
Document	N/A	1224
Word	N/A	348,908
Entity	N/A	21,616
Event	N/A	24,967

[a]BioNLP-ST: BioNLP shared task.

[b]GE: Genia event.

[c]N/A: not applicable.

[d]Represents the number of arguments >1.

Hyperparameter Setting

For the hyperparameter settings of our experiment, we used 768 dimensions for the BioBERT embeddings and set 64 dimensions for the POS-tagging and entity label embeddings. We applied a 1-layer BiLSTM with 128 hidden units and used a 2-layer GCN and 2-head self-attention for our model. The dropout rate was 0.3, the learning rate was 0.01, and the optimization function was stochastic gradient descent (SGD). The training of our CPJE model was based on the operating system of Ubuntu 20.04, using PyTorch (version 1.9.0) and Python (version 3.8.8).

The graphics processing unit was an NVIDIA TITAN Xp with 12 GB of memory.

Results

Overall Performance on MLEE

We compare our performance with the baselines shown in Textbox 3.

Textbox 3. Baselines for performance.

EventMine

Pyysalo et al [6] applied a pipeline-based event extraction system, mainly relying on support vector machine classifiers to implement trigger recognition and event extraction.

Semisupervised learning

This is a semisupervised learning framework proposed by Zhou et al [30], which can use unannotated data to extract biomedical events.

Convolutional neural network

Wang et al [3] used convolutional neural networks and multiple distributed feature vector representations to achieve event extraction tasks.

mdBLSTM (bidirectional long short-term memory with a multilevel attention mechanism and dependency-based word embeddings)

He et al [5] proposed a bidirectional long short-term memory neural network based on a multilevel attention mechanism and dependency-based word embeddings to extract biomedical events.

Reinforcement learning+knowledge bases

Zhao et al [32] proposed a framework of reinforcement learning with external biomedical knowledge bases for extracting biomedical events.

DeepEventMine

Trieu et al [35] proposed an end-to-end neural model. It uses a multioverlapping directed acyclic graph to detect nested biomedical entities, triggers, roles, and events.

Hierarchical artificial neural network

Zhao et al [36] proposed a 2-level modeling method for document-level joint biomedical event extraction.

Table 4 illustrates the overall performance against the state-of-the-art methods with gold standard entities. As seen in this table, our CPJE model achieved only a slight improvement in the trigger recognition task. For the event extraction task, the F_1 score was significantly better than the other baselines. Notably, the gap between the precision and recall of our model was much smaller than that of the mdBLSTM (bidirectional long short-term memory with a multilevel attention mechanism and dependency-based word embeddings) model, and the precision was much better than that of the RL+KBs model. This indicates that our model had a better effect on reducing cascading errors than the pipeline models. In addition, the hierarchical artificial neural network (HANN) model was also a joint extraction model; however, its performance is disappointing. This is because the HANN model focuses on extracting document-level biomedical events, which contain many cross-sentence entities, triggers, and events. However, other models aim to extract sentence-level events; therefore, the performance of these models is better than that of the HANN model.

Table 4. Overall performance on multilevel event extraction compared with the state-of-the-art methods with gold standard entities.

Method	Trigger recognition (%)			Event extraction (%)		
	Precision	Recall	F_1 score	Precision	Recall	F_1 score
EventMine[a]	70.79	81.69	75.84	62.28	49.56	55.20
SSL[a,b]	72.17	82.26	76.89	55.76	59.16	57.41
CNN[a,c]	80.92	75.23	77.97	60.56	56.23	58.31
mdBLSTM[a,d]	82.79	76.56	79.55	90.24	44.50	59.61
RL[e]+KBs[a,f]	N/A[g]	N/A	N/A	63.78	56.81	60.09
DeepEventMine[h]	N/A	N/A	N/A	69.91	55.49	61.87
HANN[h,i]	N/A	N/A	N/A	63.91	56.08	59.74
Our model[h]	82.20	78.25	80.18	72.26	55.23	62.80[j]

[a]Pipeline model.

[b]SSL: semisupervised learning.

[c]CNN: convolutional neural network.

[d]mdBLSTM: bidirectional long short-term memory with a multilevel attention mechanism and dependency-based word embeddings

[e]RL: reinforcement learning.

[f]KB: knowledge base

[g]N/A: not applicable.

[h]Joint model.

[i]HANN: hierarchical artificial neural network.

[j]The best value compared with baselines.

The Performance for Nested Events on MLEE

To evaluate the effectiveness of our model for improving the nested biomedical event extraction, we split the test set into 2 parts (*simple* and *nested*). *Simple* means that 1 event only regards the entities as its arguments; *nested* means that one of the arguments of an event may be another event. In general, nested events are present in regulation, positive regulation, negative regulation, and planned process events.

Table 5 illustrates the performance (F_1 scores) of the CNN model [3], the RL+KBs model [32], the DeepEventMine [35] model, the HANN [36] model, and our model in the trigger recognition and event extraction subtasks. In the *simple* and *nested* data of triggers, our framework was 0.44% and 1.25% better than the CNN model, which demonstrates that our model can improve the performance of trigger recognition. However, there is no significant difference between simple and nested triggers. In the *nested* data of events, our model was 6.97% higher than the CNN model, 2.57% higher than the RL+KBs model, 9.53% higher than the DeepEventMine model, and 15.8% higher than the HANN model, which illustrates that our CPJE model of using a gate GCN and an attention mechanism helps to enhance the performance of extracting nested events.

Table 5. The F1 score performance on simple events, nested events, and all events on the multilevel event extraction corpus.

Subtask and model	Simple (%)	Nested (%)	All (%)
Trigger			
CNN[a]	79.52	78.80	78.52
RL[b]+KBs[c]	N/A[d]	N/A	N/A
DeepEventMine	N/A	79.12	N/A
HANN[e]	N/A	N/A	N/A
Our model	79.96[f]	80.05[f]	80.18[f]
Event			
CNN	61.33	54.29	58.87
RL+KBs	N/A	58.69	60.09
DeepEventMine	N/A	51.73	61.87
HANN	77.08[f]	45.46	59.74
Our model	64.85	61.26[f]	62.80[f]

[a]CNN: convolutional neural network.

[b]RL: reinforcement learning.

[c]KB: knowledge base.

[d]N/A: not applicable.

[e]HANN: hierarchical artificial neural network.

[f]The best value compared with other models.

The Performance for All Events on MLEE

To illustrate the impact of our framework on different events in more detail, Table 6 presents the event extraction performance for all event types. From this table, we obtain the best extraction performance for dephosphorylation events and the worst performance for transcription events. In addition, the catabolic events had the best extraction precision, and the phosphorylation events had the best extraction recall rate.

Table 6. The extraction performance for different events on multilevel event extraction corpus.

Events	Precision (%)	Recall (%)	F_1 score (%)
Cell proliferation	62.50	58.57	60.47
Development	51.82	66.43	58.22
Blood vessel development	90.42	72.66	80.57
Growth	78.02	50.58	61.37
Death	79.12	44.32	56.81
Breakdown	71.30	48.30	57.59
Remodeling	85.71	58.32	69.41
Synthesis	48.00	20.30	28.53
Gene expression	74.72	82.42	78.38
Transcription	16.67	33.33	22.22
Catabolism	100.00	50.00	66.67
Phosphorylation	90.00	100.00	94.74
Dephosphorylation	100.00	100.00	100.00
Localization	76.86	49.98	60.57
Binding	74.52	51.23	60.71
Regulation	63.82	51.49	56.99
Positive regulation	78.28	50.66	61.51
Negative regulation	64.35	54.69	59.13
Planned process	69.57	51.86	59.42
All	64.85	61.26	62.80

Overall Performance on BioNLP-ST 2011 GE

To improve persuasion, we extended our experiment to the BioNLP-ST 2011 GE corpus. We compared our event extraction results with those of previous systems using the same corpus, as shown in Table 7. Among them, the Turku Event Extraction System (TEES) [42], EventMine [6], and stacked generalization [25] systems are based on support vector machines with designed features. The TEES-CNNs [4] are CNNs integrated into the TEES system to extract relations and events. The DeepEventMine [35] is based on bidirectional transformers and an overlapping directed acyclic graph to jointly extract biomedical events. The HANN [36] model relies on the GCN and hypergraph to obtain local and global contexts. The KB-driven tree LSTM [33] depends on KB concept embedding to improve the pretrained distributed word representations. The Graph Edge-conditioned Attention Networks with Science BERT (GEANet-SciBERT) [34] adopts a hierarchical graph representation encoded by graph edge-conditioned attention networks to incorporate domain knowledge from the Unified Medical Language System into a pretrained language model. Table 7 illustrates that except for the DeepEventMine, our approach outperformed all previous methods.

Table 7. The performance of biomedical event extraction on the BioNLP shared task 2011 Genia event corpus.

Method and event type	Precision (%)	Recall (%)	F_1 score (%)
TEES[a,b]			
Event total[c]	57.65	49.56	53.30
EventMine[a]			
Event total	63.48	53.35	57.98
Stacked generalization[a]			
Event total	66.46	48.96	56.38
TEES-CNNs[a,d]			
Event total	69.45	49.94	58.07
HANN[e,f]			
Event total	71.73	53.21	61.10
KB[g]**-driven tree LSTM**[e,h]			
Simple total[i]	85.95	72.62	78.73
Binding	53.16	37.68	44.10
Regulation total[j]	55.73	41.73	47.72
Event total	67.10	52.14	58.65
GEANet-SciBERT[e,k]			
Regulation total	55.21	47.23	50.91
Event total	64.61	56.11	60.06
DeepEventMine[e]			
Regulation total	62.36	51.88	56.64[l]
Event total	76.28	55.06	63.96[l]
Our model[e]			
Simple total	82.23	78.88	80.52
Binding	55.12	37.48	44.62
Regulation total	57.82	46.39	51.48
Event total	72.62	53.33	61.50

[a]Pipeline model.

[b]TEES: Turku Event Extraction System.

[c]Represents the overall performance on the test set.

[d]CNN: convolutional neural network.

[e]Joint model.

[f]HANN: hierarchical artificial neural network.

[g]KB: knowledge base.

[h]LSTM: long short-term memory.

[i]Represents the overall performance for simple events on the test set.

[j]Represents the overall performance for nested events on the test set (including regulation, positive regulation, and negative regulation subevents).

[k]GEANet-SciBERT: Graph Edge-conditioned Attention Networks with Science BERT.

[l]The best value compared with other models.

The KB-driven tree LSTM and GEANet-SciBERT both draw on the KB to enhance the semantic representation of words to improve the extraction performance of nested (regulation) events. However, the KB-driven tree LSTM only leverages traditional static word embedding, which cannot deeply integrate information from the KB; thus, its performance on nested events is unsatisfactory.

Unlike the KB-driven tree LSTM method, the GEANet-SciBERT model uses a specialized medical KB and scientific information to enrich the dynamic semantic representation of Bidirectional Encoder Representation from Transformers (BERT) and enhances the capability of inferring nested events via a novel GNN. Thus, the F_1 scores for the nested event extraction were significantly boosted.

Interestingly, the DeepEventMine had an outstanding performance for extracting nested biomedical events on BioNLP-ST 2011 GE but had a passive performance on MLEE. There are three reasons for this fact. First, the DeepEventMine model jointly learns 4 biomedical information tasks (entity detection, trigger detection, role detection, and event detection), which can share more biomedical features and knowledge when model training. Second, the DeepEventMine model uses a more complex graph structure (multiple overlapping directed acyclic graphs) to obtain rich syntactic information. (Finally, the BioNLP-ST 2011 GE data set size is larger than that of the MLEE data set; thus, the DeepEventMine model can be fully trained on a large corpus and enhance the performance of extracting nested events.

Discussion

In this section, we will study and discuss the performance of our CPJE model using the MLEE corpus.

Ablation Study

The Impact of the BiLSTM

Although the output of BioBERT contains rich semantic information, it has some noise impact on semantic information after concatenating POS embedding, entity embedding, and BioBERT embedding. In addition, the dimension of the BioBERT output is 768, and the total size after concatenation is more extensive, which tends to cause the phenomenon of combination explosion in the feature space. Therefore, we considered using a BiLSTM, which reduces the total dimension and integrates other information with the BioBERT information to obtain a richer semantic representation.

If we remove the BiLSTM layer, the trigger recognition precision is dropped from 82.20% to 75.64%, and the trigger recognition F_1 score is dropped from 80.18% to 76.39%, which further affects the event extraction performance (the event extraction F_1 score is fell from 62.80% to 58.02%).

The Impact of Softmax Probability

To evaluate the contribution of the softmax probability distribution after trigger prediction to the event extraction task, we used the traditional joint extraction method (as shown in equation 10), which only uses source information when extracting candidate trigger vectors and event argument vectors.

If we only use the source information (soft trigger) for joint extraction, the event extraction task lacks the probability distribution information after trigger recognition, which results in a decline in the recall rate of the model and further affects the F_1 scores (the event extraction F_1 score is dropped from 62.80% to 60.09%). However, the overall result is still slightly higher than the pipeline baseline, which also reflects that joint extraction can eliminate cascading errors.

The Impact of GCN

We removed the syntactic structure to evaluate the importance of the GCN network; therefore, the GCN module was useless in our model. If the model lacks the GCN component, the performance of trigger recognition is slightly degraded (the trigger recognition F_1 score is fell from 80.18% to 78.78%), and the result of event extraction is significantly worse than that of the proposed model (the event extraction F_1 score is fell from 62.80% to 58.40%).

As the syntactic structure can provide significant potential information for event extraction, the GCN model can be aware of the direction of information flow in syntactic structures and capture these features effectively. Therefore, the GCN model is vital for event extraction.

The Impact of Dice Loss

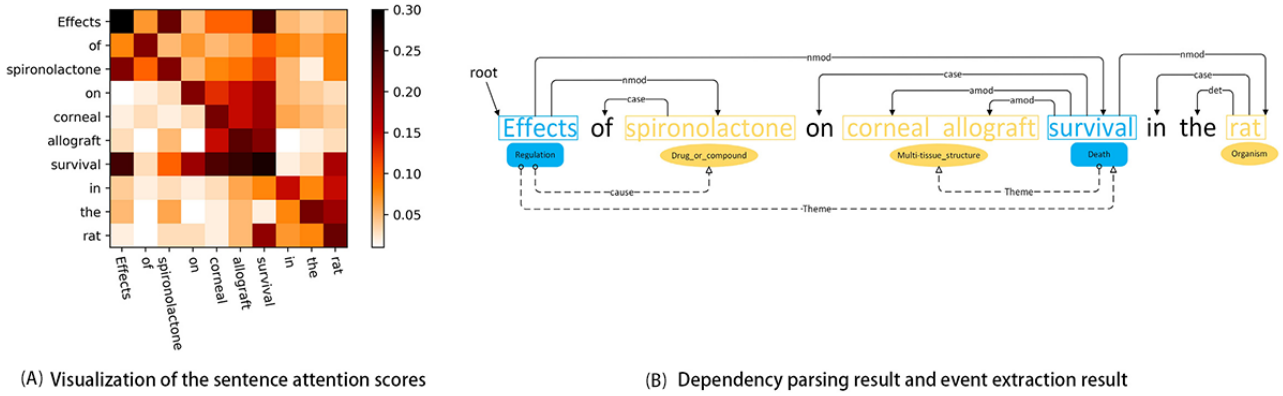
In the face of an imbalance in biomedical corpora, we used the Dice loss function. To verify that the Dice loss function had a better effect on event extraction, we used the cross-entropy loss function for comparison.

A significantly large number of negative examples in the data set indicates that easy-negative examples are extensive. A large number of straightforward examples overwhelmed the training, making the model insufficient to distinguish between positive and hard-negative examples. As the cross-entropy loss is accuracy oriented and each instance contributes equally to the loss function, the precision of the model increases (the event extraction precision is risen from 72.26% to 89.26%), but the F_1 scores do not increase (the event extraction F_1 score is dropped from 62.60% to 60.30%). Dice loss is a muted version of the F_1 score—the harmonic mean of precision and recall. When the positive and negative examples in the data set are unbalanced, the Dice loss will reduce the focus on the easy-negative sample and increase the attention on positive and hard-negative samples, thereby balancing the precision and recall values and increasing the F_1 scores.

Visualization

For the effectiveness of the attention-based gate GCN, we used the sentence “Effects of spironolactone on corneal allograft survival in the rat” in Figure 3 as an example to illustrate the captured interaction features. From Figure 3B, we know this sentence contains 2 events: a *regulation* event caused by *effects* and a *death* event caused by *survival*. In addition, a death event is one of the arguments for the regulation event.

Figure 3. An example of attention-based gate graph neural network effectiveness. (A) Row-wise heap map, where each row is an array of average scores of the 2 heads obtained from the multi-head attention mechanism. The darker the color, the higher the score and the stronger the interaction. (B) Dependency parsing result produced by Stanford CoreNLP and the golden relationships between event triggers and arguments, where yellow boxes represent entity type, and the blue boxes represent event type.



As we can see in Figure 3A, the *effects* row has moderately strong links with *Effects* (self), spironolactone (its argument), and *survival* (its argument and another event). Meanwhile, the *survival* row has strong links with *survival* (self), *effects* (another event), and *corneal allograft* (its argument). In addition, the words *rat* and *on* also have strong connections with *survival*, which means that the syntactic dependency information generated by parsing is propagated through the GCN.

Case Study

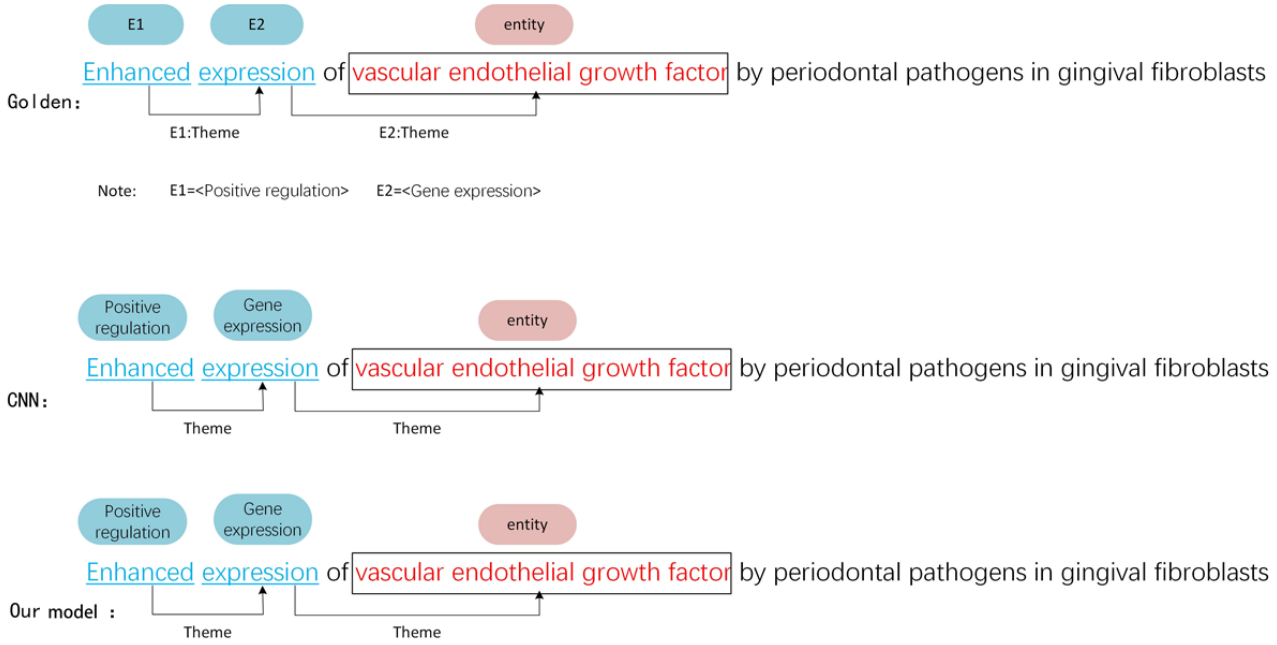
Overview

Our framework has not achieved state-of-the-art results for the BioNLP-ST 2011 GE corpus. However, the performance of extracting nested biomedical events is satisfactory, particularly in the MLEE corpus. To more intuitively demonstrate the performance of our model in extracting nested biomedical events, we analyzed 3 examples of nested events selected from the MLEE test set to study the strengths and weaknesses of our model compared with the CNN [3].

Case 1

As shown in Figure 4, case 1 is a simple nested event, where the role type of event argument is only the *theme*. It is a nested event; however, both the CNN and our model obtained correct event extraction results. This is because this sentence does not have a complete component, and perhaps, it is only a part of a complete sentence. The simpler the sentence structure is, the easier it is for the model to extract practical features. Therefore, the extraction performance for such nested events is generally favorable.

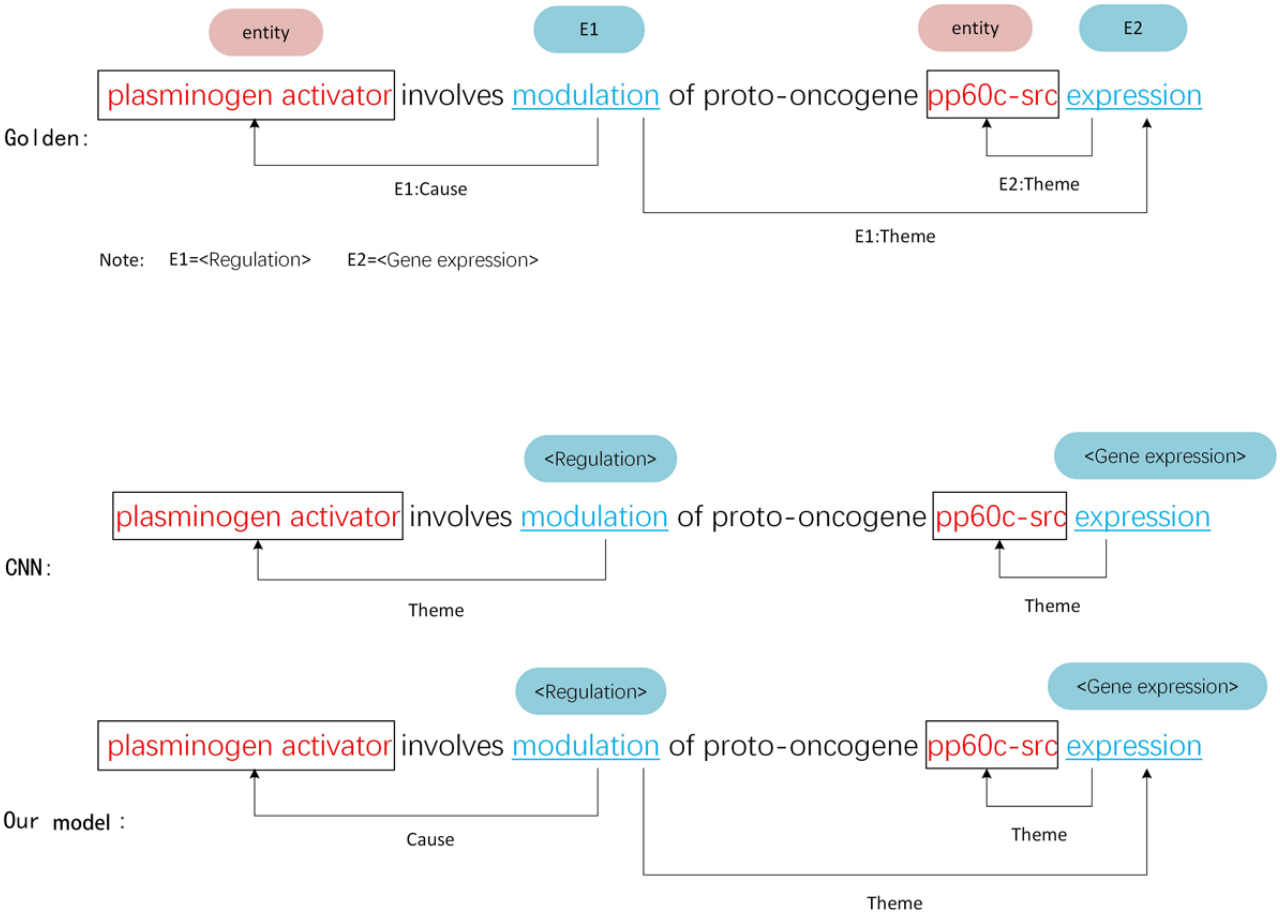
Figure 4. Case study for a simple nested event on the multilevel event extraction corpus. CNN: convolutional neural network.



Case 2

Case 2 is a general nested event whose sentence component is complete, and the role types of event arguments are *theme* and *cause*. As shown in Figure 5, the CNN model detects all correct event triggers but cannot detect the correct event arguments. The CNN model is a pipeline approach that considers trigger recognition and argument detection tasks in a cascade rather than a parallel relationship. In general, they first input the text into the CNN model to identify the triggers in the sentence. Then, they construct <trigger, entity> or <trigger, trigger> candidate pairs and input them into the CNN model again to detect the arguments. Finally, rule-based or machine learning-based methods are used to postprocess triggers and arguments to construct complete biomedical events. If there is an error in some of these steps, it will directly affect the performance of event extraction. However, our joint method regards trigger recognition and argument detection as parallel tasks that can provide valid information. Thus, we trained both tasks jointly with one model, and errors could only be generated during the model training.

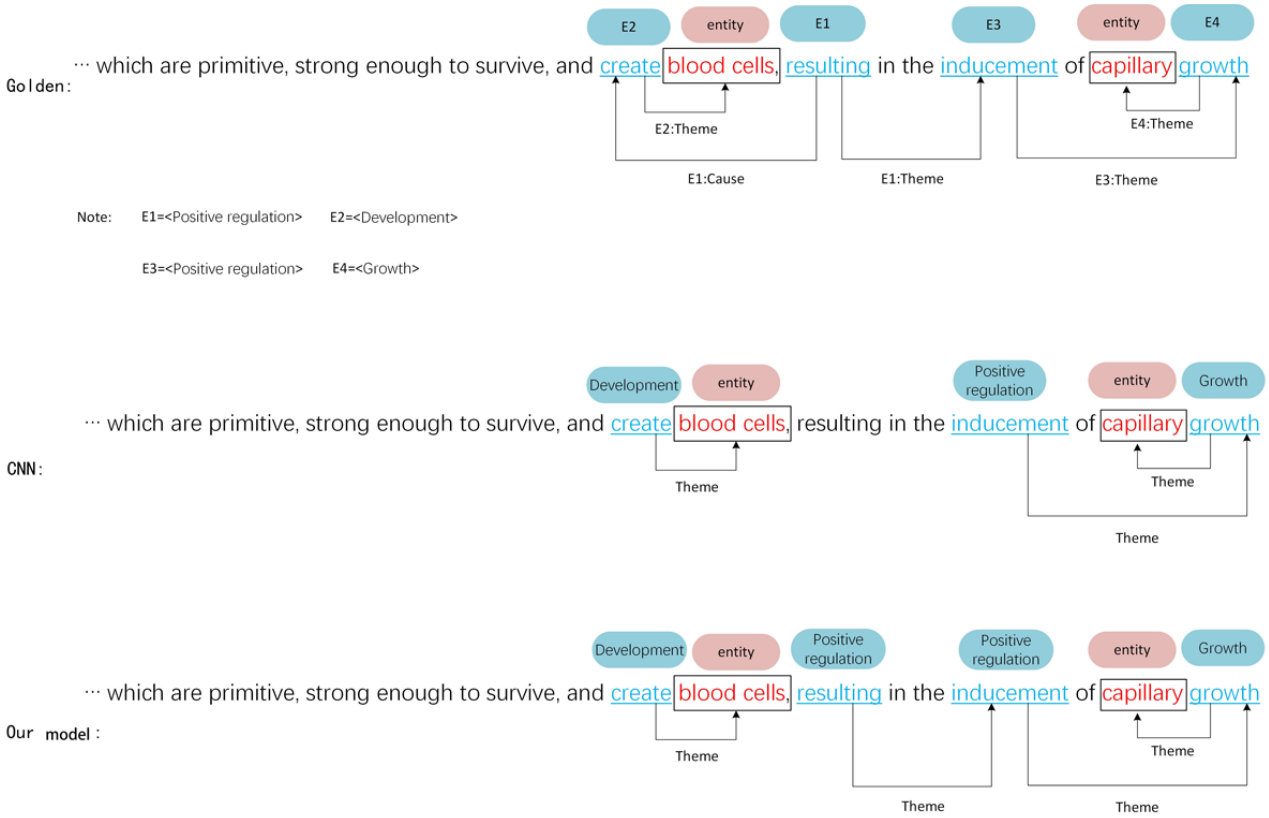
Figure 5. Case study for a common nested event on multilevel event extraction corpus. CNN: convolutional neural network.



Case 3

Case 3 is a cross-sentence nested event, as shown in Figure 6. From this example, we can determine what needs to be improved. As multiple events are nested in each other, and some of these events are not in the same sentence, this prevents the model from extracting all events efficiently and accurately. Compared with the CNN model, although our model can identify the *positive regulation* event triggered by *resulting*, it is not in the same clause as the *development* event triggered by *create*, which causes the *positive regulation* event to lack an event argument.

Figure 6. Case study for an across-sentence nested event on multilevel event extraction corpus. CNN: convolutional neural network.



Conclusions

In this study, a CPJE framework based on a multi-head attention graph CNN is proposed to achieve biomedical event extraction tasks. The cascading errors between the 2 subtasks were reduced because of the use of the joint extraction framework. With the help of the attention-based gate GCN, syntactic dependency information and the interrelations between triggers and related entities were effectively learned; thus, the extraction performance of nested biomedical events improved. The Dice loss replaced the cross-entropy loss, which weakened the negative impact of the imbalanced data set. Overall, the model obtained the best F_1 score in the MLEE biomedical event extraction corpus and achieved favorable performance on the BioNLP-ST 2011 GE corpus. In the future, we will consider integrating external resource knowledge to allow the model to learn richer information and improve the performance of cross-sentence nested events.

Acknowledgments

This study was funded by grants from the National Natural Science Foundation of China (number 62072070).

Authors' Contributions

YW proposed the study of biomedical event extraction, implemented and verified the effectiveness of the joint extraction framework, and wrote the first draft. JW put forward constructive suggestions for revising this draft. H Lu read the final manuscript and provided some useful suggestions. H Lin read and approved the final manuscript. BX read and approved the final manuscript. YZ helped to review and revise the draft. SKB helped revise the draft.

Conflicts of Interest

None declared.

References

1. McDonald RT, Pereira FC, Kulick SN, Winters R, Jin Y, White PS. Simple algorithms for complex relation extraction with applications to biomedical IE. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. 2005 Presented at: ACL '05; June 25-30, 2005; Ann Arbor, MI, USA p. 491-498. [doi: 10.3115/1219840.1219901]
2. Kilicoglu H, Bergler S. Effective bio-event extraction using trigger words and syntactic dependencies. Comput Intell 2011 Nov 27;27(4):583-609. [doi: 10.1111/j.1467-8640.2011.00401.x]

3. Wang A, Wang J, Lin H, Zhang J, Yang Z, Xu K. A multiple distributed representation method based on neural network for biomedical event extraction. BMC Med Inform Decis Mak 2017 Dec 20;17(Suppl 3):171 [FREE Full text] [doi: 10.1186/s12911-017-0563-9] [Medline: 29297321]
4. Björne J, Salakoski T. Biomedical event extraction using convolutional neural networks and dependency parsing. In: Proceedings of the BioNLP 2018 workshop. 2018 Presented at: BioNLP '18; July 19, 2018; Melbourne, Australia p. 98-108. [doi: 10.18653/v1/w18-2311]
5. He X, Li L, Song X, Huang D, Ren F. Multi-level attention based BLSTM neural network for biomedical event extraction. IEICE Trans Inf Syst 2019;E102.D(9):1842-1850. [doi: 10.1587/transinf.2018edp7268]
6. Pyysalo S, Ohta T, Miwa M, Cho H, Tsujii J, Ananiadou S. Event extraction across multiple levels of biological organization. Bioinformatics 2012 Sep 15;28(18):i575-i581 [FREE Full text] [doi: 10.1093/bioinformatics/bts407] [Medline: 22962484]
7. Kim JD, Wang Y, Takagi T, Yonezawa A. Overview of Genia event task in BioNLP shared task 2011. In: Proceedings of BioNLP Shared Task 2011 Workshop. 2011 Presented at: BioNLP '11; June 24, 2011; Portland, OR, USA p. 7-15. [doi: 10.1186/1471-2105-13-s11-s1]
8. Zheng S, Hao Y, Lu D, Bao H, Xu J, Hao H, et al. Joint entity and relation extraction based on a hybrid neural network. Neurocomputing 2017 Sep 27;257:59-66. [doi: 10.1016/j.neucom.2016.12.075]
9. Ye ZX, Ling ZH. Distant supervision relation extraction with intra-bag and inter-bag attentions. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL '19; June 2-7, 2019; Minneapolis, MN, USA p. 2810-2819. [doi: 10.48550/arXiv.1904.00143]
10. Feng J, Huang M, Zhao L, Yang Y, Zhu X. Reinforcement learning for relation classification from noisy data. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018 Feb Presented at: AAAI '18; Feb 2-7, 2018; New Orleans, LA, USA.
11. Fu TJ, Li PH, Ma WY. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: ACL '19; July 28-August 2, 2019; Florence, Italy p. 1409-1418. [doi: 10.18653/v1/p19-1136]
12. Guo Z, Zhang Y, Lu W. Attention guided graph convolutional networks for relation extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: ACL '19; July 28-August 2, 2019; Florence, Italy p. 241-251. [doi: 10.18653/v1/p19-1024]
13. Li Q, Ji H, Huang L. Joint event extraction via structured prediction with global features. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013 Aug Presented at: ACL '13; August 4-9, 2013; Sofia, Bulgaria p. 73-82.
14. Keith KA, Handler A, Pinkham M, Magliozzi C, McDuffie J, O'Connor B. Identifying civilians killed by police with distantly supervised entity-event extraction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017 Sep Presented at: EMNLP '17; September 7-8, 2017; Copenhagen, Denmark p. 1547-1557. [doi: 10.18653/v1/d17-1163]
15. Reichart R, Barzilay R. Multi-event extraction guided by global constraints. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012 Jun Presented at: NAACL '12; June 3-8, 2012; Montreal, Canada p. 70-79.
16. Lu W, Roth D. Automatic event extraction with structured preference modeling. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. 2012 Jul Presented at: ACL '12; July 8-14, 2012; Jeju Island, Korea p. 835-844.
17. Sha L, Qian F, Chang B, Sui Z. Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018 Presented at: AAAI '18; February 2-7, 2018; New Orleans, LA, USA.
18. Nguyen TH, Grishman R. Modeling skip-grams for event detection with convolutional neural networks. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016 Presented at: EMNLP '16; November 1-5, 2016; Austin, TX, USA p. 886-891. [doi: 10.18653/v1/d16-1085]
19. Liu X, Luo Z, Huang H. Jointly multiple events extraction via attention-based graph information aggregation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 Presented at: EMNLP '18; October 31-November 4, 2018; Brussels, Belgium p. 1247-1256. [doi: 10.18653/v1/d18-1156]
20. Kim JD, Ohta T, Pyysalo S, Kano Y, Tsujii J. Overview of BioNLP'09 shared task on event extraction. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. 2009 Presented at: BioNLP '09; June 05, 2009; Boulder, CO, USA p. 1-9. [doi: 10.3115/1572340.1572342]
21. Bossy R, Golik W, Ratkovic Z, Bessières P, Nédellec C. Bionlp shared task 2013 - an overview of the bacteria biotope task. In: Proceedings of the BioNLP Shared Task 2013 Workshop. 2013 Presented at: BioNLP '13; August 09, 2013; Sofia, Bulgaria p. 161-169. [doi: 10.18653/v1/W16-3002]
22. Miwa M, Saetre R, Kim JD, Tsujii J. Event extraction with complex event classification using rich features. J Bioinform Comput Biol 2010 Feb;8(1):131-146. [doi: 10.1142/s0219720010004586] [Medline: 20183879]

23. Miwa M, Thompson P, Ananiadou S. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. Bioinformatics 2012 Jul 01;28(13):1759-1765 [FREE Full text] [doi: 10.1093/bioinformatics/bts237] [Medline: 22539668]
24. Björne J, Salakoski T. TEES 2.1: automated annotation scheme learning in the BioNLP 2013 shared task. In: Proceedings of the BioNLP Shared Task 2013 Workshop. 2013 Presented at: BioNLP '13; August 9, 2013; Sofia, Bulgaria p. 16-25. [doi: 10.18653/v1/w16-3009]
25. Majumder A, Ekbal A, Naskar SK. Biomolecular event extraction using a stacked generalization-based classifier. In: Proceedings of the 13th International Conference on Natural Language Processing. 2016 Presented at: ICNLP '16; December 17-20, 2016; Varanasi, India p. 55-64.
26. Riedel S, McCallum A. Robust biomedical event extraction with dual decomposition and minimal domain adaptation. In: Proceedings of BioNLP Shared Task 2011 Workshop. 2011 Presented at: BioNLP '11; June 24, 2011; Portland, OR, USA p. 46-50.
27. Venugopal D, Chen C, Gogate V, Ng V. Relieving the Computational Bottleneck: joint inference for event extraction with high-dimensional features. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014 Presented at: EMNLP '14; October 25-29, 2014; Doha, Qatar p. 831-843. [doi: 10.3115/v1/d14-1090]
28. Nguyen DQ, Verspoor K. From POS tagging to dependency parsing for biomedical event extraction. BMC Bioinformatics 2019 Feb 12;20(1):72 [FREE Full text] [doi: 10.1186/s12859-019-2604-0] [Medline: 30755172]
29. Zhou D, Zhong D. A semi-supervised learning framework for biomedical event extraction based on hidden topics. Artif Intell Med 2015 May;64(1):51-58. [doi: 10.1016/j.artmed.2015.03.004] [Medline: 25863986]
30. Rao S, Marcu D, Knight K, Daumé III H. Biomedical event extraction using abstract meaning representation. In: Proceedings of the BioNLP 2017 workshop. 2017 Presented at: BioNLP '17; August 04, 2017; Vancouver, Canada p. 126-135. [doi: 10.18653/v1/w17-2315]
31. Yan S, Wong KC. Context awareness and embedding for biomedical event extraction. Bioinformatics 2020 Jan 15;36(2):637-643. [doi: 10.1093/bioinformatics/btz607] [Medline: 31392318]
32. Zhao W, Zhao Y, Jiang X, He T, Liu F, Li N. A novel method for multiple biomedical events extraction with reinforcement learning and knowledge bases. In: Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine. 2020 Presented at: BIBM '20; December 16-19, 2020; Seoul, South Korea p. 402-407. [doi: 10.1109/bibm49941.2020.9313214]
33. Li D, Huang L, Ji H, Han J. Biomedical event extraction based on knowledge-driven tree-LSTM. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL '19; June 2-7, 2019; Minneapolis, MN, USA p. 1421-1430. [doi: 10.18653/v1/N19-1145]
34. Huang KH, Yang M, Peng N. Biomedical event extraction with hierarchical knowledge graphs. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020 Presented at: EMNLP '20; November 16-20, 2020; Virtual p. 1277-1285. [doi: 10.18653/v1/2020.findings-emnlp.114]
35. Trieu HL, Tran TT, Duong KN, Nguyen A, Miwa M, Ananiadou S. DeepEventMine: end-to-end neural nested event extraction from biomedical texts. Bioinformatics 2020 Dec 08;36(19):4910-4917 [FREE Full text] [doi: 10.1093/bioinformatics/btaa540] [Medline: 33141147]
36. Zhao W, Zhang J, Yang J, He T, Ma H, Li Z. A novel joint biomedical event extraction framework via two-level modeling of documents. Inf Sci 2021 Mar;550:27-40. [doi: 10.1016/j.ins.2020.10.047]
37. Ramponi A, van der Goot R, Lombardo R, Plank B. Biomedical event extraction as sequence labeling. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020 Presented at: EMNLP '20; November 16-20, 2020; Virtual p. 5357-5367. [doi: 10.18653/v1/2020.emnlp-main.431]
38. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
39. Klein D, Manning CD. Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. 2003 Presented at: ACL '03; July 7-12, 2003; Sapporo, Japan p. 423-430. [doi: 10.3115/1075096.1075150]
40. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of Annual Conference on Advances in Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA, USA.
41. Li X, Sun X, Meng Y, Liang J, Wu F, Li J. Dice loss for data-imbalanced NLP tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020 Presented at: ACL '20; July 5-10, 2020; Virtual p. 465-476. [doi: 10.18653/v1/2020.acl-main.45]
42. Björne J, Salakoski T. Generalizing biomedical event extraction. In: Proceedings of BioNLP Shared Task 2011 Workshop. 2011 Presented at: BioNLP '11; June 24, 2011; Portland, OR, USA p. 183-191.

Abbreviations

BERT: Bidirectional Encoder Representation From Transformers
BiLSTM: bidirectional long short-term memory
BioBERT: Biomedical Bidirectional Encoder Representation From Transformers
BioNLP: biomedical natural language processing
BioNLP-ST: biomedical natural language processing shared task
CNN: convolutional neural network
CPJE: conditional probability joint extraction
GCN: graph convolutional network
GE: Genia event
GEANet-SciBERT: Graph Edge-conditioned Attention Networks with Science BERT
GNN: graph neural network
HANN: hierarchical artificial neural network
KB: knowledge base
LSTM: long short-term memory
mdBLSTM: bidirectional long short-term memory with a multilevel attention mechanism and dependency-based word embeddings
MLEE: multilevel event extraction
POS: parts of speech
RL: reinforcement learning
SGD: stochastic gradient descent
TEES: Turku Event Extraction System

Edited by T Hao; submitted 08.03.22; peer-reviewed by T Zhang, Y An; comments to author 06.04.22; revised version received 15.04.22; accepted 19.04.22; published 07.06.22.

Please cite as:
Wang Y, Wang J, Lu H, Xu B, Zhang Y, Banbhrani SK, Lin H
Conditional Probability Joint Extraction of Nested Biomedical Events: Design of a Unified Extraction Framework Based on Neural Networks
JMIR Med Inform 2022;10(6):e37804
URL: https://medinform.jmir.org/2022/6/e37804
doi:10.2196/37804
PMID:35671070

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org