JMIR Medical Informatics

Impact Factor (2020): 2.96
Volume 9 (2021), Issue 8 ISSN: 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Viewpoints

Potential Uses of Blockchain Technology for Outcomes Research on Opioids (e16293)
Aldren Gonzales, Scott Smith, Prashila Dullabh, Lauren Hovey, Krysta Heaney-Huls, Meagan Robichaud, Roger Boodoo. 4

A System to Support Diverse Social Program Management (e23219)
Mollie McKillop, Jane Snowdon, Van Willis, Shira Alevy, Rubina Rizvi, Karen Rewalt, Charlyne Lefebvre-Paillé, William Kassler, Gretchen Purcell Jackson. 19

Original Papers

Using Electronic Medical Record Data for Research in a Healthcare Information and Management Systems Society (HIMSS) Analytics Electronic Medical Record Adoption Model (EMRAM) Stage 7 Hospital in Beijing: Cross-sectional Study (e24405)
Rui Li, Yue Niu, Sarah Scott, Chu Zhou, Lan Lan, Zhigang Liang, Jia Li. 29

Predicting Patients' Intention to Use a Personal Health Record Using an Adapted Unified Theory of Acceptance and Use of Technology Model: Secondary Data Analysis (e30214)
Consuela Yousef, Teresa Salgado, Ali Farooq, Keisha Burnett, Laura McClelland, Abin Thomas, Ahmed Alenazi, Laila Abu Esba, Aeshah AlAzmi, Abrar Alhameed, Ahmed Hattan, Sumaya Elgadi, Saleh Almekhloof, Mohammed AlShammary, Nazzal Alanezi, Hani Alhamdan, Sahal Khoshhal, Jonathan DeShazo. 39

Usage Patterns of Web-Based Stroke Calculators in Clinical Decision Support: Retrospective Analysis (e28266)
Benjamin Kummer, Lubaina Shakir, Rachel Kwon, Joseph Habboushe, Nathalie Jetté. 54

Quality of Hospital Electronic Health Record (EHR) Data Based on the International Consortium for Health Outcomes Measurement (ICHOM) in Heart Failure: Pilot Data Quality Assessment Study (e27842)
Hannelore Aerts, Dipak Kalra, Carlos Sáez, Juan Ramírez-Anguita, Miguel-Angel Mayer, Juan Garcia-Gomez, Marta Durà-Hernández, Geert Thienpont, Pascal Coorevits. 64

Communicating the Implementation of Open Notes to Health Care Professionals: Mixed Methods Study (e22391)
Karin Jonnergård, Lena Petersson, Gudbjörg Erlingsdóttir. 78

Classification of Electronic Health Record–Related Patient Safety Incidents: Development and Validation Study (e30470)
Sari Palojoki, Kaija Saranto, Elina Reponen, Noora Skants, Anne Vakkuri, Riikka Vuokko. 92

XSL•FO
RenderX

Gender Prediction for a Multiethnic Population via Deep Learning Across Different Retinal Fundus Photograph Fields: Retrospective Cross-sectional Study (e25165)
Bjorn Betzler, Henrik Yang, Sahil Thakur, Marco Yu, Ten Quek, Zhi Soh, Geunyoung Lee, Yih-Chung Tham, Tien Wong, Tyler Rim, Ching-Yu Cheng. 101

A Worker-Centered Personal Health Record App for Workplace Health Promotion Using National Health Care Data Sets: Design and Development Study (e29184)
Hyun Park, Kwang Kim, Ho-Young Chung, Sungmoon Jeong, Jae Soh, Young Hyun, Hwa Kim. 113

Team Dynamics in Hospital Workflows: An Exploratory Study of a Smartphone Task Manager (e28245)
Danula Hettiachchi, Lachie Hayes, Jorge Goncalves, Vassilis Kostakos. 153

Foodborne Disease Risk Prediction Using Multigraph Structural Long Short-term Memory Networks: Algorithm Design and Validation Study (e29433)
Yi Du, Hanxue Wang, Wenjuan Cui, Hengshu Zhu, Yunchang Guo, Fayaz Dharejo, Yuanchun Zhou. 162

A Deep Neural Network for Estimating Low-Density Lipoprotein Cholesterol From Electronic Health Records: Real-Time Routine Clinical Application (e29331)
Sangwon Hwang, Chanwoo Gwon, Dong Seo, Jooyoung Cho, Jang-Young Kim, Young Uh. 174

Ranking Rule-Based Automatic Explanations for Machine Learning Predictions on Asthma Hospital Encounters in Patients With Asthma: Retrospective Cohort Study (e28287)
Xiaoyi Zhang, Gang Luo. 186

Current-Visit and Next-Visit Prediction for Fatty Liver Disease With a Large-Scale Dataset: Model Development and Performance Comparison (e26398)
Cheng-Tse Wu, Ta-Wei Chu, Jyh-Shing Jang. 208

Development and Validation of an Arterial Pressure-Based Cardiac Output Algorithm Using a Convolutional Neural Network: Retrospective Study Based on Prospective Registry Data (e24762)
Hyun-Lim Yang, Chul-Woo Jung, Seong Yang, Min-Soo Kim, Sungho Shim, Kook Lee, Hyung-Chul Lee. 232

Using a Convolutional Neural Network to Predict Remission of Diabetes After Gastric Bypass Surgery: Machine Learning Study From the Scandinavian Obesity Surgery Register (e25612)
Yang Cao, Ingmar Näslund, Erik Näslund, Johan Ottosson, Scott Montgomery, Erik Stenberg. 244

Patient-Level Cancer Prediction Models From a Nationwide Patient Cohort: Model Development and Validation (e29807)
Eunsaem Lee, Se Jung, Hyung Hwang, Jaewoo Jung. 256

Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning (e23230)
Pei-Fu Chen, Ssu-Ming Wang, Wei-Chih Liao, Lu-Cheng Kuo, Kuan-Chih Chen, Yu-Cheng Lin, Chi-Yu Yang, Chi-Hao Chiu, Shu-Chih Chang, Feipei Lai. 268

An Artificial Neural Network–Based Pediatric Mortality Risk Score: Development and Performance Evaluation Using Data From a Large North American Registry (e24079)
Niema Ghanad Poor, Nicholas West, Rama Sreepada, Srinivas Murthy, Matthias Görges. 281

Factors to Effective Telemedicine Visits During the COVID-19 Pandemic: Cohort Study (e27977)
Kristin Gmunder, Jose Ruiz, Dido Franceschi, Maritza Suarez. 296

XSL•FO
RenderX

Improving Human Happiness Analysis Based on Transfer Learning: Algorithm Development and Validation (e28292)
Lele Yu, Shaowu Zhang, Yijia Zhang, Hongfei Lin. 309

Matching Biomedical Ontologies: Construction of Matching Clues and Systematic Evaluation of Different Combinations of Matchers (e28212)
Peng Wang, Yunyan Hu, Shaochen Bai, Shiyi Zou. 321

Review

The Unified Medical Language System at 30 Years and How It Is Used and Published: Systematic Review and Content Analysis (e20675)
Xia Jing. 135

Corrigenda and Addendas

Correction: Predicting Antituberculosis Drug–Induced Liver Injury Using an Interpretable Machine Learning Method: Model Development and Validation Study (e32415)
Tao Zhong, Zian Zhuang, Xiaoli Dong, Ka Wong, Wing Wong, Jian Wang, Daihai He, Shengyuan Liu. 292

Correction: The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities (e32869)
Muhammad Ayaz, Muhammad Pasha, Mohammed Alzahrani, Rahmat Budiarto, Deris Stiawan. 294

XSL•FO
RenderX

Viewpoint

Potential Uses of Blockchain Technology for Outcomes Research on Opioids

Aldren Gonzales[1], MSc; Scott R Smith[1], PhD; Prashila Dullabh[2], MD; Lauren Hovey[2], MA; Krysta Heaney-Huls[2], MPH; Meagan Robichaud[2], MPH; Roger Boodoo[3], MD

[1]US Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation, Office of Health Policy, Washington, DC, United States

[2]NORC at the University of Chicago, Chicago, IL, United States

[3]Department of Defense, Defense Health Agency, United States Navy, Falls Church, VA, United States

Corresponding Author:
Aldren Gonzales, MSc
US Department of Health and Human Services
Office of the Assistant Secretary for Planning and Evaluation
Office of Health Policy
200 Independence Ave SW
Washington, DC, 20201
United States
Phone: 1 2028707414
Email: aldren.gonzales@hhs.gov

Abstract

The scale and severity of the opioid epidemic call for innovative, multipronged solutions. Research and development is key to accelerate the discovery and evaluation of interventions that support pain and substance use disorder management. In parallel, the use and integration of blockchain technology within research networks holds the potential to address some of the unique challenges facing opioid research. This paper discusses the applications of blockchain technology and illustrates potential ways in which it could be applied to strengthen the validity of outcomes research on the opioid epidemic. We reviewed published and gray literature to identify useful applications of blockchain, specifically those that address the challenges faced by opioid research networks and programs. We then convened a panel of experts to discuss the strengths, limitations, and feasibility of each application. Blockchain has the potential to address some of the issues surrounding health data management, including data availability, data sharing and interoperability, and privacy and security. We identified five primary applications of blockchain to opioids: clinical trials and pharmaceutical research, incentivizing data donation and behavior change, secure exchange and management of e-prescriptions, supply chain management, and secondary use of clinical data for research and public health surveillance. The published literature was limited, leading us to rely on gray literature, which was also limited in its discussion of the technical aspects of implementation. The technical expert panel provided additional context and an assessment of feasibility that was lacking in the literature. Research on opioid use and misuse is challenging because of disparate data stored across different systems, data and system interoperability issues, and legal requirements. These areas must be navigated to make data accessible, timely, and useful to researchers. Blockchain technologies have the potential to act as a facilitator in this process, offering a more efficient, secure, and privacy-preserving solution for data exchange. Among the 5 primary applications, we found that clinical trial research, supply chain management, and secondary use of data had the most examples in practice and the potential effectiveness of blockchain. More discussions and studies should focus on addressing technical questions concerning scalability and tackling practical concerns such as cost, standards, and governance around the implementation of blockchain in health care. Policy concerns related to balancing the need for data accessibility that also protects patient privacy and autonomy in revoking consent should also be examined.

(JMIR Med Inform 2021;9(8):e16293) doi:10.2196/16293

KEYWORDS

blockchain; distributed ledger; opioid crisis; outcomes research; patient-centered outcomes research; mobile phone

XSL•FO
RenderX

Background

The Opioids Problem

Prevalent misuse and overdose related to prescription opioids have created a public health crisis in the United States [1]. In 2017, approximately 11.4 million people misused prescription opioids [2]. In addition, the rise of heroin use and the increase in the availability of illicit, highly potent synthetic opioids have fueled the crisis [1]. The urgency of the problem has become more evident with the increasing number of drug overdose deaths in the country. In 2018 alone, approximately 70% of more than 67,000 drug overdose deaths recorded involved opioids. Approximately 67% or 2 of 3 opioid-related overdose deaths involved synthetic opioids [3]. The Centers for Disease Control and Prevention also estimates that, on average, 130 Americans die every day from opioid overdose [4].

Although preliminary data from 2017 to 2018 showed a 2.8% decrease in opioid overdose deaths, there is broad recognition that the crisis is far from over and continued attention and additional research are needed [5]. Responding to the opioid epidemic will require innovative multidimensional solutions coordinated across sectors that take advantage of emerging technologies. The use of existing health data has been recognized as a key component in addressing the opioid epidemic. These data have the potential to support research that advances current knowledge about pain and addiction and leads to the discovery of new treatment options and interventions [6,7]. However, research efforts and interventions that target opioid use disorder are challenged by gaps in data and information exchange across sectors, causing a real obstacle to addressing the epidemic. Although several solutions are being implemented, technology is still considered an underused asset to address the crisis [8].

Specific to the opioid epidemic, different innovative health information technology (IT) solutions have been developed and implemented. The prescription drug monitoring program (PDMP) system is one tool that state governments have invested in to provide prescribers and pharmacists access to critical information regarding patients' controlled substance prescription history. Other examples of health IT tools being implemented to support opioid-related management include the expansion of e-prescriptions [9], clinical decision support tools in electronic health record (EHR) systems [10], telehealth for addiction treatment services [11], and smartphone apps to support recovery [12].

Blockchain

Blockchain is another rapidly evolving technology with potential applications to the opioid crisis and health system–level research issues related to data availability, interoperability, and privacy and security [13]. Blockchain is a type of distributed ledger technology that uses a peer-to-peer network to provide "a shared, immutable, and transparent append-only register of all the actions that have happened to all the participants [called "nodes," which can be any organization or an individual who take part in the digital business transaction] of the network" [14,15].

Traditionally, records are managed and verified by a central authority. With blockchain, recording is decentralized, allowing all authorized users to keep an identical copy of transactions (also called *blocks*). To illustrate how it works, consider a spreadsheet document containing transactions that are duplicated and stored in a network of authorized computers (*nodes*) and are updated periodically. When a new transaction is made, the spreadsheet needs to be updated. This new transaction will be represented as a *block* on the web and will be distributed to authorized computers for verification. If the whole network says that the transaction is valid, that *block* will be appended to the chain and all the copies of the spreadsheet stored in the network will be reconciled with the new permanent record. This is the concept of a blockchain.

The technical details of blockchain technology are beyond the scope of this paper. However, to provide a backbone for discussing the use cases, the blockchain's core features and multiple advantages over traditional distributed or centralized databases are summarized here. First, each block in the chain is connected using cryptography, which prevents tampering and malicious attacks. Second, the blockchain provides a full transaction history as part of its data *blocks*. This allows each node to validate the accuracy of transaction data and reach a consensus before adding another block to the blockchain, safeguarding transparency and reliability. Third, the system architecture is distributed across members of the network, allowing members to maintain control of their data while still contributing information to the collective. Fourth, because all data and transactions (eg, entering new information and updating or deleting a record) are available to those authorized in the network, blockchain brings about trust and transparency among the participants to the records [16,17].

There are two categories of blockchain systems: public and private. A public blockchain has several member nodes connected in a decentralized way that allows anyone to participate, read, and write data to the chain. As the network is public, nodes can operate maliciously to manipulate the assets within the blockchain. On the other hand, private blockchains allow only authorized nodes to participate in the network and exchange digital assets [18,19].

Although blockchain's popularity started in 2009 in the financial sector with the use of cryptocurrencies (eg, Bitcoin), it has grown and expanded to different industries, including health care, legal, security, and government [20,21]. In health care, various proof of concepts and pilots are in progress to innovate processes and address longstanding problems in handling data [22]. As a distributed digital ledger, blockchain is believed to have the potential to mitigate some of the intractable issues in health information exchange and data management [23]. With the growing interest and development in this field, both from the private industry and government, there is a need to better understand how blockchain technologies could support opioid-related health outcome studies.

In this paper, we explore the different applications of blockchain and how it may be strategically deployed to facilitate opioid-related research. This paper can serve as a helpful resource for researchers, health IT innovators, and other

XSL•FO
RenderX

stakeholders in exploring blockchain technologies to address data challenges and data infrastructure gaps. This paper also aims to stimulate discussions on potential implementation challenges and encourage more research to generate evidence on blockchain's applicability in health care research.

Methods

This paper is informed by two primary data collection activities: (1) a search and review of peer-reviewed and gray literature and (2) a technical expert panel (TEP).

Literature Review

A literature review was conducted to identify challenges in conducting opioid-related research and to understand the potential uses of blockchain to address those challenges, including its limitations and other implementation considerations. Given that the application of blockchain to health care is an emerging field, peer-reviewed and gray literature was assessed in this paper. Search strings used included terms related to longitudinal health records, data sharing, and research applications for blockchain, as well as direct references to blockchain and opioids. Searches were conducted using PubMed, Google Scholar, and Google search engine. Supplemental literature was obtained from three additional sources: (1) the TEP and a subject matter expert advising the team, (2) government papers and reports [16], and (3) white papers from the industry and academia [24]. We conducted a title and abstract review of the 458 search results, followed by a full-text review that yielded 104 relevant articles.

TEP Engagement

As the use of blockchain in health care is new and emerging, the literature contains limited technical information on the technologies and minimal assessment of their feasibility. As such, 3 experts were recruited from the field of health IT, blockchain technology development, and opioids research to offer an *on-the-ground* perspective. In particular, we discussed research challenges associated with opioid outcomes research, how blockchain applications relate to opioid research in particular, and challenges and recommendations for using blockchain in opioid research.

Challenges in Opioid Research

The number of opioid-relevant data sources, their diversity, and natural heterogeneity creates challenges for opioid researchers. To assess the effectiveness of interventions and programs directed to help address the crisis, researchers and clinicians need data that are accessible, high quality, robust, and timely [25]. They also need improved tools and services that allow researchers to better manage opioid-related data, as well as improve the efficiency of data integration throughout the research life cycle [6]. The challenges identified in the literature also highlight the need to harmonize and link data sources for analysis [26] and overcome systemic barriers to interoperability [27]. Opioid use disorder patient data tend to be scattered across institutions and service points (eg, criminal justice, health care, and substance abuse treatment systems) that are involved in providing care, interventions, and assistance. As these institutions have information systems that are not designed to interoperate with other systems, they create barriers to effective care coordination for clinicians and longitudinal data access for researchers [25].

Another unique challenge in opioid research is the 42 Code of Federal Regulations (CFR) restriction on the disclosure and use of records of patients with substance use disorder (SUD), which are maintained in connection with a part 2 program. Part 2 programs are federally assisted programs for individuals, entities other than a general medical facility, or an identified unit within a general medical facility that *holds itself out* as providing SUD diagnosis, treatment, or referral for treatment [28]. As patients with SUD can potentially face stigma and discrimination, their data are protected by stringent privacy policies [29]. Although privacy policies such as the Health Insurance Portability and Accountability Act of 1996 and 42 CFR Part 2 encourage patients to seek care without thinking of potential negative consequences, they inhibit communication among providers and sharing of information through health information exchange unless very narrow circumstances are met and patient consent for data sharing is obtained [30]. As a result, researchers have limited access to real-world data [31] that are essential for tracking patient trajectories, conducting risk prediction, and outcome analysis.

Textbox 1 highlights the challenges identified in the literature review. The textbox also provides more context and descriptions of the challenges.

XSL•FO
RenderX

Textbox 1. List of challenges in conducting opioid research.

Recruitment and Retention of Study Participants

- Patient skepticism about research and the health care system and concern about confidentiality and privacy [32] often result in lower consent rates.
- Difficulty and cost of collecting longitudinal health data about substance use disorder, particularly because study attrition rates are often high.

Data Integrity, Accuracy, or Completeness

- Lack of completeness in electronic health records and other data sources due to absence in alignment between workflow and documentation, user error, and use of nonstandardized free-text fields that cannot easily be converted into analyzable data.
- If data are not entered into the electronic health record, the data are not available for research (eg, patient pain agreements that outline opioid use disorder preventive actions and paper prescriptions) [8].
- Underreporting of opioid use and deaths due to challenges associated with collecting vital statistics and hospital data [33]:
  - Death certificates do not always specify drugs that contributed to death, causing researchers to underestimate opioid overdose deaths [34].
  - Neonatal abstinence syndrome surveillance relies primarily on hospital discharge data, which may not capture cases of opioid use disorder diagnosed during other points in pregnancy [35,36].
- Limitations of self-reported opioid use in nationally representative surveys stemming from participants' lack of knowledge on opioid misuse, overly broad questions (eg, on overall use, rather than specific opioids), and lack of awareness about exposure to adulterated drugs [37].

Timeliness of Data

- Reporting to prescription drug monitoring programs can vary across states, between 5 minutes and 8 days [38]. The latest report from the National Vital Statistics System on drug overdose deaths and the current National Survey on Drug Use and Health data are still from 2018 [3,39].
- Some Medicare data files have a lag time of 2-4 years [40].

The Need for Linked Data

- Opioid data exist in silos across health systems [41]: first responder organizations, medical examiners or coroners, law enforcement entities, criminal justice entities, treatment providers, and other stakeholders.
- Differences in unique respondent identifiers between data sets can make it difficult to match respondents' data.

Data Security and Privacy

- Substance use disorder or opioid use disorder data are subject to additional regulations as "sensitive protected health information," including opt-in consent policies (42 Code of Federal Regulation Part 2) and Health Insurance Portability and Accountability Act.
- Disclosure of information that would identify opioid use disorder requires written consent, which limits the data available for opioid research [42].

Data Sharing and Interoperability

- Data sharing agreements among public health departments, providers, health systems, and federal agencies are complex and time intensive to create [43].
- The high cost of data exchange is a barrier [44].
- Lack of interoperable electronic health records prevents information sharing on prescriptions, and prescription drug monitoring programs or electronic health record integration has not been widely implemented.
- Prescription drug monitoring programs are operated individually by state governments, and requirements around who is able to or who is required to access them vary widely [38].

Data Collection Gaps at the Point of Care and Secondary Use of Clinical Data

- Adoption and awareness of opioid prescribing guidelines vary widely [8]
  - Providers may not consider nonopioid alternatives for pain management.
  - Providers may not prescribe appropriate opioid types, doses, and quantities or durations tailored to the patients' specific types of pain (eg, acute vs chronic).

XSL•FO
RenderX

Blockchain Applications

Overview

There is growing interest among health IT innovators in the application of blockchain in health care. Specifically, blockchain is used to address issues related to data availability, interoperability, and privacy and security, as they relate to care management and health research [45-47].

The value of blockchain in research can be demonstrated through its potential in improving health research at multiple levels by improving data quality, reducing cost, decreasing administrative delays by fast-tracking vendor proposal review and contract purchasing, and accelerating the time needed to translate research to practice [48]. When applied to specific points in the research lifecycle, blockchain could potentially increase access to research studies by making more publications available through its network, streamline processes for securing funding [49], and improve transparency to prevent tampering and misreporting of data findings [50].

Although blockchain cannot be characterized as a panacea for the many challenges in opioid research, the results of the literature review and inputs from the TEP indicate that it may be strategically deployed to solve several major challenges. The 5 applications that emerged consistently in the literature as promising areas for applying blockchain are summarized in Table 1 with details about the problem area, ways blockchain could potentially address the problem, and a real-world example.

Table 1. Blockchain applications to address opioid research challenges.

Blockchain application themes	Needs or problem areas identified	How blockchain can be used to address opioid research challenges	Real-world example
Clinical trials and pharmaceutical research	Management of and transparency in reporting clinical trials and consent management	Creating a trusted record to support clinical trials with public data and findings and facilitating data movement via secure sharing and consent	Simulation of a clinical trial study (using actual raw data) on the efficacy and safety of omalizumab (asthma and chronic idiopathic urticaria drug) using a blockchain-based system to test the resilience of the data to tampering and improve traceability [51].
Incentivizing data donation and behavior change	Sharing of data for research and healthy behavior change	Use of cryptocurrency to distribute incentives that target data donation and adoption of healthy behavior	A secure and transparent distributed personal data marketplace (using blockchain) that allows users to sell their biomedical data and customers to buy data for research and analysis using cryptocurrency [52].
Secure exchange and management of electronic opioid prescriptions	Inaccurate prescription data, multiple active opioid prescriptions, and the need for better tools and systems for monitoring opioid prescriptions	Reducing fraud in e-prescribing and improving timeliness and ease of data reporting and surveillance	A solution that uses blockchain to track specific prescriptions using a machine-readable code from the time a drug was prescribed to distribution by the pharmacy. The code, which serves as the unique identifier, is associated with the prescription information, thereby augmenting PDMP[a] data and allowing pharmacists to verify its accuracy and eligibility to be filled [53].
Supply chain management	Theft or diversion of drugs, the introduction of counterfeit medicines, and contamination of drugs during production and distribution	Improving drug traceability	A pilot project uses blockchain technology to assist the US Food and Drug Administration and members of the pharmaceutical distribution supply chain in the development of an electronic, interoperable system to identify and trace certain prescription drugs as they are distributed within the United States [54].
Secondary use of clinical data	Patient data stored in different information systems, interoperability, availability of longitudinal data	Increasing researchers access to longitudinal and population-level outcome data that support research and near real-time public health surveillance	A blockchain-based information management system that allows secure access and sharing of patient records from one EHR[b] to another [55].

[a]PDMP: prescription drug monitoring program.

[b]EHR: electronic health record.

Clinical Trials and Pharmaceutical Research

Among the different applications of blockchain in health care, its use in clinical trials and pharmaceutical research is cited as among the most promising [56-59]. Research is important for developing the tools and understanding to identify new treatment options, identify ways to prevent adverse outcomes, and identify other viable opioid alternatives to manage pain [60,61]. Furthermore, trust in the validity of research data and analysis is critical for translating clinical trial results to quality clinical care. Consent collection and management is one area where blockchain can be used to support research. Using *smart contracts*, a simple computer program that is used in blockchain to digitally facilitate, verify, and enforce contracts [62], can help researchers capture all aspects of data that might be subject to manipulation including trial registration, protocol, subject registration, and clinical measurement. This technology can be

XSL•FO
RenderX

used to record patient consent to participate and allow consent to be audited to ensure adherence to recruitment guidelines [63].

Blockchain can also support clinical trial management and reporting. Incomplete and inaccurate reporting of clinical trial data [64,65] can lead to problems for regulators such as the Food and Drug Administration in auditing data, real-time oversight, and adverse event reporting [23,51]. Compliance with 21 CFR Part 11 for late stage preclinical and clinical research can be cost-prohibitive for companies, which might compromise data integrity; however, distributed clinical trial management and trial data via blockchain can create an immutable audit trail that allows users to ensure that the results have not been tampered with [64,66,67]. For phase IV clinical trials that look into drug or device safety over time, regulators could use blockchain smart contracts to automatically query clinical trial sites for adverse events [23].

Blockchain can allow researchers to maintain ownership of study data while publishing results in real time. This could encourage faster dissemination of results and potentially seed collaboration with other researchers working on the same topic [68]. Specific to opioids, blockchain "could help the research and development of opiate alternatives by laying the groundwork for a decentralized database of [laboratory] test results with free access to this data" [59], which could lead to more cost-efficient drug development [67].

Several prototypes use blockchain technology as a data platform, in addition to existing distributed clinical trial infrastructures. One example is a prototype developed by a university that enables access to research trial data abstracted from EHRs. It uses blockchain technologies such as cryptography and smart contracts to record patient consent and to track distributed researcher queries from trial data repositories stored *off-chain* [67]. A proof of concept was developed, which makes data collection within the trial life cycle immutable yet traceable and potentially more trustworthy [51]. Another blockchain-based platform was developed to assist pharmaceutical and biotechnology industries in simplifying recording processes and ensuring data integrity and fidelity in all phases of the research process [63,69].

Incentivizing Data Donation and Behavior Change

Blockchain technologies can encourage patients to share their medical information with researchers, both through secure sharing and incentives. These data increase researchers' access to longitudinal data, which then support better outcome research to study the opioid crisis [70,71]. The use of blockchain can reward users through cryptocurrency to participate in networks [72]. It could also shift data stewardship from centralized authorities (eg, the National Institutes of Health, research networks, or academic research centers) so that patients and researchers can manage their data access rights [73].

One example is a platform for storing patient data that leverages artificial intelligence and blockchain technology to support a marketplace for individuals and biobanks to store, manage, and control access to genomic and other health data [52]. A number of companies [74,75] also manage DNA marketplaces where individuals can share DNA data in exchange for cryptocurrency, which are then purchased by researchers.

In addition to incentivizing data sharing, blockchain technologies can drive users to follow health recommendations. Examples of existing platforms target the adoption of wellness activities [76,77] and reduction of doctor appointment no-shows [78]. Another platform rewards patients with cancer through cryptocurrency to report side effects, medication adherence, and healthy lifestyle choices. Oncologists can also receive rewards to create content to help monitor the patient's inputs [79].

The idea of providing nonmonetary incentives in the area of substance abuse is not new, and many studies have already demonstrated positive outcomes [80-82]. Although the identified examples are not opioid-specific, they illustrate the blockchain's potential in this use case. Researchers can implement blockchain-enabled technology to incentivize participation in studies, attend follow-up sessions, submit patient-reported measures, and encourage adherence to medication and treatment.

Secure Exchange and Management of Electronic Opioid Prescriptions

Moving from paper to electronic prescribing reduces (but does not eliminate) prescription theft and forgery and facilitates tracking of prescription histories for prescribers, pharmacists, and patients [8]. When securely stored on a blockchain, e-prescriptions can be made tamperproof, which is further secured by monitoring for potential misuse [83].

Blockchain-enabled systems are believed to be beneficial for the real-time capture and verification of prescriptions. One company developed a platform that uses a process to ensure e-prescription fidelity using blockchain. First, every new prescription is assigned a unique identifier in the form of a machine-readable symbol that is associated with a block of prescription details (eg, drug, dosage, anonymized patient identifier, and timestamp). Second, pharmacists can scan the code, verify that the data block matches the prescription details, and document that it has been filled. The blocks are stored in multiple places as part of the distributed ledger system, encrypted during transmission, and only accessible with the correct cryptographic key, and therefore more trustworthy [84].

Leveraging electronic prescription technology, PDMPs are intended to capture and disseminate real-time information about opioid prescribing practices and prevent *doctor and pharmacy shopping* in which a patient seeks the same prescription from multiple providers or attempts to fill prescriptions at multiple pharmacies to decrease fill denial [85]. Having access to timely PDMP data provides researchers the opportunity to analyze prescribing patterns, which are central in combatting opioid misuse and addiction. Unfortunately, PDMP reporting intervals vary widely, ranging from 5 minutes to 8 days [36]. A blockchain-enabled monitoring system would allow real-time verification of previous prescriptions by doctors as they consider prescribing new opioids, followed by real-time reporting of new prescriptions, and real-time verification by pharmacists filling that prescription. Reducing the time and friction of data transfer

XSL•FO
RenderX

also reduces costs and can be deployed nationwide instead of state by state [86].

Supply Chain Management

Addressing supply chain issues such as theft or diversion, the introduction of counterfeit medicines, and contamination of medicines during manufacturing, storage, or distribution [87] by ensuring the provenance and authenticity of the drugs are crucial to patient safety. Blockchain can help with 2 traceability issues. First, it allows companies to track their products down the supply chain, creating a circuit that is secure and difficult to penetrate by counterfeit products. Second, it allows stakeholders, especially laboratories, to identify the exact location of their drugs in case of a problem [88].

Several studies have proposed applying blockchain to supply chain information exchange and data storage [87,89-91]. Transaction information regulated by the 2013 Drug Supply Chain Security Act [89], that is, product information, transaction history, and ownership, can be stored and managed on a blockchain [87]. This allows network participants to track and validate a drug pill by pill [88] from manufacturing to dispensing, detecting anomalies and identifying missing drug products and unauthorized data insertions [92]. The private industry has shown great interest in this use case and has developed several working platforms and proof-of-concept designs exploring the feasibility of blockchain in addressing supply chain issues [89,93,94]. Another pilot project focused on applying blockchain-based technology to supply chain management by identifying unused oral cancer drugs and giving them to patients who cannot afford them [95].

Specific to opioids, the US government and private companies are testing the use of blockchain-based supply chains to help improve the security of prescription drug supply and distribution and to allow real-time monitoring of pharmaceutical products [96]. Various efforts are also being implemented to track dispensing [97] and expand access [98] to naloxone, a drug used to counteract opioid overdoses. Although policies allowing naloxone dispensing through standing order have shown a significant reduction in opioid-related deaths [99-101], this could create information gaps regarding its distribution and use. As naloxone is distributed to first aid responders and laypersons, the recording of its use may not always be consistent. Having access to this information could help public health officials and researchers identify trends in opioid use using naloxone as a marker for opioid overdose and assess the impact of local policies related to naloxone distribution efforts [102,103].

Secondary Use of Clinical Data

A comprehensive and accurate view of a patient's trajectory over time, across providers, and across health care and non–health care settings is crucial to the ability to precisely answer research questions and conduct near real-time public health surveillance. However, patient health data are often collected and stored in disparate information systems (eg, emergency department registries and first responder data systems), greatly reducing researchers' access to longitudinal records for real-world evidence generation for opioid-related treatment [55,104] and public health surveillance [19]. Accessing data from disparate health information systems requires high overhead costs, and systems often lack basic computer security protocols to authenticate patient data.

Aside from access concerns, dealing with substance use data is unique because it requires special and careful handling. Privacy and security are critically important, given that SUD-like opioid misuse remains highly stigmatized and is therefore classified as *sensitive protected health information*. This means that they require extra protection and consent requirements under 42 CFR Part 2 [105], creating obvious barriers for opioid researchers [40] and information exchange [106]. Data from nonsubstance abuse treatment providers only provide part of the picture of a patient's trajectory. SUD and associated SUD treatment data are crucial to an understanding of patient outcomes in relation to treatment settings (eg, detox and residential) and treatment types (eg, medication-assisted treatment). Blockchain can provide structure and security for improved data sharing by creating a concurrent, distributed, redundant, and secure system that facilitates record linkage and improves interoperability between data sharing partners [107] for research and surveillance use cases.

An example of how blockchain technology can be implemented to facilitate record linkage is a platform [55] that was developed to gather and link information from disparate patient records without central data storage. It allows authorized users to upload encrypted clinical summaries for cross-system sharing and to easily search and retrieve patient health information that has been shared across systems. Government health agencies are also exploring ways to further maximize the potential of information in EHRs and how blockchain can be used for public health surveillance [108]. Blockchain is seen as a technology that could complement the public health's complicated peer-to-peer model for data sharing "to more efficiently manage data during a crisis or to better track opioid abuse" [109].

Discussion

Overview

The abovementioned applications demonstrate the potential of blockchain to support opioid research. However, blockchain is a relatively new technology in health care, necessitating a critical assessment and testing of its suitability for the opioid research challenges identified in this paper. It is also recognized that blockchain is not only and may not be the best solution for each of the identified gaps. On the basis of the literature, the application of blockchain in clinical trials and pharmaceutical research is ready for more real-world implementation. Other applications could provide evidence of data fidelity and provenance to improve the broad and timely sharing of clinical trial data and accelerate the pace of research [63].

Input from the TEP regarding blockchain applications with the most potential impact on opioid research include increasing access to longitudinal and population-level outcome data for research and surveillance [55] and supply chain management [57,88], with an emphasis on monitoring administration of opioid overdose reversal medication. The TEP also highlighted areas where blockchain could potentially support the

XSL•FO
RenderX

administrative side of health care research: grant proposal processing and review; financial distribution of research funds and longitudinal tracking of dollars to demonstrate return on investment; regulatory tracking and auditing of research that lower admin cost while ensuring compliance; and facilitating more rapid dissemination of findings.

Challenges and Limitations

Overview

Blockchain applications in health care are still in their infancy. There is a need to further study and test the true feasibility for health care applications in general and in areas that require a high degree of privacy and confidentiality protection (eg, SUD). Near-term opportunities to support growth in the blockchain market should focus on real-world implementation to assess challenges and limitations.

Technical Challenges

There are several technical challenges in using blockchain for health care research. First, existing data infrastructures may require new mechanisms to interface with blockchain. For example, collecting and managing data with blockchain solutions may require technical upgrades to existing systems. Depending on the data sources (eg, EHRs, distributed research network data marts, clinical trial registries, and PDMPs), different technical solutions may be required at varying costs and complexities [110]. Relatedly, the TEP identified a need for discussion on internet requirements and capabilities for running these new blockchain applications.

Second, the lack of metadata standards for information stored in the blockchain may challenge interoperability. This includes a lack of standards around smart contracts, which are needed before the blockchain can be applied consistently [63].

Third, blockchain does not address all existing challenges in data interoperability and validity. There is wide variability in the standardization of electronic health information, which is not solved by blockchain. Data validity concerns are also not fully resolved by blockchain because data stored off-chain can still be manipulated before being added to the ledger [66].

Fourth, not all use cases have a clear business model to incent implementation and participation. Despite the potential for many of these use cases, they all require investments in infrastructure and incentives for adoption and use [111], and for networks, it is necessary to determine how to distribute costs.

Finally, scalability with regard to processing power has been identified as a key challenge to implementation [112]. The use of blockchain to store vast amounts of data (eg, millions of patient records, multi-institutional data, and global research records) can incur equally vast storage costs. Blockchain networks have defined data size limits that may be quickly exceeded, depending on the use case [113].

Policy Challenges

A complicated legal and regulatory framework governs the use and disclosure of patient health information, with additional federal and state laws governing specially protected health information, such as substance use (eg, 42 CFR Part 2, section 3221 of the Coronavirus Aid, Relief, and Economic Security Act) and genetic data (eg, Genetic Information Nondiscrimination Act and the US National Institutes of Health Genomic Data Sharing policy). Some state and international laws have requirements regarding data destruction upon revocation of patient consent [114,115]. This has prompted some organizations to exclude specially protected health information from their sharable records. Understanding how blockchain implementations can link and integrate SUD treatment records with other health care data while managing confidentiality requirements is critical to the application of blockchain to opioid research and treatment. Examining blockchain implementations in Europe may provide some early lessons learned regarding how implementers manage blockchain immutability while complying with the General Data Protection Regulation.

One of the key characteristics of blockchain is that it promotes trust in transactions and records—a significant attribute given patients' expectations of privacy and confidentiality. Ambiguities in current and future policies governing cryptocurrency may limit its potential and challenge users to comply with legal requirements (eg, if cryptocurrency or blockchain tokens are classified as security, they will become subject to Securities and Exchange Commission rules) [62]. Federal regulators should consider their role in ensuring trust in the larger clinical trial ecosystem and other data donation initiatives by encouraging the private industry to adopt strong privacy and confidentiality requirements. For example, the CARIN Alliance has developed a trust framework and voluntary code of conduct for stakeholders and organizations entrusted with personally identifiable information [116].

The industry also needs data governance rules for which entities can write data to an official chain. For example, blockchain can reduce the burden of credentialing for providers by enabling providers to self-assert rather than requiring an intermediary such as a medical board to issue the claim on behalf of the provider [86,117]. In some instances, this may require changes to existing laws regarding the use of digital signatures, which support writing information in a chain.

In terms of policies related to blockchain, most regulatory discussions are happening at the federal agency level around its use for cryptocurrency. This is despite the recognition of blockchain applications in other areas. In 2019, there were 27 state bills and resolutions relating to blockchain, which have been enacted or adopted. These resolutions tackle more applications outside of cryptocurrency, such as examining blockchain's use for elections (Connecticut and New York), state administrative transactions (Connecticut), and health care use cases (Virginia) [118,119].

Potential Opportunities

Solving these technical and policy challenges requires a coordinated approach between the private and public industries, as illustrated by the real-world examples presented herein. To further explore blockchain to opioid-related research, the TEP encouraged work in the following areas:

XSL•FO
RenderX

1. There is a general lack of education and understanding of blockchain, which challenges its application in research. Raising awareness among key research stakeholders will increase knowledge and build competencies, priming the research community for implementation. Given the technical and policy challenges, researchers should be encouraged to use industry guidance to assess the feasibility of blockchain applications. The National Institute of Standards and Technology may offer one such tool to focus on researchers and developers working to identify near-term opportunities that are fit for blockchain [16].
2. There is a lack of standards for smart contracts. Given that many blockchain applications rely on smart contracts, developing standards and policies can improve interoperability within a network [63].
3. The legal and regulatory environment for the use and disclosure of substance use treatment information poses challenges for data sharing [29]. Studies related to the technical solutions for implementing blockchain in this complex ecosystem (considering various policy and privacy requirements) are needed to realize the potential of integrating SUD treatment records with other health care data to track patient outcomes over time.

Conclusions

Researchers are currently faced with a number of challenges with access to and use of opioid-related data. Aside from the need for high-quality, accessible, robust, and timely data, opioid researchers are also confronted by siloed information systems and the privacy requirements for SUD data. Considering the features and capabilities of blockchain and its current application in other industries, it has the potential to act as a facilitator to address these challenges by offering a more efficient, secure, and privacy-preserving solution to the research process, data management, and data exchange. Among the 5 primary applications that we identified, its use in clinical trial research, supply chain management, and secondary use of data for research and public health surveillance had the most evidence for implementation opportunities and potential for the effectiveness of blockchain. Although these blockchain applications present great potential, future work should understand and address concerns related to standards, infrastructure, scalability, implementation cost, sustainability, and governance. Policy concerns related to balancing the need to create high-fidelity data that also protect patient privacy and patient autonomy in revoking consent to use their data for research and treatment should also be addressed. Discussion and evidence generation efforts should focus on addressing these challenges to evaluate the feasibility and at the same time maximize the potential of blockchain technology.

Acknowledgments

This paper was partially prepared under contract #HHS23320160020I (task order number HHSP23337001T) between the Department of Health and Human Services' Office of the Assistant Secretary for Planning and Evaluation - Office of Health Policy and NORC at the University of Chicago. The contract was funded by the Office of the Secretary–Patient-Centered Outcomes Research Trust Fund. The views expressed in this paper are those of the authors and do not necessarily represent the views of the US Department of Health and Human Services, NORC at the University of Chicago, the Defense Health Agency, the Department of Defense, or the United States Government. The authors would like to thank Debbi Bucci (Lead IT Architect at the US Department of Health and Human Services, Office of Standards and Interoperability, Office of the National Coordinator), Kristin A Lyman, JD, MHA (Associate Director, at the Louisiana Public Health Institute), and Sean Manion, PhD (Chief Executive Officer of Science Distributed and Co-Chief Editor of Frontiers: Blockchain for Science) for their valuable contributions as TEP members.

Conflicts of Interest

RB owns nontrivial amounts of Bitcoin, Litecoin, Ethereum, and other cryptocurrencies. The other authors have no conflicts to declare.

References

1. Skolnick P. The Opioid Epidemic: Crisis and Solutions. Annu Rev Pharmacol Toxicol 2018 Jan 06;58:143-159. [doi: 10.1146/annurev-pharmtox-010617-052534] [Medline: 28968188]
2. What is the U.S. Opioid Epidemic? Department of Health and Human Services (US). URL: https://www.hhs.gov/opioids/about-the-epidemic/index.html [accessed 2019-09-12]
3. Drug Overdose Deaths. Centers for Disease Control Prevention (US). 2020. URL: https://www.cdc.gov/drugoverdose/data/statedeaths.html [accessed 2020-05-18]
4. Opioids: Understanding the Epidemic. Centers for Disease Control and Prevention (US). 2018. URL: https://www.cdc.gov/drugoverdose/epidemic/index.html [accessed 2019-07-18]
5. Ehley B. U.S. 'turning the tide' on the opioid crisis, health secretary says. Politico. URL: https://www.politico.com/story/2018/10/23/opioid-crisis-health-secretary-932332 [accessed 2019-09-11]
6. National Science and Technology Council. Health Research and Development to Stem the Opioid Crisis: A National Roadmap. 2018. URL: https://www.nih.gov/sites/default/files/Health-RD-to-Stem-Opioid-Crisis-2018-Roadmap-for-Public-Comment.pdf [accessed 2019-08-01]

XSL•FO
RenderX

7. Department of Health and Human Services (US). Strategy to Combat Opioid Abuse, Misuse, and Overdose: A Framework Based on the Five Point Strategy. 2018. URL: https://www.hhs.gov/opioids/sites/default/files/2018-09/opioid-fivepoint-strategy-20180917-508compliant.pdf [accessed 2019-09-12]
8. Addressing Data and Information Gaps Contributing to Opioid Use Disorder (Policy Brief). Network for Excellence in Health Education. 2018. URL: https://www.nehi-us.org/writable/publication_files/file/nehi_opioids_policy_brief_final.pdf [accessed 2019-09-04]
9. Gawande AA. It's Time to Adopt Electronic Prescriptions for Opioids. Ann Surg 2017 Apr;265(4):693-694. [doi: 10.1097/SLA.0000000000002133] [Medline: 28067675]
10. Finley EP, Schneegans S, Tami C, Pugh MJ, McGeary D, Penney L, et al. Implementing prescription drug monitoring and other clinical decision support for opioid risk mitigation in a military health care setting: a qualitative feasibility study. J Am Med Inform Assoc 2018 May 01;25(5):515-522 [FREE Full text] [doi: 10.1093/jamia/ocx075] [Medline: 29025024]
11. Eibl JK, Gauthier G, Pellegrini D, Daiter J, Varenbut M, Hogenbirk JC, et al. The effectiveness of telemedicine-delivered opioid agonist therapy in a supervised clinical setting. Drug Alcohol Depend 2017 Jul 01;176:133-138 [FREE Full text] [doi: 10.1016/j.drugalcdep.2017.01.048] [Medline: 28535455]
12. Steinkamp JM, Goldblatt N, Borodovsky JT, LaVertu A, Kronish IM, Marsch LA, et al. Technological Interventions for Medication Adherence in Adult Mental Health and Substance Use Disorders: A Systematic Review. JMIR Ment Health 2019 Mar 12;6(3):e12493 [FREE Full text] [doi: 10.2196/12493] [Medline: 30860493]
13. Mackey TK, Kuo T, Gummadi B, Clauson KA, Church G, Grishin D, et al. 'Fit-for-purpose?' - challenges and opportunities for applications of blockchain technology in the future of healthcare. BMC Med 2019 Mar 27;17(1):68 [FREE Full text] [doi: 10.1186/s12916-019-1296-7] [Medline: 30914045]
14. Dubovitskaya A, Novotny P, Xu Z, Wang F. Applications of Blockchain Technology for Data-Sharing in Oncology: Results from a Systematic Literature Review. Oncology 2020;98(6):403-411 [FREE Full text] [doi: 10.1159/000504325] [Medline: 31794967]
15. Hau YS, Lee JM, Park J, Chang MC. Attitudes Toward Blockchain Technology in Managing Medical Information: Survey Study. J Med Internet Res 2019 Dec 09;21(12):e15870 [FREE Full text] [doi: 10.2196/15870] [Medline: 31815676]
16. National Institute of Standards and Technology. Blockchain Technology Overview (NISTIR 8202). 2018. URL: https://nvlpubs.nist.gov/nistpubs/ir/2018/NIST.IR.8202.pdf [accessed 2019-07-16]
17. Alladi T, Chamola V, Rodrigues JJPC, Kozlov SA. Blockchain in Smart Grids: A Review on Different Use Cases. Sensors (Basel) 2019 Nov 08;19(22):4862 [FREE Full text] [doi: 10.3390/s19224862] [Medline: 31717262]
18. Koštál K, Helebrandt P, Belluš M, Ries M, Kotuliak I. Management and Monitoring of IoT Devices Using Blockchain. Sensors (Basel) 2019 Feb 19;19(4):856 [FREE Full text] [doi: 10.3390/s19040856] [Medline: 30791392]
19. Chattu VK, Nanda A, Chattu SK, Kadri SM, Knight AW. The Emerging Role of Blockchain Technology Applications in Routine Disease Surveillance Systems to Strengthen Global Health Security. BDCC 2019 May 08;3(2):25. [doi: 10.3390/bdcc3020025]
20. Marr B. Here Are 10 Industries Blockchain Is Likely To Disrupt. Forbes. 2018. URL: https://www.forbes.com/sites/bernardmarr/2018/07/16/here-are-10-industries-blockchain-is-likely-to-disrupt/#6dace152b5a2 [accessed 2019-08-20]
21. Leeming G, Ainsworth J, Clifton D. Blockchain in health care: hype, trust, and digital health. The Lancet 2019 Jun;393(10190):2476-2477 [FREE Full text] [doi: 10.1016/s0140-6736(19)30948-1]
22. Peterson K, Deeduvanu R, Kanjamala P, Boles K. A Blockchain-Based Approach to Health Information Exchange Networks. 2016. URL: https://www.healthit.gov/sites/default/files/12-55-blockchain-based-approach-final.pdf [accessed 2019-09-05]
23. Zhuang Y, Sheets L, Shae Z, Tsai JJP, Shyu C. Applying Blockchain Technology for Health Information Exchange and Persistent Monitoring for Clinical Trials. AMIA Annu Symp Proc 2018;2018:1167-1175 [FREE Full text] [Medline: 30815159]
24. Office of the National Coordinator of Health IT. Blockchain Technology and the Potential for Its Use in Health IT and/or Healthcare Related Research Data. Blockchain Challenge on ONC TechLab. 2017. URL: https://oncprojectracking.healthit.gov/wiki/display/TechLabI/Blockchain+Challenge+on+ONC+Tech+Lab [accessed 2019-09-02]
25. Smart R, Kase CA, Meyer A, Stein B. Data Sources and Data-Linking Strategies to Support Research to Address the Opioid Crisis - Final Report. 2018. URL: https://aspe.hhs.gov/system/files/pdf/259641/OpioidDataLinkage.pdf [accessed 2019-08-28]
26. Torrance N, Mansoor R, Wang H, Gilbert S, Macfarlane GJ, Serpell M, et al. Association of opioid prescribing practices with chronic pain and benzodiazepine co-prescription: a primary care data linkage study. Br J Anaesth 2018 Jun;120(6):1345-1355 [FREE Full text] [doi: 10.1016/j.bja.2018.02.022] [Medline: 29793600]
27. Monica K. CMS Opioid Roadmap Prioritizes Healthcare Interoperability. EHR Intelligence. 2018. URL: https://ehrintelligence.com/news/cms-opioid-roadmap-prioritizes-healthcare-interoperability [accessed 2019-09-02]
28. Office of the National Coordinator of Health IT, Substance Abuse and Mental Health Services Administration. Disclosure of Substance Use Disorder Patient Records: Does Part 2 Apply to Me?. URL: https://www.samhsa.gov/sites/default/files/does-part2-apply.pdf [accessed 2019-09-02]
29. Hu LL, Sparenborg S, Tai B. Privacy protection for patients with substance use problems. Subst Abuse Rehabil 2011;2:227-233 [FREE Full text] [doi: 10.2147/SAR.S27237] [Medline: 24474860]

XSL•FO
RenderX

30. McCarty D, Rieckmann T, Baker RL, McConnell KJ. The Perceived Impact of 42 CFR Part 2 on Coordination and Integration of Care: A Qualitative Analysis. Psychiatr Serv 2017 Mar 01;68(3):245-249 [FREE Full text] [doi: 10.1176/appi.ps.201600138] [Medline: 27799017]
31. Sarata AK, Redhead CS. Privacy Protections for Individuals with Substance Use Disorders: The Part 2 Final Rule in Brief. Congressional Research Service. 2018. URL: https://fas.org/sgp/crs/misc/R44790.pdf [accessed 2019-09-09]
32. Neale J, Tompkins CNE, McDonald R, Strang J. Improving recruitment to pharmacological trials for illicit opioid use: findings from a qualitative focus group study. Addiction 2018 Jun;113(6):1066-1076 [FREE Full text] [doi: 10.1111/add.14163] [Medline: 29356208]
33. Buchanich JM, Balmert LC, Williams KE, Burke DS. The Effect of Incomplete Death Certificates on Estimates of Unintentional Opioid-Related Overdose Deaths in the United States, 1999-2015. Public Health Rep 2018;133(4):423-431 [FREE Full text] [doi: 10.1177/0033354918774330] [Medline: 29945473]
34. Giroir B. Expanding Access to Treatment for Opioid Use Disorder: HHS Opioid Epidemic Update. URL: https://tinyurl.com/49fwd8by [accessed 2019-07-05]
35. Jilani SM, Frey MT, Pepin D, Jewell T, Jordan M, Miller AM, et al. Evaluation of State-Mandated Reporting of Neonatal Abstinence Syndrome - Six States, 2013-2017. MMWR Morb Mortal Wkly Rep 2019 Jan 11;68(1):6-10 [FREE Full text] [doi: 10.15585/mmwr.mm6801a2] [Medline: 30629576]
36. Haight SC, Ko JY, Tong VT, Bohm MK, Callaghan WM. Opioid Use Disorder Documented at Delivery Hospitalization - United States, 1999-2014. MMWR Morb Mortal Wkly Rep 2018 Aug 10;67(31):845-849 [FREE Full text] [doi: 10.15585/mmwr.mm6731a1] [Medline: 30091969]
37. Palamar JJ. Barriers to accurately assessing prescription opioid misuse on surveys. Am J Drug Alcohol Abuse 2019;45(2):117-123 [FREE Full text] [doi: 10.1080/00952990.2018.1521826] [Medline: 30230924]
38. State PDMP Profiles and Contacts. Prescription Drug Monitoring Program Training Technical Assistance Center. URL: https://www.pdmpassist.org/State [accessed 2019-07-05]
39. Key substance use mental health indicators in the United States: Results from the 2018 PEP19-5068, NSDUH Series H-54) National Survey on Drug Use Health (HHS Publication No. Substance Abuse and Mental Health Services Administration. 2019. URL: https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHNationalFindingsReport2018/NSDUHNationalFindingsReport2018.pdf.2019 [accessed 2019-12-10]
40. Master Beneficiary Summary File (MBSF) Base. Research Data Assistance Center (ResDAC). URL: https://www.resdac.org/cms-data/files/mbsf-base [accessed 2019-07-05]
41. O'Brien M. Notes from the Field: Opioid Crisis Understanding the Opioid Crisis Through Data and All-Stakeholder Reviews. National Institute of Justice. URL: https://nij.ojp.gov/topics/articles/notes-field-opioid-crisis-understanding-opioid-crisis-through-data-and-all [accessed 2019-08-06]
42. Department of Health and Human Services (US). Confidentiality of Substance Use Disorder Patient Records. Federal Register. URL: https://www.federalregister.gov/documents/2017/01/18/2017-00719/confidentiality-of-substance-use-disorder-patient-records [accessed 2019-09-01]
43. Bresnick J. Five Blockchain Use Cases for Healthcare Payers, Providers. Health IT Analytics. URL: https://healthitanalytics.com/news/five-blockchain-use-cases-for-healthcare-payers-providers [accessed 2019-07-15]
44. Ekblaw A, Azaria A, Halamka J, Lippman A. A Case Study for Blockchain in Healthcare: MedRec prototype for electronic health records and medical research data. 2016. URL: https://www.healthit.gov/sites/default/files/5-56-onc_blockchainchallenge_mitwhitepaper.pdf [accessed 2019-07-15]
45. Agbo C, Mahmoud Q, Eklund J. Blockchain Technology in Healthcare: A Systematic Review. Healthcare (Basel) 2019 Apr 04;7(2):A [FREE Full text] [doi: 10.3390/healthcare7020056] [Medline: 30987333]
46. Dimitrov DV. Blockchain Applications for Healthcare Data Management. Healthc Inform Res 2019 Jan;25(1):51-56 [FREE Full text] [doi: 10.4258/hir.2019.25.1.51] [Medline: 30788182]
47. Vazirani AA, O'Donoghue O, Brindley D, Meinert E. Implementing Blockchains for Efficient Health Care: Systematic Review. J Med Internet Res 2019 Feb 12;21(2):e12439 [FREE Full text] [doi: 10.2196/12439] [Medline: 30747714]
48. Manion S. Advancing Health Research with Blockchain. In: Dhillon V, Bass J, Hooper M, Metcalf D, Cahana A, editors. Blockchain in Healthcare: Innovations That Empower Patients, Connect Professionals and Improve Care. New York: Productivity Press; Jun 30, 2021.
49. Günther V, Chirita A. "Scienceroot" Whitepaper. Scienceroot. 2018 Jun 13. URL: https://www.scienceroot.com/wp-content/uploads/2020/11/whitepaper.pdf [accessed 2019-09-09]
50. Filippova E. Blockchain Solutions for Scientific Publishing. Medium. 2018. URL: https://medium.com/crypto3conomics/blockchain-solutions-for-scientific-publishing-ef4b4e79ae2 [accessed 2019-09-09]
51. Wong DR, Bhattacharya S, Butte AJ. Prototype of running clinical trials in an untrustworthy environment using blockchain. Nat Commun 2019 Feb 22;10(1):917 [FREE Full text] [doi: 10.1038/s41467-019-08874-y] [Medline: 30796226]
52. Lovett L. Nebula Genomics, Longenesis team up to create life data ecosystem, data sharing platforms. Mobihealthnews. 2018 May 16. URL: https://www.mobihealthnews.com/content/nebula-genomics-longenesis-team-create-life-data-ecosystem-data-sharing-platforms [accessed 2021-07-20]

XSL•FO
RenderX

53. Engelhardt MA. Hitching Healthcare to the Chain: An Introduction to Blockchain Technology in the Healthcare Sector. TIM Review 2017 Oct 27;7(10):22-34. [doi: 10.22215/timreview/1111]
54. DSCSA Pilot Project Program. Food and Drug Administration (US). URL: https://www.fda.gov/drugs/drug-supply-chain-security-act-dscsa/dscsa-pilot-project-program [accessed 2020-05-23]
55. Fan K, Wang S, Ren Y, Li H, Yang Y. MedBlock: Efficient and Secure Medical Data Sharing Via Blockchain. J Med Syst 2018 Jun 21;42(8):136. [doi: 10.1007/s10916-018-0993-7] [Medline: 29931655]
56. Novotny M. Blockchain In Clinical Trials: A New Era For Our Data. Clinical Research News. 2018 Sep 19. URL: https://www.clinicalresearchnewsonline.com/news/2018/09/19/blockchain-in-clinical-trials-a-new-era-for-our-data [accessed 2019-08-03]
57. Clauson K, Breeden E, Davidson C, Mackey T. Leveraging Blockchain Technology to Enhance Supply Chain Management in Healthcare: An exploration of challenges and opportunities in the health supply chain. Blockchain in Healthcare Today 2018 Mar 23;1:1-12 [FREE Full text] [doi: 10.30953/bhty.v1.20]
58. Gordon W, Wright A, Landman A. Blockchain in Health Care: Decoding the Hype. NEJM Catalyst. URL: https://catalyst.nejm.org/decoding-blockchain-technology-health/ [accessed 2019-07-17]
59. Kuo T, Kim H, Ohno-Machado L. Blockchain distributed ledger technologies for biomedical and health care applications. J Am Med Inform Assoc 2017 Nov 01;24(6):1211-1220 [FREE Full text] [doi: 10.1093/jamia/ocx068] [Medline: 29016974]
60. Skolnick P, Volkow ND. Re-energizing the Development of Pain Therapeutics in Light of the Opioid Epidemic. Neuron 2016 Oct 19;92(2):294-297 [FREE Full text] [doi: 10.1016/j.neuron.2016.09.051] [Medline: 27764663]
61. Baumann MH, Kopajtic TA, Madras BK. Pharmacological Research as a Key Component in Mitigating the Opioid Overdose Crisis. Trends Pharmacol Sci 2018 Dec;39(12):995-998. [doi: 10.1016/j.tips.2018.09.006] [Medline: 30454770]
62. McGhin T, Choo KR, Liu CZ, He D. Blockchain in healthcare applications: Research challenges and opportunities. Journal of Network and Computer Applications 2019;135:62-75 [FREE Full text] [doi: 10.1016/j.jnca.2019.02.027]
63. Nugent T, Upton D, Cimpoesu M. Improving data transparency in clinical trials using blockchain smart contracts. F1000Res 2016;5:2541 [FREE Full text] [doi: 10.12688/f1000research.9756.1] [Medline: 28357041]
64. Mackey TK, Kuo T, Gummadi B, Clauson KA, Church G, Grishin D, et al. 'Fit-for-purpose?' - challenges and opportunities for applications of blockchain technology in the future of healthcare. BMC Med 2019 Mar 27;17(1):68 [FREE Full text] [doi: 10.1186/s12916-019-1296-7] [Medline: 30914045]
65. Who's sharing their clinical trial results? FDAAA TrialsTracker. URL: http://fdaaa.trialstracker.net [accessed 2019-08-23]
66. Maslove DM, Klein J, Brohman K, Martin P. Using Blockchain Technology to Manage Clinical Trials Data: A Proof-of-Concept Study. JMIR Med Inform 2018 Dec 21;6(4):e11949 [FREE Full text] [doi: 10.2196/11949] [Medline: 30578196]
67. Shrier AA, Chang A, Diakun-thibault N, Forni L, Landa F, Mayo J, et al. Blockchain and Health IT: Algorithms, Privacy, and Data (White Paper). 2016 Aug 08. URL: https://www.healthit.gov/sites/default/files/1-78-blockchainandhealthitalgorithmsprivacydata_whitepaper.pdf.2016 [accessed 2019-07-13]
68. Ganti L. Why blockchain is a dream come true for pharma researchers. Pharmaphorum. 2018 Jul 17. URL: https://pharmaphorum.com/views-and-analysis/why-blockchain-is-a-dream-come-true-for-pharma-researchers/ [accessed 2019-08-21]
69. Gruber B. Scientist.com using blockchain tech to ensure data integrity. Outsourcing-Pharma.com. 2018 May 03. URL: https://www.outsourcing-pharma.com/Article/2018/05/03/Scientist.com-using-blockchain-tech-to-ensure-data-integrity [accessed 2019-07-29]
70. Baara M, Lipset C, Kudumala A, Fox J, Israel A. Blockchain opportunities for patient data donation and clinical research. Deloitte. URL: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/process-and-operations/us-cons-blockchain-opportunities-patient-data-donation-clinical-research.pdf.2018 [accessed 2019-07-03]
71. Mamoshina P, Ojomoko L, Yanovich Y, Ostrovski A, Botezatu A, Prikhodko P, et al. Converging blockchain and next-generation artificial intelligence technologies to decentralize and accelerate biomedical research and healthcare. Oncotarget 2018 Jan 19;9(5):5665-5690 [FREE Full text] [doi: 10.18632/oncotarget.22345] [Medline: 29464026]
72. Ozercan HI, Ileri AM, Ayday E, Alkan C. Realizing the potential of blockchain technologies in genomics. Genome Res 2018 Sep;28(9):1255-1263 [FREE Full text] [doi: 10.1101/gr.207464.116] [Medline: 30076130]
73. Shabani M. Blockchain-based platforms for genomic data sharing: a de-centralized approach in response to the governance problems? J Am Med Inform Assoc 2019 Jan 01;26(1):76-80 [FREE Full text] [doi: 10.1093/jamia/ocy149] [Medline: 30496430]
74. Encrypgen. Gene-Chain: The DNA Data Marketplace. URL: https://encrypgen.com/ [accessed 2019-07-10]
75. Luna DNA - Researchers. LunaDNA. URL: https://www.lunadna.com/researchers/ [accessed 2019-07-19]
76. Singer N. Can Sweatcoin, a Hot Fitness App, Keep You Off the Couch? New York Times. 2018 Jan 07. URL: https://www.nytimes.com/2018/01/07/technology/sweatcoin-fitness-app.html [accessed 2019-07-15]
77. Clinicoin. URL: https://clinicoin.io/en [accessed 2019-07-15]
78. Healthereum. URL: https://healthereum.com/ [accessed 2019-07-15]
79. Travers J. Blockchain and Cancer: How This Tech Is Changing Research and Treatment. Labroots. 2019 Jun 10. URL: https://www.labroots.com/trending/cancer/14933/blockchain-cancer-tech-changing-research-treatment [accessed 2019-08-16]

XSL•FO
RenderX

80. Fletcher JB, Shoptaw S, Peck JA, Reback CJ. Contingency Management Reduces Symptoms of Psychological and Emotional Distress among Homeless, Substance-dependent Men Who Have Sex with Men. Ment Health Subst Use 2014 Nov 01;7(4):420-430 [FREE Full text] [doi: 10.1080/17523281.2014.892897] [Medline: 25364379]
81. Kaminer Y, Burleson JA, Burke R, Litt MD. The efficacy of contingency management for adolescent cannabis use disorder: a controlled study. Subst Abus 2014;35(4):391-398. [doi: 10.1080/08897077.2014.933724] [Medline: 25010430]
82. Mackillop J, Murphy CM, Martin RA, Stojek M, Tidey JW, Colby SM, et al. Predictive Validity of a Cigarette Purchase Task in a Randomized Controlled Trial of Contingent Vouchers for Smoking in Individuals With Substance Use Disorders. Nicotine Tob Res 2016 May;18(5):531-537 [FREE Full text] [doi: 10.1093/ntr/ntv233] [Medline: 26498173]
83. Linn L, Koo M. Blockchain for health data and its potential use in health IT and health care related research. ONC/NIST Use of Blockchain for Healthcare and Research Workshop. 2016. URL: https://www.healthit.gov/sites/default/files/11-74-ablockchainforhealthcare.pdf [accessed 2019-06-03]
84. Engelhardt MA. Hitching Healthcare to the Chain: An Introduction to Blockchain Technology in the Healthcare Sector. Technology Innovation Mananagement Review 2017 Oct 27;7(10):22-34. [doi: 10.22215/timreview/1111]
85. Checking the PDMP: An Important Step to Improving Opioid Prescribing Practices. Centers for Disease Control and Prevention (US). URL: https://www.cdc.gov/drugoverdose/pdf/pdmp_factsheet-a.pdf [accessed 2020-10-10]
86. Raghavendra M. Can Blockchain technologies help tackle the opioid epidemic: A Narrative Review. Pain Med 2019 Oct 01;20(10):1884-1889. [doi: 10.1093/pm/pny315] [Medline: 30848821]
87. Scott T, Post A, Quick J, Rafiqi S. Evaluating Feasibility of Blockchain Application for DSCSA Compliance. SMU Data Science Review 2018;1:1-25 [FREE Full text]
88. Miliard M. Blockchain being put to work by IBM, Intel, CDC to combat opioid epidemic. Healthcare IT News. URL: https://www.healthcareitnews.com/news/blockchain-being-put-work-ibm-intel-cdc-combat-opioid-epidemic [accessed 2019-08-15]
89. Clauson KA, Breeden EA, Davidson C, Mackey TK. Leveraging Blockchain Technology to Enhance Supply Chain Management in Healthcare:. Blockchain in Healthcare Today 2018 Mar 23:1-12. [doi: 10.30953/bhty.v1.20]
90. Li P, Nelson S, Malin B, Chen Y. DMMS: A Decentralized Blockchain Ledger for the Management of Medication Histories. Blockchain in Healthcare Today 2019;2:1-5 [FREE Full text] [doi: 10.30953/bhty.v2.38]
91. Evans J. Improving the Transparency of the Pharmaceutical Supply Chain through the Adoption of Quick Response (QR) Code, Internet of Things (IoT), and Blockchain Technology: One Result: Ending the Opioid Crisis. tlp 2019 Mar 07;19(1):593-600. [doi: 10.5195/tlp.2019.227] [Medline: 5637758]
92. Sylim P, Liu F, Marcelo A, Fontelo P. Blockchain Technology for Detecting Falsified and Substandard Drugs in Distribution: Pharmaceutical Supply Chain Intervention. JMIR Res Protoc 2018 Sep 13;7(9):e10163 [FREE Full text] [doi: 10.2196/10163] [Medline: 30213780]
93. BlockVerify. URL: http://www.blockverify.io/ [accessed 2019-07-05]
94. Mediledger. URL: https://www.mediledger.com/ [accessed 2019-07-19]
95. Benniche S. Using blockchain technology to recycle cancer drugs. Lancet Oncol 2019 Jun;20(6):e300. [doi: 10.1016/S1470-2045(19)30291-8] [Medline: 31085048]
96. Monegain B. IBM Watson, FDA align to boost public health with blockchain. Healthcare IT News. URL: https://www.healthcareitnews.com/news/ibm-watson-fda-align-boost-public-health-blockchain [accessed 2019-08-28]
97. Dolatshahi J, Maldjian L, Welch A, Fulmer C, Winkelstein E. Tracking Community Naloxone Dispensing: Part of a Strategy to Reduce Overdose Deaths. Online Journal of Public Health Informatics 2019;11:e445. [doi: 10.5210/ojphi.v11i1.9932]
98. Kim D, Irwin K, Khoshnood K. Expanded access to naloxone: options for critical response to the epidemic of opioid overdose mortality. Am J Public Health 2009 Mar;99(3):402-407. [doi: 10.2105/AJPH.2008.136937] [Medline: 19150908]
99. Rees D, Sabia J, Argys L, Dave D, Latshaw J. With a Little Help from My Friends: The Effects of Naloxone Access and Good Samaritan Laws on Opioid-Related Deaths. National Bureau of Economic Research 2019;62(1):1-27 [FREE Full text] [doi: 10.1086/700703]
100. Abouk R, Pacula RL, Powell D. Association Between State Laws Facilitating Pharmacy Distribution of Naloxone and Risk of Fatal Overdose. JAMA Intern Med 2019 Jun 01;179(6):805-811 [FREE Full text] [doi: 10.1001/jamainternmed.2019.0272] [Medline: 31058922]
101. McClellan C, Lambdin BH, Ali MM, Mutter R, Davis CS, Wheeler E, et al. Opioid-overdose laws association with opioid use and overdose mortality. Addict Behav 2018 Nov;86:90-95. [doi: 10.1016/j.addbeh.2018.03.014] [Medline: 29610001]
102. Wheeler E, Jones TS, Gilbert MK, Davidson PJ, Centers for Disease Control Prevention (CDC). Opioid Overdose Prevention Programs Providing Naloxone to Laypersons - United States, 2014. MMWR Morb Mortal Wkly Rep 2015 Jun 19;64(23):631-635 [FREE Full text] [Medline: 26086633]
103. Frank R, Humphreys K, Pollack H. Does Naloxone Availability Increase Opioid Abuse? The Case For Skepticism. HealthAffairs. 2018. URL: https://www.healthaffairs.org/do/10.1377/hblog20180316.599095/full/ [accessed 2019-08-18]
104. Bean R. Will Blockchain Transform Healthcare? Forbes. 2018 Aug 05. URL: https://www.forbes.com/sites/ciocentral/2018/08/05/will-blockchain-transform-healthcare/?sh=23cf7247553d [accessed 2019-07-13]
105. Department of Health and Human Services (US), Office of the Surgeon General. Facing Addiction in America: The Surgeon General's Report on Alcohol, Drugs, and Health. Washington, DC: HHS; 2016. URL: https://www.ncbi.nlm.nih.gov/books/n/surgaddict/pdf/ [accessed 2019-10-15]

XSL•FO
RenderX

106. Mannatt R, Dworkowitz A. Overcoming Data-Sharing Challenges in the Opioid Epidemic: Integrating Substance Use Disorder Treatment in Primary Care. California Health Care Foundation. 2018. URL: https://www.chcf.org/wp-content/uploads/2018/07/OvercomingDataSharingChallengesOpioid.pdf [accessed 2019-08-28]
107. Nichol P, Dailey W. Micro-Identities Improve Healthcare Interoperability with Blockchain: Deterministic Methods for Connecting Patient Data to Uniform Patient Identifiers. ResearchGate. 2016. URL: https://tinyurl.com/npxxrh45 [accessed 2019-08-18]
108. Sweeney E. CDC eyes blockchain for public health surveillance. Fierce Healthcare. 2017. URL: https://www.fiercehealthcare.com/mobile/cdc-blockchain-public-health-surveillance-data-sharing [accessed 2019-08-20]
109. Ocrcutt M. Why the CDC Wants in on Blockchain. MIT Technology Review. 2017. URL: https://www.technologyreview.com/s/608959/why-the-cdc-wants-in-on-blockchain/ [accessed 2019-08-23]
110. O'Donoghue O, Vazirani AA, Brindley D, Meinert E. Design Choices and Trade-Offs in Health Care Blockchain Implementations: Systematic Review. J Med Internet Res 2019 May 10;21(5):e12426 [FREE Full text] [doi: 10.2196/12426] [Medline: 31094344]
111. Gordon W, Catalini C. Blockchain Technology for Healthcare: Facilitating the Transition to Patient-Driven Interoperability. Comput Struct Biotechnol J 2018;16:224-230 [FREE Full text] [doi: 10.1016/j.csbj.2018.06.003] [Medline: 30069284]
112. Blockchain Performance, Throughput and Scalability. HIMMS. URL: https://www.himss.org/library/blockchain-performance-throughput-and-scalability [accessed 2019-09-05]
113. Zhang P, Schmidt D, White J, Lenz G. Chapter One - Blockchain Technology Use Cases in Healthcare. In: Raj P, Deka GC, editors. Advances in Computers. Amsterdam: Elsevier; 2018:1-41.
114. 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 2019-09-03]
115. Millard C. Blockchain and law: Incompatible codes? Computer Law & Security Review 2018;34:843-846 [FREE Full text] [doi: 10.1016/j.clsr.2018.06.006]
116. Trust Framework and Code of Conduct - The CARIN Alliance Code of Conduct. CARIN Alliance. URL: https://www.carinalliance.com/our-work/trust-framework-and-code-of-conduct/ [accessed 2019-09-05]
117. Funk E, Riddell J, Ankel F, Cabrera D. Blockchain Technology: A Data Framework to Improve Validity, Trust, and Accountability of Information Exchange in Health Professions Education. Acad Med 2018 Dec;93(12):1791-1794. [doi: 10.1097/ACM.0000000000002326] [Medline: 29901658]
118. Morton H. Blockchain 2019 Legislation. National Conference of State Legislatures. 2019. URL: https://www.ncsl.org/research/financial-services-and-commerce/blockchain-2019-legislation.aspx2019 [accessed 2020-11-20]
119. Insider Intelligence. Insider. 2021. URL: https://www.businessinsider.com/blockchain-cryptocurrency-regulations-us-global [accessed 2021-03-01]

Abbreviations

CFR: Code of Federal Regulation
EHR: electronic health record
IT: information technology
PDMP: prescription drug monitoring program
SUD: substance use disorder
TEP: technical expert panel

Edited by C Lovis; submitted 10.06.20; peer-reviewed by J Hoppe, C McClellan, TT Kuo; comments to author 07.07.20; revised version received 11.03.21; accepted 26.03.21; published 27.08.21.

Please cite as:
Gonzales A, Smith SR, Dullabh P, Hovey L, Heaney-Huls K, Robichaud M, Boodoo R
Potential Uses of Blockchain Technology for Outcomes Research on Opioids
JMIR Med Inform 2021;9(8):e16293
URL: https://medinform.jmir.org/2021/8/e16293
doi:10.2196/16293
PMID:34448721

©Aldren Gonzales, Scott R Smith, Prashila Dullabh, Lauren Hovey, Krysta Heaney-Huls, Meagan Robichaud, Roger Boodoo. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 27.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR

XSL•FO
RenderX

Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

A System to Support Diverse Social Program Management

Mollie McKillop[1], MPH, PhD; Jane Snowdon[1], PhD; Van C Willis[1], PhD; Shira Alevy[1], EdM; Rubina Rizvi[1], MD, PhD; Karen Rewalt[1], MBA; Charlyne Lefebvre-Paillé[1], MA; William Kassler[1], MPH, MD; Gretchen Purcell Jackson[1,2], MD, PhD

[1]IBM Watson Health, Cambridge, MA, United States

[2]Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
Mollie McKillop, MPH, PhD
IBM Watson Health
75 Binney Street
Cambridge, MA, 02142
United States
Phone: 1 3322073519
Email: mollie.mckillop@ibm.com

Abstract

Background: Social programs are services provided by governments, nonprofits, and other organizations to help improve the health and well-being of individuals, families, and communities. Social programs aim to deliver services effectively and efficiently, but they are challenged by information silos, limited resources, and the need to deliver frequently changing mandated benefits.

Objective: We aim to explore how an information system designed for social programs helps deliver services effectively and efficiently across diverse programs.

Methods: This viewpoint describes the configurable and modular architecture of Social Program Management (SPM), a system to support efficient and effective delivery of services through a wide range of social programs and lessons learned from implementing SPM across diverse settings. We explored usage data to inform the engagement and impact of SPM on the efficient and effective delivery of services.

Results: The features and functionalities of SPM seem to support the goals of social programs. We found that SPM provides fundamental management processes and configurable program-specific components to support social program administration; has been used by more than 280,000 caseworkers serving more than 30 million people in 13 countries; contains features designed to meet specific user requirements; supports secure information sharing and collaboration through data standardization and aggregation; and offers configurability and flexibility, which are important for digital transformation and organizational change.

Conclusions: SPM is a user-centered, configurable, and flexible system for managing social program workflows.

(JMIR Med Inform 2021;9(8):e23219) doi:10.2196/23219

KEYWORDS

other clinical informatics applications; process management tools; requirements analysis and design; consumer health informatics; public health

Introduction

Government and community-based organizations are responsible for delivering social services to clients through social programs provided at the national, state, county, and city levels. Social programs are critical to the health and welfare of many citizens, as they provide a wide range of benefits [1], such as (1) health and human services for health insurance, prevention services, child welfare, and nutrition assistance; (2) workforce services for unemployment insurance programs and job training; and (3) social security programs offering income support and benefits [2].

In contrast to commercial entities, government agencies administering social programs face unique challenges regarding service delivery and their operational processes, including (1) information silos that limit decision-making abilities, (2) requirements to balance privacy with data sharing and transparent use of public funds, (3) reduced financial resources but growing demand for services, (4) legislative and organizational influences on eligibility and entitlement, and (5)

XSL•FO
RenderX

the need to document the delivery of mandated services for beneficiaries across multiple categorical programs [3,4]. These challenges are exacerbated by societal shifts, such as increasing income inequality, aging populations, unemployment, ongoing changes in government policies, and resource constraints [5-7].

Opportunities exist for information technology to address these challenges by improving the efficiency and transparency of social program workflows and enabling collaboration among stakeholders. These goals can be achieved by centralizing critical data, streamlining eligibility determination and case management, and improving communication within and across organizations [8]. Previous preliminary research indicates that the use of information systems to support high-quality and efficient service delivery is promising but limited in scope and does not meet the unique needs of social programs [8-10]. Comprehensive systems specifically designed for social programs have not been previously described in the literature.

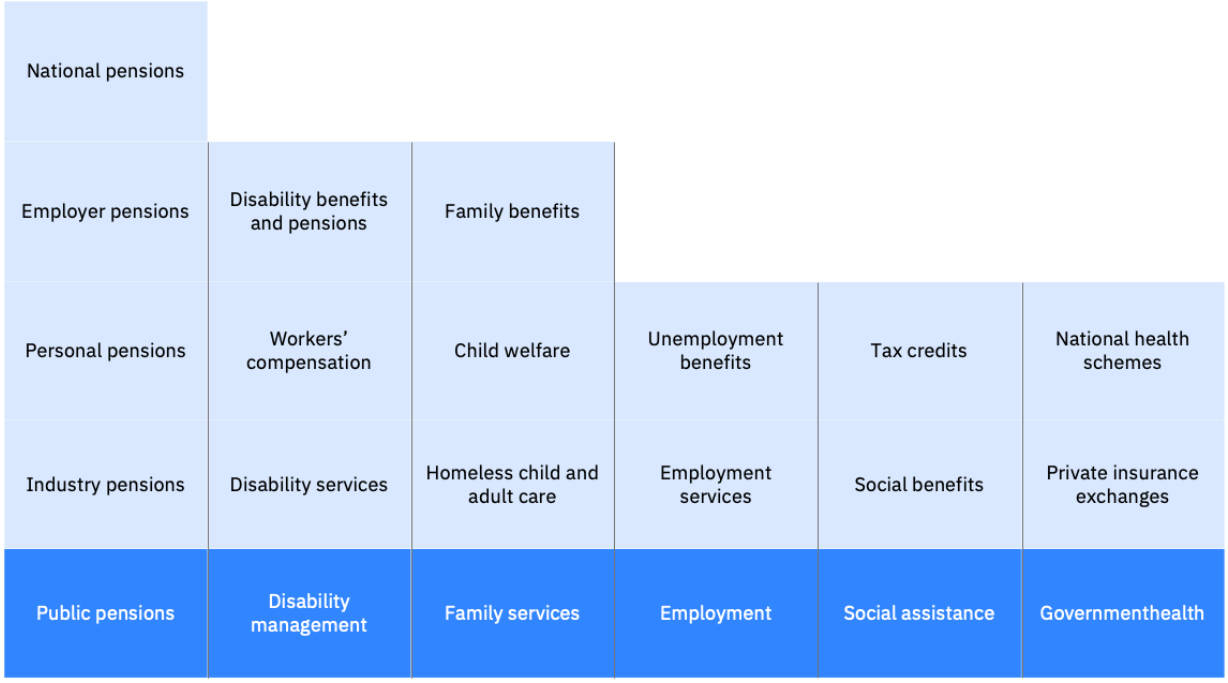
This viewpoint describes the design, functionality, and selected applications of a software solution for social services, which combines domain-specific business processes with a flexible open architecture to allow for needed configurability while standardizing data elements. Specific design principles that aim to address the unique challenges encountered by social programs, along with examples of implementation and usage, are described.

Our system was designed and developed by subject matter experts in social programs, including industry specialists, service professionals experienced in social programs, and social program product developers. The system was developed for social service and human service agencies to advance digital transformation, intending to support a wide range of constituents across the health and human services enterprise. Potential users include, but are not limited to, Medicaid program managers, directors, analysts, caseworkers and care managers, clients, and beneficiaries. The system aims to prioritize the needs of users and beneficiaries to unify multidisciplinary teams. User-centered research methods, including user shadowing, interviews, surveys, and scenario testing, were used to identify user needs, and human-centered design leveraging iterative co-development and prototyping were used to develop the system.

System Scope

Social Program Management (SPM) supports two basic types of social programs: (1) programs in which eligibility is determined primarily based on need and (2) programs where eligibility for benefits and services is determined based on previous contributions [11,12]. SPM also supports care and protection programs, such as child welfare programs, where eligibility is based on practice models and assessments; Figure 1 describes the full scope and scale of service organizations that SPM supports.

Figure 1. Scope of services supported by Social Program Management. Social Program Management serves health and social service organizations at all levels of government and nonprofit organizations.



These diverse social programs share similar goals and challenges with the aim of improving service delivery to and outcomes for beneficiaries. To meet these goals, social programs require systems that (1) support complex eligibility and entitlement, (2) provide beneficiaries with easy access to services, (3) enable efficient management of high case volumes, (4) provide decision support and knowledge management tools, (5) allow flexibility for changes in policies and processes, and (6) reduce the potential for fraud and abuse. These needs must be met throughout triage, initial contract and registration, determination of eligibility for benefits, service planning and delivery, and outcome evaluation. Delivering services across these program stages relies on secure data gathering, documentation, retrieval, validation, auditing, and analysis. SPM has features and capabilities that address each of these needs (Table 1) [13].

XSL•FO
RenderX

Table 1. Features and capabilities to address needs.

Challenge and need	Features and functionalities
Information and programs are managed in silos	
Integrate service delivery	• Centralized repository for the integrated management of cases maintained in the system and external systems • Multidisciplinary team portal provides a means for cross-agency and cross-program teams to collaborate on cases • Indexing on cases and participants provides master data management, which allows for identifying individuals across systems, especially web services • Citizen portal and multichannel access allows clients to check on status, submit applications, and manage benefits on agency websites
Consolidate infrastructure	• The system data model supports needs-based and contribution-based programs • Program-specific modules for social security, health and human services, and workforce services based on a common data model
Provide standards-based integration	• The system data model can be deployed as web services • Preconfigured adapters and enterprise application integration connectors to facilitate integration with existing systems
Data protection and privacy	
Prevent unauthorized access to sensitive data	• Role-based and data field level security for personal and case data
Protect the integrity and privacy of personal data while sharing data responsibly	• Configurable security levels to prevent unauthorized access to data while allowing multiple stakeholders to view a client's data with the appropriate level of access • Auditing and tracking of transactions involving sensitive data • Auditing and traceability for logging time, date, and the user responsible for any read, update, and delete actions for any specified participant or case data elements
Reduce inaccurate or duplicate data	• Integrated case management module supports centralized or distributed maintenance and sharing of case and participant data across programs
Shrinking budgets	
Provide standard programs and eligibility and entitlement rules with opportunities for customization	• Eligibility and entitlement rules for income support and assessments for child welfare programs • Configurable, packaged connectors and adapters for integration to existing and service-oriented applications
Leverage and consolidate cost-effective, existing infrastructure to deploy new program solutions	• Support for open standards and de facto standards to ensure deployment on the widest possible range of operating systems, hardware platforms, and middleware
Provide tools to allow incremental approaches to implementation	• Web services and configurable business processes allow for maximum flexibility in deployment and implementation options
Increasing demand for services	
Deliver high performance and scalability	• Eligibility and entitlement engine optimized for processing and calculating high volumes of rules-based assessments and complex reassessments where information may need to be retroactively changed for large populations of clients • Supervisor workspace for real-time analysis and dynamic allocation of workloads across a department or agency
Offer broad access and reliability	• Multichannel access through a device-independent web-based user interface • Configurable citizen portal for access to cross-program screening and eligibility
Legislative and organizational change	
Supply configurable systems that can adapt to legislation without reprogramming and support different organizational structures	• Configurable eligibility and entitlement engine supporting complex reassessments • Configurable regional administration segregates duties by location or organization • Concurrent execution of reassessment batch jobs with ongoing web-based transactions
Balancing accuracy, consistency, and outcome focus	

XSL•FO
RenderX

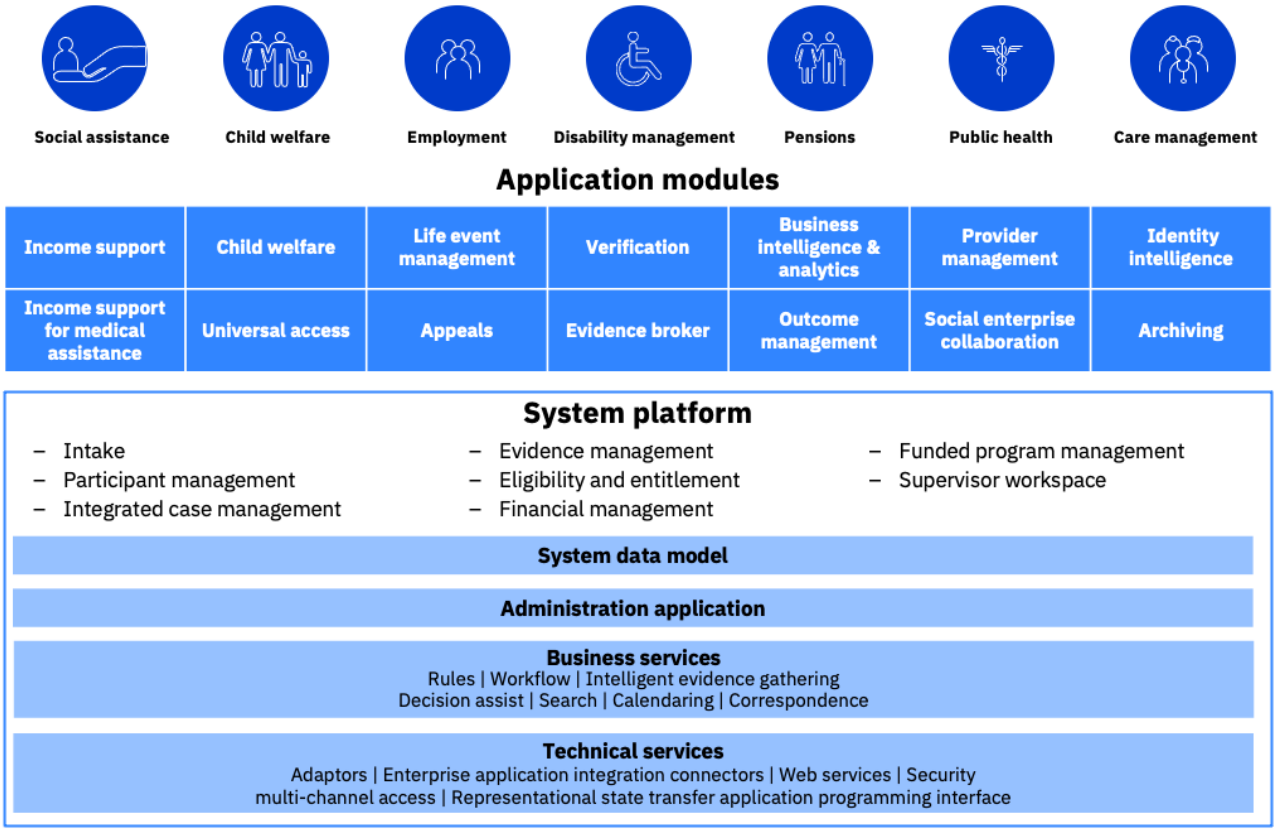
Challenge and need	Features and functionalities
Comply with legislation in the delivery of benefits across large populations while also individualizing services for clients and families	• Integrated service-planning templates based on best practices, such as structured decision-making assessments, which are developed by the National Council on Crime and Delinquency • Intelligent evidence gathering module scripts and templates for consistent, structured capture of information for caseworkers and beneficiaries • Decision assist module, a configurable rules-based matrix designed to assure consistency and accuracy in rendering decisions • The social enterprise collaboration module provides a common platform and set of tools for multidisciplinary collaboration in social program organizations

System Architecture

Overview

The features and functionalities described in Table 1 are organized within a single user-centered system comprising modules, the SPM data model, administration application, and business and technical services. The architecture for SPM version 7.0.9, the latest version, is outlined in Figure 2, and the modules and applications of the SPM are described in the following sections (see Multimedia Appendix 1 for technical aspects of SPM).

Figure 2. Social Program Management design. Social Program Management provides core processing and infrastructure components for social program management through its platform. Business services provide support for management needs common across all types of organizations. Modules complement the system platform and provide additional functionality and system configurability.



SPM Modules

The SPM modules presented in Table 2 support repeatable processes that are common across programs. These processes include managing client information and data regarding benefits, automatic assessment of eligibility and entitlement, management of tasks, communication, and scheduling. Each module is supported by the SPM platform, including the data model and administrative, business, and technical services. The SPM modularization supports the incremental modernization of systems used by social programs. As the data model is application-agnostic, the integration of future modules or functionalities is supported. These modules support a comprehensive range of functionalities required for service delivery in social programs. Modules are divided by those that exist for all implementations of SPM (enterprise-wide applications) and those specific to a particular implementation and user agreement (implementation-specific modules).

XSL•FO
RenderX

Table 2. Social Program Management modules and applications.

Name	Type	Description
Intake	Enterprise-wide	The intake module facilitates integration with external systems that manage the client. The intake module notifies an external system when an intake is approved. The external system can then retrieve the details of the intake for further processing.
Participant management	Enterprise-wide	A participant is the system term for any individual or organization about which a social enterprise wants to record information. Participant management provides for the creation and maintenance of all relevant basic information such as contact details, addresses, communications, demographic information, and alternate names and identifiers.
Integrated case management	Enterprise-wide	Integrated case management provides functionality to facilitate the creation, management, and tracking of cases in support of social program service delivery. Interactions between participants and the agency and any associated program or service delivery are recorded. These interactions include assessment, eligibility determination, case approval, program delivery, outcome evaluation, and closure.
Evidence management	Enterprise-wide	Evidence is any data collected in support of a case. In general, such information is program-specific, although various types of evidence may be shared through a number of programs. Typically, the primary use of evidence is for the determination of program eligibility and entitlement. Evidence management provides capabilities to standardize and simplify the process of defining, creating, and maintaining such program-specific, temporal data.
Eligibility and entitlement	Enterprise-wide	The rules for determining both eligibility and entitlement are typically dictated by a combination of legislation, policy, and operating procedure. For many programs, such rules are determined by federal, state, or local governments. These rules often change. The system provides two mechanisms to help social enterprises deal with the problem of changing evidence: • The ability to detect the change and the ability to initiate a reassessment where required • Overpayment and underpayment processing where the system compares the old and new situations automatically detects any overpayments or underpayments and initiates appropriate action
Financial management	Enterprise-wide	Financial management manages and tracks the financial transactions associated with program delivery, including benefit payments and liability recovery. Financial management generates, manages, and tracks the financial transactions associated with cases and participants. It also supports the issue of payments and the creation of liabilities as determined by assessment and entitlement processing.
Funded program management	Enterprise-wide	Funded program management is used to manage funds that can be obligated to clients in need of assistance or to provide payment to providers for services.
Supervisor workspace	Enterprise-wide	Supervisor workspace provides dashboard-style views of a social program's workload. It allows managers to monitor workloads through supervisor dashboard views of staff assignments, real-time display and access to information, centralized management of cases and tasks, prioritization, and allocation of workloads.
Income support	Implementation-specific modules	Income support delivers health and social program components, business processes, toolsets, and interfaces on a dynamically configurable architecture that allows an administrator to change rules without writing code. Income support is designed for programs that provide food, cash, and medical assistance.
Child welfare	Implementation-specific modules	Child welfare provides case management tools that support agencies that work to safeguard children, promote well-being, and support child permanency. Child services facilitate intake, ongoing case management, child abuse investigations, removal of children from unsafe situations, and the adoption of children.
Life event management	Implementation-specific modules	Life event management helps the caseworker to collect evidence and provide guidance that is based on a client's life event, such as the birth of a child, marriage, divorce, or change in employment.
Verification	Implementation-specific modules	The verification engine streamlines the process of verifying evidence that is used in determining eligibility and entitlement as part of program delivery. It provides the functions that are needed for efficient management of verifications where policy or legislation mandates that evidence is verified as a prerequisite for eligibility.
Business intelligence and analytics	Implementation-specific modules	Business intelligence and analytics is a decision support solution that helps social program organizations analyze the effectiveness of their programs and gain insight into the efficiency of their operations. It is scalable from the program to enterprise level. It consists of embedded analytics, domain-specific dashboards, extract, transform, and load functions, and tool-independent, predefined, domain-specific (only for social program management) data marts.

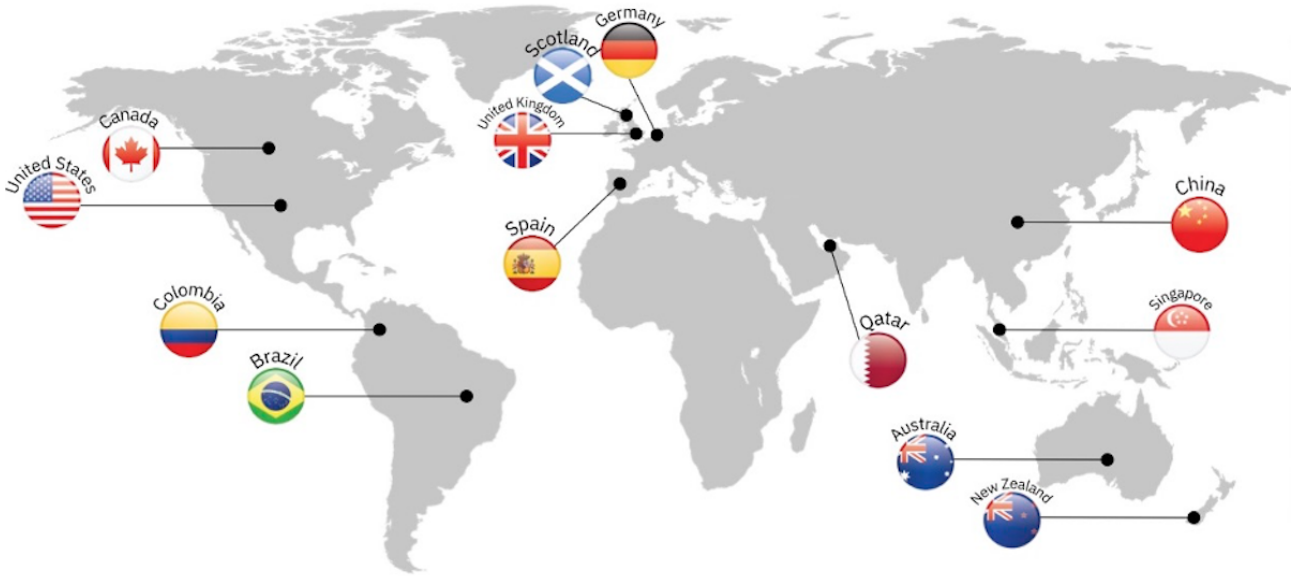
XSL•FO
RenderX

Name	Type	Description
Provider management	Implementation-specific modules	Provider management manages the interactions between the agency and its outside providers, such as foster families, housing facilities, and other vendors.
Identity intelligence	Implementation-specific modules	Identity intelligence aims to give caseworkers the confidence that an applicant is who they say they are and are not duplicated within the system, which might result in duplicate benefits. The product solves this problem by applying analytics to client data and verifies information with limited worker involvement.
Income support for medical assistance	Implementation-specific modules	Income support for medical assistance is specifically built to provide business tools and processes for the management of traditional medical assistance programs, plus the Affordable Care Act and modified adjusted gross income-based Medicaid programs.
Universal access	Implementation-specific modules	Universal access is a fully configurable web-based citizen-facing application that enables agencies to offer a web self-service solution to their clients. Universal access can provide a greater number of clients with access to programs and services by allowing clients to complete key tasks on the web without the assistance of a worker.
Appeals	Implementation-specific modules	Appeals is an automated solution that provides support for the appeals and fair hearings process. Appeals automates the intake, hearings, and decision processes and manages participants in the appeals process. Appeals supports multilevel appeals in which multiple issues for an appellant and respondent can be viewed at a single appeal hearing.
Evidence broker	Implementation-specific modules	Evidence broker facilitates flexible data sharing between different case types and between agencies.
Outcome management	Implementation-specific modules	Outcome management provides organizations that deliver social programs with a framework and automated tools to create and manage outcome plans for clients and their families. Outcome management is designed to help organizations assess needs, establish goals, plan for goal attainment, and track progress.
Social enterprise collaboration	Implementation-specific modules	Social enterprise collaboration is a common platform and set of tools for multidisciplinary collaboration in social programs. Multidisciplinary teams are involved in supporting the needs of clients and families, including other agencies, local providers, and interested community partners.
Archiving	Implementation-specific modules	As database size grows, performance can degrade rapidly. A large percentage of data in a social program database is unlikely to be accessed daily. Performance can be greatly improved if infrequently accessed data are removed from the production environment. Archiving stores and maintains inactive data in a repository so that it can be retrieved when necessary.

System Usage and Case Reports

The latest version of SPM has been used by more than 50 programs in 13 countries (Figure 3) and 13 languages. In total, more than 280,000 caseworkers worldwide have used SPM with 30 million beneficiaries.

Figure 3. Worldwide usage of Social Program Management.



XSL•FO
RenderX

Case Report 1: User-Centered Design for Social Program Efficiency

Eligibility and entitlement determination are often complex, time-consuming, and one of the most frustrating aspects of social program service delivery for administrators and clients. A large US city department of human resources used SPM's universal access module to redesign their Supplemental Nutrition Assistance Program (SNAP) [14] to digitize services and reduce wait times compared with manual processing of applications and limit the need for in-person visits.

In April 2018, the city implemented a citizen-facing portal, available in 7 languages, to deliver the SNAP and other benefits via clients' own desktop and mobile phones. The design of the portal was informed by 30 shadowing sessions, 10 interviews, and 20 scenario testing sessions. Iterative prototyping and co-design workshops were also conducted to refine the portal design.

A retrospective pre-post comparison of SPM metrics (ie, web and mobile log-ins, web-based applications and recertifications, and calls to update profile information) was conducted using monthly data collected in April 2018 and April 2019 (Table 3). The median age of the SNAP recipients was 25 to 44 years, and most were women (870,000/1,523,502, 57.11%). With regard to ethnic background, 27.83% (424,050/1,523,502) were Black only, 24.43% (372,245/1,523,502) were Hispanic and White only, 10.35% (157,723/1,523,502) were Black and Hispanic only, 11.23% (171,157/1,523,502) were multiethnic, 14.6% (222,495/1,523,502) were White, 10.29% (156,813/1,523,502) were Asian, and 1.25% (19,019/1,523,502) were of other ethnicities.

Table 3. Social Program Management universal access pre- and postimplementation system metrics.

Metric	Predeployment (April 2018)	Postdeployment (April 2019)	Percentage change (%)
System log-ins (web and mobile)	915,532	1,684,248	+83.96
Applications and recertifications received	33,421	40,198	+20.28
Profile update calls[a]	9696	17,547	+80.97

[a]Profile update calls to center staff are required to update profile information.

In the assessed periods, application rejections because of failure to provide documentation were reduced by 20% from 2674 in April 2018 to 2139 in April 2019 and center visits were reduced by 37% from 71,116 in April 2018 to 44,803 in April 2019. Of the 30,000 SNAP applications submitted in August 2019, 80% were submitted on the web through client-mobile devices via SPM's universal access module. Client experience satisfaction was also measured with a web-based 5-star rating survey emailed to participants after they completed the application process. Responses (27,128) were collected with an average rating of 4.31 out of 5 (5 being the highest; 1 being the lowest) for the SNAP application and 4.44 out of 5 for the SNAP recertification.

Case Report 2: Flexibility and Standardization in Digital Transformation

A US state's health and human services department designed a program to improve the way state county departments provided services to families and allow caseworkers to spend less time on administrative tasks and more time helping individuals and families. Specifically, the state wanted real-time data sharing and aggregation across different health and human services divisions. At the time, most families were served through multiple categorical programs. Concurrently, the state wanted to limit the amount of system customization for each county yet be flexible enough to allow each to use legacy systems in a consolidated system managing all benefits and services until legacy systems could be sunset.

SPM was implemented in 2012, and information technology systems were modernized in more than 100 counties for 8 years. Deployments of SPM were designed to be interoperable with legacy systems so that incremental rollout could occur [15]. SPM has replaced or is in the process of replacing approximately 20 legacy systems with a single one. SPM's common data model provides data sharing; therefore, caseworkers no longer need to enter data into multiple systems, spending less time on administrative tasks and more time assisting families. Participants' administrative data can be viewed and shared in real time to support a holistic view of the client, their needs, and the analysis of progress toward goals and programmatic outcomes. Currently, at least 3.5 million individuals across the state have been provided benefits through SPM [16].

Case Report 3: Data Aggregation, Sharing, and Collaboration

In a large city in Germany, 7 local districts and 40 regional agencies are responsible for protecting children from abuse. Two special government organization units support these districts and agencies in defining and monitoring policies and providing financial and technical resources. In an effort to provide better outcomes and serve clients more efficiently, the government sought to use SPM to improve processes for client intake, case management, and communication between agencies. Specific needs included improving upon reports of abuse, traditionally done through telephone and fax machines, and better data sharing and collaboration. The government leveraged the SPM platform to address these needs, including applications for verification of evidence and provider management for managing interactions between the government and local agencies, such as foster care. Implementation-specific modules included the social enterprise collaboration for supporting multidisciplinary teams and the child welfare module for managing child abuse cases. The solution was implemented in stages where standards for social services relevant to child abuse cases, such as foster care, were defined; an interface among SPM, legacy systems, and systems of local agencies was then deployed. In the second stage, modules for managing clients

XSL•FO
RenderX

were implemented. These modules provided local agencies and the government with case management capabilities for documenting and sharing information related to interactions with children and families, services, contacts, worker visits, and judicial processes.

SPM allows caseworkers to view and manage a wide range of information in one place, from the initial receipt of an allegation through the final case outcome. This information enables bidirectional information sharing among local agencies and can then be used to document outcomes digitally. For example, when an allegation is made, supporting information such as police reports is automatically imported into SPM. As of June 2019, SPM in this city had 1300 users, 370,000 clients, and more than 200,000 cases and processed approximately 50,000 transactions per month since its implementation in 2014. This represents an average increase of 60% in the number of cases processed per year compared with the legacy system.

Discussion

Overview

We presented a configurable and modular system that delivers integrated cross-program and cross-agency solutions for needs-based and contribution-based social service programs [3,17-21]. Although social programs provide a wide variety of benefits across distinct and heterogeneous populations, fundamental needs are shared across programs that deliver services. SPM provides a set of functionalities that are generalizable and support government agencies and their beneficiaries across diverse program types and locations. These functionalities are configurable and adaptable to address the unique context of each social program.

The flexibility of SPM provides advantages in implementation and change management for the digital transformation of social programs. Most legacy social program systems have automated key program processes. Usually developed ad hoc, these older systems are complex, heterogeneous, and high maintenance. At the same time, these organizations tend to be risk-averse and often dependent on specific products or platforms [22]. Rather than imposing a need to redevelop the entire technology infrastructure, SPM can provide an interface between new solutions and legacy systems through modularity and open standards.

Finally, data aggregation and sharing are important for improving outcomes and tracking program success by reducing information gaps and providing a holistic view of clients. Previous research has demonstrated that integrated case management with a multidisciplinary approach may improve positive outcomes for clients [23]. SPM supports comprehensive data on service delivery for accountability and caseworker decision-making through a common data model. At the same time, a combination of business and infrastructure security mechanisms keeps client data protected and secure during program or agency collaboration, supporting trust in social programs.

This system description is limited in that the implementations described are on-premise solutions currently. Data quality and silos can limit the extent of insights and analytics, and how these data are presented influences decision-making. Improving program performance and achieving better health outcomes requires bringing together and presenting data visually to enhance decision-making abilities. These are common challenges for any enterprise-wide solution. We have made the platform more flexible and portable by moving more core processes of SPM to the cloud. We have released SPM to work on an open-source container application platform so that users can secure and use their data across multiple environments, including public and private clouds. We are also providing analytic capabilities across social programs and in a visual format at the point of decision-making to better assess the impact of social programs on health outcomes.

This viewpoint highlights the features of SPM with use cases selected to illustrate its generalizable features, such as benefits management, health and human services administration, and case coordination. These case studies were limited by organizations that were willing to share data and participate in the research. We were not able to design metric collection a priori, so the data for each use case are based on what could be provided by participating organizations.

Conclusions

SPM is a user-centered, configurable, and flexible system designed to manage social program workflows. Its features and functionalities support the goals of social programs through improved service delivery to beneficiaries with functionalities and features for complex eligibility and entitlement, convenient access to services, complex case management, organizational and policy change management, and program transparency. More than 50 government organizations, 280,000 caseworkers, and 30 million beneficiaries are served through SPM, demonstrating the flexibility and scalability across social program types and settings in designing administrative systems that support a streamlined workflow.

Acknowledgments

The authors would like to thank the teams at Diona, Accenture, and KPMG with whom they closely collaborated on the design, development, and execution of various aspects of case report one. The authors would also like to acknowledge the following people for their contributions to the project: Graham Harper, Haeun Jin, David Way, Ciara Layden, Ann Murphy, Pragya Singh, Shane McFadden, Alan Kenny, Shira Alevy, David Brotman, Brett South, Hannah Helmy, and Joanne Hastings.

Conflicts of Interest

The authors of this manuscript are employed by IBM.

XSL•FO
RenderX

Multimedia Appendix 1
Description of Social Program Management architecture.
[DOCX File , 14 KB - medinform_v9i8e23219_app1.docx]

References

1. Kerlin JA. Social Enterprise: A Global Comparison. Lebanon, New Hampshire: University Press of New England; 2009:1-240.
2. Government benefits. USA.gov. URL: https://www.usa.gov/benefits [accessed 2019-12-10]
3. Parton N. Changes in the form of knowledge in social work: From the 'social' to the 'informational'? Br J Soc Work 2006 Nov 08;38(2):253-269. [doi: 10.1093/bjsw/bcl337]
4. Gillingham P, Graham T. Designing electronic information systems for the future: Social workers and the challenge of New Public Management. Crit Soc Policy 2015 Dec 10;36(2):187-204. [doi: 10.1177/0261018315620867]
5. Kulik CT, Ryan S, Harper S, George G. Aging populations and management. Acad Manage J 2014 Aug;57(4):929-935. [doi: 10.5465/amj.2014.4004]
6. Saez E. Income and wealth inequality: evidence and policy implications. Contemp Econ Policy 2016 Oct 14;35(1):7-25. [doi: 10.1111/coep.12210]
7. Norström T, Grönqvist H. The great recession, unemployment and suicide. J Epidemiol Community Health 2015 Feb 22;69(2):110-116 [FREE Full text] [doi: 10.1136/jech-2014-204602] [Medline: 25339416]
8. Carrilio TE. Accountability, evidence, and the use of information systems in social service programs. J Soc Work 2008 Apr 01;8(2):135-148. [doi: 10.1177/1468017307088495]
9. Scurlock-Evans L, Upton D. The role and nature of evidence: a systematic review of social workers' evidence-based practice orientation, attitudes, and implementation. J Evid Inf Soc Work 2015 Mar 06;12(4):369-399. [doi: 10.1080/15433714.2013.853014] [Medline: 25747891]
10. Grundy J, Grundy J. A survey of Australian human services agency software usage. J Technol Hum Serv 2013 Jan;31(1):84-94. [doi: 10.1080/15228835.2012.751297]
11. Wallace LS. A view of health care around the world. Ann Fam Med 2013 Jan 14;11(1):84 [FREE Full text] [doi: 10.1370/afm.1484] [Medline: 23319511]
12. Sigerist HE. From Bismarck to Beveridge: Developments and trends in social security legislation. J Public Health Policy 1999;20(4):474. [doi: 10.2307/3343133]
13. Cúram business application suite on IBM System Z. IBM Redbooks. 2009. URL: http://www.redbooks.ibm.com/abstracts/sg247715.html?Open [accessed 2019-12-10]
14. Supplemental Nutrition Assistance Program (SNAP). USDA-FNS. URL: https://www.fns.usda.gov/snap/supplemental-nutrition-assistance-program [accessed 2019-11-03]
15. Joint Legislative Oversight Committee, on Health and Human Services. Department of Health and Human Services NC FAST. 2016. URL: https://www.ncleg.gov/documentsites/committees/JLOCHHS/Handouts%20and%20Minutes%20by%20Interim/2016-17%20Interim%20JLOC-HHS%20Handouts/September%2027,%202016/Item%20V-DHHS-ITUpdates-JLOCSept27.pdf [accessed 2021-07-23]
16. Gibbs S, Perry-Manning S. Department of Health and Human Services NC FAST update. Joint Legislative Oversight Committee on Information Technology. 2018. URL: https://www.ncleg.gov/DocumentSites/committees/JLOCIT/03-08-2018/JLOC-IT_NC%20FASTUpdate_20180308.pdf [accessed 2021-07-23]
17. Colvin AD, Bullock AN. Technology acceptance in social work education: implications for the field practicum. J Teach Soc Work 2014 Oct 14;34(5):496-513. [doi: 10.1080/08841233.2014.952869]
18. Lagsten J, Andersson A. Use of information systems in social work – challenges and an agenda for future research. Euro J Soc Work 2018 Jan 19;21(6):850-862. [doi: 10.1080/13691457.2018.1423554]
19. Devlieghere J, Roose R. Electronic Information Systems: In search of responsive social work. J Soc Work 2018 Feb 08;18(6):650-665. [doi: 10.1177/1468017318757296]
20. Sarwar A, Harris M. Children's services in the age of information technology: what matters most to frontline professionals. J Soc Work 2018 Jul 13;19(6):699-718. [doi: 10.1177/1468017318788194]
21. Carrilio TE, Packard T, Clapp JD. Nothing in—nothing out. Admin Soc Work 2004 Feb 04;27(4):61-75. [doi: 10.1300/j147v27n04_05]
22. Flemig S, Osborne S, Kinder T. Risky business—reconceptualizing risk and innovation in public services. Public Money Manage 2016 Jul 11;36(6):425-432. [doi: 10.1080/09540962.2016.1206751]
23. Busse R, Stahl J. Integrated care experiences and outcomes in Germany, the Netherlands, and England. Health Aff (Millwood) 2014 Sep;33(9):1549-1558. [doi: 10.1377/hlthaff.2014.0419] [Medline: 25201659]

Abbreviations

SNAP: Supplemental Nutrition Assistance Program
SPM: Social Program Management

XSL•FO
RenderX

Edited by C Lovis; submitted 05.08.20; peer-reviewed by E Chiou, I Mircheva; comments to author 26.10.20; revised version received 08.12.20; accepted 05.06.21; published 30.08.21.

Please cite as:
McKillop M, Snowdon J, Willis VC, Alevy S, Rizvi R, Rewalt K, Lefebvre-Paillé C, Kassler W, Purcell Jackson G
A System to Support Diverse Social Program Management
JMIR Med Inform 2021;9(8):e23219
URL: https://medinform.jmir.org/2021/8/e23219
doi:10.2196/23219
PMID:34459741

©Mollie McKillop, Jane Snowdon, Van C Willis, Shira Alevy, Rubina Rizvi, Karen Rewalt, Charlyne Lefebvre-Paillé, William Kassler, Gretchen Purcell Jackson. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 30.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using Electronic Medical Record Data for Research in a Healthcare Information and Management Systems Society (HIMSS) Analytics Electronic Medical Record Adoption Model (EMRAM) Stage 7 Hospital in Beijing: Cross-sectional Study

Rui Li[1*], PhD; Yue Niu[2*], MM; Sarah Robbins Scott[3], MM; Chu Zhou[3], MD; Lan Lan[4], MD; Zhigang Liang[1], MD; Jia Li[1], MD

[1]Information Center, Xuanwu Hospital, Capital Medical University, Beijing, China

[2]Statistical Procedure Department, Blueballon (Beijing) Medical Research Co, Ltd, Beijing, China

[3]National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China

[4]West China Biomedical Big Data Center, West China Hospital, Sichuan University, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Jia Li, MD
Information Center
Xuanwu Hospital
Capital Medical University
45 Changchun Street
Beijing, 100053
China
Phone: 86 10 83929211
Email: lij@xwhosp.org

Abstract

Background: With the proliferation of electronic medical record (EMR) systems, there is an increasing interest in utilizing EMR data for medical research; yet, there is no quantitative research on EMR data utilization for medical research purposes in China.

Objective: This study aimed to understand how and to what extent EMR data are utilized for medical research purposes in a Healthcare Information and Management Systems Society (HIMSS) Analytics Electronic Medical Record Adoption Model (EMRAM) Stage 7 hospital in Beijing, China. Obstacles and issues in the utilization of EMR data were also explored to provide a foundation for the improved utilization of such data.

Methods: For this descriptive cross-sectional study, cluster sampling from Xuanwu Hospital, one of two Stage 7 hospitals in Beijing, was conducted from 2016 to 2019. The utilization of EMR data was described as the number of requests, the proportion of requesters, and the frequency of requests per capita. Comparisons by year, professional title, and age were conducted by double-sided chi-square tests.

Results: From 2016 to 2019, EMR data utilization was poor, as the proportion of requesters was 5.8% and the frequency was 0.1 times per person per year. The frequency per capita gradually slowed and older senior-level staff more frequently used EMR data compared with younger staff.

Conclusions: The value of using EMR data for research purposes is not well studied in China. More research is needed to quantify to what extent EMR data are utilized across all hospitals in Beijing and how these systems can enhance future studies. The results of this study also suggest that young doctors may be less exposed or have less reason to access such research methods.

(JMIR Med Inform 2021;9(8):e24405) doi:10.2196/24405

KEYWORDS

electronic medical records; data utilization; medical research; China

XSL•FO
RenderX

Introduction

Electronic medical records (EMRs), or digitized versions of patient medical charts, are often considered a key component of a hospital or health care system's health information system [1]. EMR systems have transformed data and record keeping in the medical field, and they enable providers to more systematically track patient information over time, promote a more holistic approach to patient care, support the streamlining of preventative screening, support the monitoring of patients, and improve overall quality [2,3]. For these reasons, there has been rapid growth in the implementation of EMR systems in health care settings throughout the world in recent decades [4-9]. Subsequently, the amount and availability of clinical data automatically collected by EMRs are increasing at an exponential rate [10,11], and EMRs have been recognized as a valuable resource for observational data and for large-scale analyses [12,13]. As such, EMR data are often used for research purposes in many universities and organizations around the world [14,15]. Using EMR data for medical research [16,17] has several benefits, such as being low cost, having a large volume of data, and saving time because there is no need to recruit and retain participants [18-21]. Thus, it is believed that using EMRs to obtain clinical information has the potential to revolutionize medical research in the coming years [22,23].

In China, the EMR system has become the core system for the collection and management of hospital information, as the National Electronic Medical Record System has been promoted across the country since 2011 [24-26]. Furthermore, with many hospitals implementing the Healthcare Information and Management Systems Society (HIMSS) Analytics Electronic Medical Record Adoption Model (EMRAM) standards, numerous Chinese hospitals have become international standard and accredited hospitals [27]. One result of this shift has been that increasing numbers of western institutions are collaborating with China on medical research using EMR data [28].

As research using EMR data has become increasingly prevalent, researchers have been pondering how to better explore the technical value of EMR data. In addition, there exists a growing body of literature on the feasibility and efficacy of using electronic health records for research purposes. Electronic health records (EHRs) are inclusive of a broader view of patient care, including diagnoses, medications, immunizations, family medical history, and provider contact information. EMR data, however, are digital versions of patient charts. They contain notes and information collected by and for clinicians in that particular care setting and are mostly used by providers for diagnosis and treatment [3]. In China and abroad, studies on the topic of using EMR or EHR data for research have primarily focused on the challenges of using such systems. Researchers over a decade ago raised concerns regarding the quality and comprehensiveness of clinical data being collected in EMR systems and mentioned that there were systematic biases inherent to data collected primarily for clinical care [29]. Other studies have identified other barriers, including legal, technical, ethical, social, and resource-related issues, such as privacy protection, data security, data custodians, and the motives for collecting data, as well as a lack of incentives to share data [15,30]. An additional systematic review identified four domains of potential limitations, including data quality issues (91.7%), data preprocessing challenges (53.3%), privacy concerns (18.3%), and potential for limited generalizability (21.7%) [31]. Some studies have consequently developed a list of caveats and recommendations for overcoming such limitations [30,32-35].

Additionally, the majority of existing research focuses on the quality of EMR/EHR data and its related challenges [36-39]. These challenges can be divided into five primary areas as follows: completeness, consistency, validity, reliability, and accuracy [40-42]. Some analyses have aimed to develop assessment frameworks to ensure data quality across studies [43], but there are few studies that quantitatively explore how and to what extent EMR or EHR data are being collected and used in China. Thus, it is necessary to build EMR data quality metrics and standardize routine documentation to enable its secondary use for medical research [44-46].

The paralleled use of EMR data for medical research has been noted. In one such study, the characteristics of EMR data in China were compared against data collected in hospitals in the United States in order to understand system and cultural differences that may exist between Chinese and English clinical documents [47]. A study by van Velthoven et al [48], for example, shed light on the feasibility of extracting EMR data across a number of countries. These studies are useful for understanding how data collection systems in China and the use of EMR data for medical research may adapt to more international standards, further supporting collaboration between Chinese and foreign research institutions.

Currently, in Chinese hospitals, the data available to researchers are limited in scope to just EMRs, rather than full EHRs. In order to further promote utilizing EMR data for research, a quantitative investigation of the current status of data utilization is warranted, since understanding the status quo is a prerequisite for determining barriers and improving the existing system. It is necessary to explore the obstacles that hinder EMR data utilization for medical research from the perspective of data consumers, but there is currently no quantitative research or surveys published on the recent status of EMR data utilization for medical research in any institution or region in China. Thus, this study aimed to understand the landscape, including barriers and obstacles, of utilizing EMR data for medical research in Chinese medical institutions. This study will provide data managers and medical research managers with a broader understanding of what types of data are being used; what extent they are being utilized; and who is accessing such data, laying the groundwork for further promotion of this research method.

Methods

Study Design

A serial, cross-sectional, descriptive study was carried out at Xuanwu Hospital, Capital Medical University (XWHCMU) in Beijing, China. XWHCMU is a large 1600-bed tertiary general hospital with a complete EMR data repository and is one of the two HIMSS Analytics EMRAM Stage 7 hospitals in Beijing. The HIMSS Analytics EMRAM incorporates methodology and

XSL•FO
RenderX

algorithms to automatically score hospitals around the world relative to their EMR capabilities. A Stage 7 rating signifies the highest level of EMR function and application, achieving a near paperless environment that harnesses technology to support optimized patient care. At Xuanwu Hospital, the EMRAM data system was implemented in 2014. All employees receive training on the content and scope of the EMR data available, the permissions for EMR data utilization, and the process of requesting and obtaining EMR data.

Data Sources and Extraction

All data from the Office Information System (Office Automation) was extracted, because each EMR data extraction request in the hospital must be approved through the EMR data management module in the Office Automation. Variables of interest included data request purpose, requester ID, requester department, and data request time. If the purpose of the data request was for scientific research, it was included in the study. The requester ID was used to retrieve the age and professional title of all requesters in the hospital human resources dictionary. The requester ID was also used as the main index for data matching and integration, forming a total of 933 EMR data request records for scientific research purposes between 2016 and 2019.

The use of EMR data for research purposes by key departments in the hospital was also assessed. XWHCMU evaluates the scientific research performance of each department every year based on a set of 18 evaluation criteria, including published papers/books, transformation of scientific research results, academic events, and approved scientific research projects. The top 10 clinical departments with the highest cumulative research work performance score over the last 4 years were selected as "key departments" for this study. The performance score of each department, evaluation indicators, and standards of scientific research work can be found in Multimedia Appendix 1.

Statistical Analysis

The data were analyzed using IBM SPSS Statistics for Windows version 23.0 (IBM Corp). The data were expressed using times, frequencies, and percentages. The chi-square test was used for categorical variables, with $P<.05$ considered statistically significant. A summary of the statistical indicators, their definitions, and how they were calculated can be found in Table 1.

Table 1. Summary of the statistical indicators of the study, their corresponding definitions, and how they were calculated.

Statistical indicators	Definition	Calculation
Times	An absolute value index	The cumulative value of the number of requests for electronic medical record (EMR) data for research by professional and technical personnel in the observation unit (institution or department) during the observation period.
Frequency	An intensity index	The number of requests/∑the number of professional and technical personnel in this observation unit × time.
Proportion of requesters	A ratio indicator	∑the number of professional and technical personnel who have requested EMR data for research/∑the number of professional and technical personnel in this observation unit.
Number of departments that did not request data	A counting indicator	The number of departments that never requested EMR data for scientific research during the observation period.
Absolute increment of frequency	The absolute value of growth	Can be further divided into cumulative growth and annual growth.
Cumulative growth	The absolute value of growth	The difference between the frequency of a certain year and that at baseline (2016).
Annual growth	The absolute value of growth	The difference between the frequency of a year and that of the previous year.
Frequency growth rate	The growth rate of frequency	Divided into fixed base ratio growth rate and link ratio growth rate.
Relative ratio with fixed base	The growth rate of frequency	The net increase rate of frequency in a certain year compared with the baseline (2016), that is, the ratio of a certain year's frequency to the baseline frequency minus 100%.
Link relative	The growth rate of frequency	The net increase rate of frequency in a year compared with the frequency of the previous year, that is, the ratio of frequency of a year to that of the previous year minus 100%.

Results

EMR Data Utilization From 2016 to 2019 at XWHCMU

The frequency of EMR data utilization increased from 0.06 times per person per year (2016) to 0.1 times per person per year (2019), and the proportion of requesters increased from 3.3% (2016) to 5.8% (2019), as seen in Table 2. The majority of medical departments at the hospital are using the EMR system, with the number not using the system decreasing from 21 (2016) to 5 (2019). The fixed base ratio growth rate of the frequency of EMR data utilization was 66.67%, and the year-to-year growth rate in 2019 was zero.

XSL•FO
RenderX

The frequency at which EMR data was used for medical research increased significantly between 2016 and 2018 (Table 2). The growth rate frequency has gradually slowed down over the past 4 years, with a bottleneck occurring in 2019, during which the growth rate was 0%.

Table 2. General trends in the utilization of electronic medical records in Xuanwu Hospital, Capital Medical University, Beijing, China between 2016 and 2019.

Year	Times	Frequency	Proportion of requesters, n/N (%)	Number of departments that did not request data, n/N (%)	Absolute increment of request frequency		Request frequency growth rate, %	
					Cumulative growth	Annual growth	Relative ratio with fixed base	Link relative
2016	171	0.06	98/3060 (3.2%)	21/47 (44.7%)	N/A[a]	N/A	N/A	N/A
2017	201	0.07	119/2935 (4.1%)	19 /47 (40.4%)	0.01	0.01	16.67	16.67
2018	288	0.10	153/2883 (5.3%)	14/47 (29.8%)	0.04	0.03	66.67	42.86
2019	273	0.10	163/2667 (6.1%)	5/47 (10.6%)	0.04	0.00	66.67	0.00

[a]N/A: not applicable.

Utilization of EMR Data by Key Departments at XWHCMU From 2016 to 2019

The key departments had a per capita request frequency lower than the average per capita request frequency for the overall hospital (Table 3). The proportion of data utilization by key departments decreased from 70.0% in 2016 to 49.4% in 2019.

XSL•FO
RenderX

Table 3. Utilization of electronic medical record data in the key scientific research departments of Xuanwu Hospital, Capital Medical University, Beijing, China between 2016 and 2019.

Research score ranking	Department	2016			2017			2018			2019		
		Times	Proportion of the whole hospital request times, %	Frequency	Times	Proportion of the whole hospital request times, %	Frequency	Times	Proportion of the whole hospital request times, %	Frequency	Times	Proportion of the whole hospital request times, %	Frequency
1	Neurology	49	28.8%	0.16	57	28.4%	0.19	70	24.3%	0.23	65	23.8%	0.16
2	Neurosurgery	18	10.6%	0.08	17	8.5%	0.08	16	5.6%	0.07[a]	8	2.9%	0.03[a]
3	Radiology	7	4.1%	0.06	6	3.0%	0.05[a]	8	2.8%	0.07[a]	7	2.6%	0.06[a]
4	General Surgery	5	2.9%	0.05[a]	21	10.4%	0.20	16	5.6%	0.15	10	3.7%	0.07[a]
5	Functional Neurosurgery	1	0.6%	0.01[a]	1	0.5%	0.01[a]	4	1.4%	0.05[a]	4	1.5%	0.05[a]
6	Interventional Radiography	0	0%	0.00[a]	1	0.5%	0.03[a]	9	3.1%	0.26	5	1.9%	0.14
7	Vascular Surgery	13	7.7%	0.19	15	7.5%	0.22	13	4.5%	0.19	5	1.9%	0.07[a]
8	Anesthesiology	0	0%	0.00[a]	2	1.0%	0.01[a]	2	0.7%	0.01[a]	6	2.0%	0.03[a]
9	Pharmacy	25	14.7%	0.20	19	9.5%	0.15	33	11.5%	0.26	22	8.1%	0.17
10	Orthopedics	1	0.6%	0.02[a]	3	1.5%	0.05[a]	3	1.0%	0.05[a]	3	1.0%	0.05[a]
Total	N/A[b]	119	70.0%	N/A	142	70.8%	N/A	174	60.5%	N/A	135	49.4%	N/A
Frequency of the overall hospital	N/A	N/A	N/A	0.06	N/A	N/A	0.07	N/A	N/A	0.10	N/A	N/A	0.10

[a]The annual per capita electronic medical record data utilization frequency of this department was lower than the annual average of the whole hospital. The annual average is based on all departments.

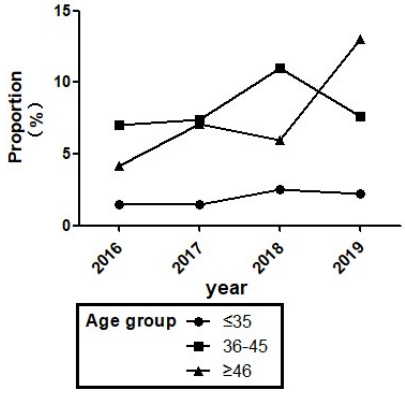
[b]N/A: not applicable.

Utilization of EMR Data by Age

As seen in Figure 1, the trend in the proportion of individuals using EMR data varied between 2016 and 2019. Those aged 36 to 45 years made up the largest proportion of researchers using EMR data from 2016 to 2018, though this trend declined in 2019, when those aged 46 years of age or older made up the larger proportion of requests. Generally speaking, those under the age of 35 years represented the smallest proportion of EMR data users at the hospital.

XSL•FO
RenderX

Figure 1. Trend in the proportion of electronic medical record data users by age group at Xuanwu Hospital, Capital Medical University, Beijing, China between 2016 and 2019.



Utilization of EMR Data by Staff Level

In 2016, the proportion of junior-level professionals using EMR data for medical research was the lowest (1.2%), while those with senior-level titles made up the largest proportion of EMR data users (8.8%). This trend continued through 2019, as seen in Table 4. Between 2016 and 2019, senior-level professionals made up the largest proportion of those requesting EMR data (255/533, 47.8%), followed by intermediate-level staff (161/533, 30.2%) and then junior-level staff (117/533, 21.9%). Over the 4-year period, the proportion of senior- and intermediate-level staff requesting EMR data increased, while there was no significant change in the junior-level staff group.

Table 4. Electronic medical record data utilization by junior-, intermediate-, and senior-level staff at Xuanwu Hospital, Capital Medical University, Beijing, China between 2016 and 2019.

Year	Professional title			Total, n/N (%)	Chi-square (*df*)	*P* value
	Junior-level requester, n/N (%)	Intermediate-level requester, n/N (%)	Senior-level requester, n/N (%)			
2016	23/1894 (1.2%)	26/658 (4.0%)	49/508 (9.6%)	98/3060 (3.2%)	84.155 (5)	<.001
2017	22/1811 (1.2%)	37/648 (5.7%)	60/476 (12.6%)	119/2935 (4.1%)	131.622 (5)	<.001
2018	38/1772 (2.1%)	44/644 (6.8%)	71/467 (15.2%)	153/2883 (5.3%)	191.04 (5)	<.001
2019	34/1755 (1.9%)	54/497 (10.9%)	75/415 (18.1%)	163/2667 (6.1%)	147.299 (5)	<.001

Discussion

Principal Findings

This study aimed to understand the landscape of EMR data utilization for medical research at XWHCMU between 2016 and 2019. In the past 4 years, the use of EMR data for medical research was quite uncommon at the hospital. Though overall utilization rates increased each year, the overall growth rate is slowing, with a frequency of just 0.1 times per person per year in 2019. More so, key research departments at the hospital are not utilizing EMR data for research purposes, while junior-level staff continue to be limited in their ability to use the system.

According to the results of this study, the proportion of hospital staff using EMR data was less than 6% and the frequency of EMR data utilization did not exceed 10 times per 100 researchers in 1 year. Meanwhile, even the top 10 research departments at Xuanwu Hospital reduced the frequency at which they used EMR data for medical research purposes. Current clinical scientific research data collection still heavily relies on semimanual input. In China, the Hospital Information System has continuously improved, with the EMR system accumulating a large amount of valuable health care data. According to the Annual Report on the Status of Chinese Hospital Informatization (2018-2019), more than one-fourth of tertiary medical institutions have invested in EMR data utilization for research [26]. Since prospective clinical research is more demanding and difficult to perform, retrospective research is an important means of obtaining clinical evidence. EMR data can be not only used as independent data, but also tied to administrative data for retrospective research [13,16,17], saving both time and money for medical institutions wishing to carry out such research studies with limited resources [18,19]. Thus, steps within the hospital should be taken to promote the awareness of this type of available research data, along with the encouragement to carry our medical research using these systems. Further evaluations are needed to gain a better understanding as to why current medical staff may not be accessing such data or why these trends may be declining.

Although the frequency of data usage has increased significantly (the fixed base ratio growth rate was 66.67%), this was not found to be significant, and a bottleneck was noted in 2019.

XSL•FO
RenderX

The reasons for this decline in data utilization over the last 3 years were not analyzed, though further follow-up studies to determine the factors influencing the decisions for EMR data utilization would be beneficial. These studies could examine if the external environment has changed, including policies for utilizing EMR data, mechanisms for data sharing, and procedures for requesting and obtaining data.

This study also found that older more senior professionals at Xuanwu Hospital were more likely to use EMR data compared to younger age groups ($P<.001$). Junior-level staff should be the main force for tapping the value of the EMR data, as they need scientific research achievements to be promoted and younger individuals tend to accept new technologies and new methods faster compared to older populations [49]. In large general hospitals in China, all professional and technical staff are required to have independent scientific research capabilities and publications. However, there is a serious contrast between actual need and actual use of EMR data among junior-level staff, as seen in this study. While this study did not evaluate such contrasts, other research has aimed to identify why such barriers to data access may exist, as noted in the Introduction section of this manuscript. The first issue of data access may be inequality, as bureaucracy has been noted as one of main barriers when using EMR data for research [48]. If this is the case at hospitals in Beijing, it is urgent to establish an equal and open EMR data utilization mechanism. Another potential barrier is whether there is a lack of awareness of the research value of EMR data among younger junior-level staff [50]. Lastly, the EMR data utilization skills of junior-level staff may be insufficient [51,52]. If awareness and skills are indeed lacking, it is required to establish systematic training and technical support services for this group [53,54].

Limitations

As this study was limited to one hospital in Beijing, China, the results cannot represent the general situation of other medical institutions in China. In addition, due to information confidentiality, more personnel-related information could not be obtained and the included indicators may not be comprehensive. For other factors that may affect the utilization of EMR data, further research is needed.

Conclusions

This is the first quantitative study considering EMR data utilization for medical research in a hospital in Beijing. It offers unique insights into the frequency of EMR data usage for medical research purposes and who is utilizing such data. The value of using EMR data for research purposes remains understudied. The results of this study also suggest that young doctors may be less exposed or have less reason to access such research methods. More research is needed to quantify to what extent EMR data are utilized across all hospitals in Beijing and how these systems can enhance future studies.

Acknowledgments

We are grateful to the Information Center, the Scientific Research Management Department, and the Human Resources Department for their cooperation throughout the study. The Information Center helped us extract the data request records; the Scientific Research Management Department provided the performance score of each department, evaluation indicators, and standards of scientific research work; and the Human Resources Department provided related personnel information. We thank Anjie Ren for guidance on this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Research performance assessment standard of XuanWu hospital.
[DOC File , 104 KB - medinform_v9i8e24405_app1.doc]

References

1. Lugn NE. Connecting for Health: Global Vision, Local Insight. J Telemed Telecare 2016 Jun 22;12(3):161-162. [doi: 10.1258/135763306776738585]
2. Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna TJ, CDC Prevention Epicenter Program. Incidence and Trends of Sepsis in US Hospitals Using Clinical vs Claims Data, 2009-2014. JAMA 2017 Oct 03;318(13):1241-1249 [FREE Full text] [doi: 10.1001/jama.2017.13836] [Medline: 28903154]
3. What are the differences between electronic medical records, electronic health records, and personal health records? Health Information Technology. URL: https://www.healthit.gov/faq/what-are-differences-between-electronic-medical-records-electronic-health-records-and-personal [accessed 2020-01-01]
4. Kawaguchi H, Koike S, Ohe K. Facility and Regional Factors Associated With the New Adoption of Electronic Medical Records in Japan: Nationwide Longitudinal Observational Study. JMIR Med Inform 2019 Jun 14;7(2):e14026 [FREE Full text] [doi: 10.2196/14026] [Medline: 31199307]
5. O'Donnell A, Kaner E, Shaw C, Haighton C. Primary care physicians' attitudes to the adoption of electronic medical records: a systematic review and evidence synthesis using the clinical adoption framework. BMC Med Inform Decis Mak 2018 Nov 13;18(1):101 [FREE Full text] [doi: 10.1186/s12911-018-0703-x] [Medline: 30424758]

XSL•FO
RenderX

6. Owens B. Family doctors call for guaranteed access to EMR data for research and quality improvement. CMAJ 2018 Jan 15;190(2):E60-E61 [FREE Full text] [doi: 10.1503/cmaj.109-5543] [Medline: 29335269]
7. Zhang XY, Zhang P. Recent perspectives of electronic medical record systems. Exp Ther Med 2016 Jun;11(6):2083-2085 [FREE Full text] [doi: 10.3892/etm.2016.3233] [Medline: 27284289]
8. Henry J. Adoption of Electronic Health Record Systems among U.S. Non-federal Acute Care Hospitals: 2008-2013. Health Information Technology. 2014. URL: https://www.healthit.gov/sites/default/files/briefs/oncdatabrief16.pdf [accessed 2020-02-01]
9. Ministry of Health and Welfare. URL: http://www.mohw.go.kr/eng/ [accessed 2020-02-05]
10. Puskarich MA, Callaway C, Silbergleit R, Pines JM, Obermeyer Z, Wright DW, et al. Priorities to Overcome Barriers Impacting Data Science Application in Emergency Care Research. Acad Emerg Med 2019 Jan;26(1):97-105 [FREE Full text] [doi: 10.1111/acem.13520] [Medline: 30019795]
11. Ehrenstein V, Nielsen H, Pedersen AB, Johnsen SP, Pedersen L. Clinical epidemiology in the era of big data: new opportunities, familiar challenges. Clin Epidemiol 2017;9:245-250 [FREE Full text] [doi: 10.2147/CLEP.S129779] [Medline: 28490904]
12. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. J Healthc Eng 2018;2018:4302425 [FREE Full text] [doi: 10.1155/2018/4302425] [Medline: 29849998]
13. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: use of electronic medical records for health outcomes research: a literature review. Med Care Res Rev 2009 Dec;66(6):611-638. [doi: 10.1177/1077558709332440] [Medline: 19279318]
14. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, et al. Using electronic health records for clinical research: the case of the EHR4CR project. J Biomed Inform 2015 Feb;53:162-173 [FREE Full text] [doi: 10.1016/j.jbi.2014.10.006] [Medline: 25463966]
15. Canaway R, Boyle DI, Manski-Nankervis JAE, Bell J, Hocking JS, Clarke K, et al. Gathering data for decisions: best practice use of primary care electronic records for research. Med J Aust 2019 Apr;210 Suppl 6:S12-S16 [FREE Full text] [doi: 10.5694/mja2.50026] [Medline: 30927466]
16. Linn G, Ying Y, Chang K. Does Computerized Physician Order Entry Benefit from Dynamic Structured Data Entry? A Quasi-Experimental Study. BMC Med Inform Decis Mak 2018 Nov 26;18(1):109 [FREE Full text] [doi: 10.1186/s12911-018-0709-4] [Medline: 30477491]
17. Reimer AP, Milinovich A, Madigan EA. Data quality assessment framework to assess electronic medical record data for use in research. Int J Med Inform 2016 Jun;90:40-47 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.03.006] [Medline: 27103196]
18. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Expert Panel. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007;14(1):1-9 [FREE Full text] [doi: 10.1197/jamia.M2273] [Medline: 17077452]
19. Weiskopf N, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. AMIA Annu Symp Proc 2013;2013:1472-1477 [FREE Full text] [Medline: 24551421]
20. Lingren T, Sadhasivam S, Zhang X, Marsolo K. Electronic medical records as a replacement for prospective research data collection in postoperative pain and opioid response studies. Int J Med Inform 2018 Mar;111:45-50 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.12.014] [Medline: 29425633]
21. Lai YS, Afseth JD. A review of the impact of utilising electronic medical records for clinical research recruitment. Clin Trials 2019 Apr;16(2):194-203. [doi: 10.1177/1740774519829709] [Medline: 30764659]
22. Duz M, Marshall JF, Parkin T. Validation of an Improved Computer-Assisted Technique for Mining Free-Text Electronic Medical Records. JMIR Med Inform 2017 Jun 29;5(2):e17 [FREE Full text] [doi: 10.2196/medinform.7123] [Medline: 28663163]
23. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013 Apr 03;309(13):1351-1352. [doi: 10.1001/jama.2013.393] [Medline: 23549579]
24. 2008-2013 white paper on the status of hospital informatization in China. Chima. URL: https://chima.org.cn/Sites/Uploaded/File/2020/07/2163730922566786578013601 17.pdf [accessed 2020-03-05]
25. 2017-2018 white paper on the status of hospital informatization in China. Chima. URL: https://chima.org.cn/Html/News/Articles/11000170.html [accessed 2020-02-05]
26. 2018-2019 white paper on the status of hospital informatization in China. Chima. URL: https://chima.org.cn/Html/News/Articles/4878.html [accessed 2020-03-05]
27. Stage 6 and 7 Achievement. HIMSS Analytics. URL: https://www.himssanalytics.org/asia-pacific/stage-6-7-achievement [accessed 2020-04-01]
28. Wu Y, Lei J, Wei WQ, Tang B, Denny JC, Rosenbloom ST, et al. Analyzing differences between chinese and english clinical text: a cross-institution comparison of discharge summaries in two languages. Stud Health Technol Inform 2013;192:662-666 [FREE Full text] [Medline: 23920639]

XSL•FO
RenderX

29. Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? Ann Intern Med 2009 Sep 01;151(5):359-360 [FREE Full text] [doi: 10.7326/0003-4819-151-5-200909010-00141] [Medline: 19638404]
30. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol 2017 Jan;106(1):1-9 [FREE Full text] [doi: 10.1007/s00392-016-1025-6] [Medline: 27557678]
31. Edmondson M, Reimer A. Challenges Frequently Encountered in the Secondary Use of Electronic Medical Record Data for Research. Comput Inform Nurs 2020 Jul;38(7):338-348. [doi: 10.1097/CIN.0000000000000609] [Medline: 32149742]
32. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013 Aug;51(8 Suppl 3):S30-S37 [FREE Full text] [doi: 10.1097/MLR.0b013e31829b1dbd] [Medline: 23774517]
33. Hersh W, Cimino J, Payne PR, Embi P, Logan J, Weiner M, et al. Recommendations for the use of operational electronic health record data in comparative effectiveness research. EGEMS (Wash DC) 2013;1(1):1018 [FREE Full text] [doi: 10.13063/2327-9214.1018] [Medline: 25848563]
34. Callahan A, Shah NH, Chen JH. Research and Reporting Considerations for Observational Studies Using Electronic Health Record Data. Ann Intern Med 2020 Jun 02;172(11 Suppl):S79-S84 [FREE Full text] [doi: 10.7326/M19-0873] [Medline: 32479175]
35. Milinovich A, Kattan MW. Extracting and utilizing electronic health data from Epic for research. Ann Transl Med 2018 Feb;6(3):42 [FREE Full text] [doi: 10.21037/atm.2018.01.13] [Medline: 29610734]
36. Ni K, Chu H, Zeng L, Li N, Zhao Y. Barriers and facilitators to data quality of electronic health records used for clinical research in China: a qualitative study. BMJ Open 2019 Jul 02;9(7):e029314 [FREE Full text] [doi: 10.1136/bmjopen-2019-029314] [Medline: 31270120]
37. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 May 02;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
38. Raman SR, Curtis LH, Temple R, Andersson T, Ezekowitz J, Ford I, et al. Leveraging electronic health records for clinical research. Am Heart J 2018 Aug;202:13-19. [doi: 10.1016/j.ahj.2018.04.015] [Medline: 29802975]
39. Ming-Yu Z. Resource Management of Big Data of Clinical Research. China Digital Medicine 2020;15(08). [doi: 10.3969/j.issn.1673-7571.2020.08.034]
40. Long JA, Richards JA, Seko CE. The Canadian Institute for Health Information (CIHI) Data Quality Framework, Version 1: A Meta-Evaluation and Future Directions. In: Proceedings of the 6th International Conference on Information Quality. 2001 Presented at: 6th International Conference on Information Quality; 2001; Cambridge, MA p. 370-383 URL: https://www.zhangqiaokeyan.com/academic-conference-foreign_proceedings-of-the-6th-internatial-cfere_thesis/020512600211.html
41. Opmeer BC. Electronic Health Records as Sources of Research Data. JAMA 2016 Jan 12;315(2):201-202. [doi: 10.1001/jama.2015.15419] [Medline: 26757473]
42. Baier AW, Snyder DJ, Leahy IC, Patak LS, Brustowicz RM. A Shared Opportunity for Improving Electronic Medical Record Data. Anesth Analg 2017 Sep;125(3):952-957. [doi: 10.1213/ANE.0000000000002134] [Medline: 28632540]
43. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 01;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
44. Brown ML. Can't you just pull the data? The limitations of using of the electronic medical record for research. Paediatr Anaesth 2016 Nov;26(11):1034-1035. [doi: 10.1111/pan.12951] [Medline: 27747978]
45. von Martial S, Brix TJ, Klotz L, Neuhaus P, Berger K, Warnke C, et al. EMR-integrated minimal core dataset for routine health care and multiple research settings: A case study for neuroinflammatory demyelinating diseases. PLoS One 2019;14(10):e0223886 [FREE Full text] [doi: 10.1371/journal.pone.0223886] [Medline: 31613917]
46. Assessing Data Quality for Healthcare Systems Data Used in Clinical Research. NIH Collaboratory. 2014. URL: https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Assessing-data-quality_V1%200.pdf [accessed 2020-04-05]
47. Wu Y, Lei J, Wei W, Tang B, Denny JC, Rosenbloom ST, et al. Analyzing differences between chinese and english clinical text: a cross-institution comparison of discharge summaries in two languages. Stud Health Technol Inform 2013;192:662-666 [FREE Full text] [Medline: 23920639]
48. van Velthoven MH, Mastellos N, Majeed A, O'Donoghue J, Car J. Feasibility of extracting data from electronic medical records for research: an international comparative study. BMC Med Inform Decis Mak 2016 Jul 13;16:90 [FREE Full text] [doi: 10.1186/s12911-016-0332-1] [Medline: 27411943]
49. Lee CC, Czaja SJ, Moxley JH, Sharit J, Boot WR, Charness N, et al. Attitudes Toward Computers Across Adulthood From 1994 to 2013. Gerontologist 2019 Jan 09;59(1):22-33 [FREE Full text] [doi: 10.1093/geront/gny081] [Medline: 29982458]
50. Zhao JW. Establishment of a clinical research big data center in a hospital in Henan province. Chin J Hosp Admin 2020;36(8):668-671. [doi: 10.3760/cma.j.cn111325-20200415-01110]
51. Haixing W. Discussion on the application of health big data in clinical research. Chinese Hospitals 2020;24(7):63-64. [doi: 10.19660/j.issn.1671-0592.2020.07.19]

XSL•FO
RenderX

52. Rui L, Lan L, Zhigang L, Zhifang G, Yan Y, Jia L. Analysis of Data Utilization of Hospital Information System for Clinical Research. Chinese Journal of Health Informatics and Management 2018;15(1):86-89. [doi: 10.3969/j.issn.1672-5166.2018.01.017]
53. Fanxiu H. Hospital Big Data Framework System Construction and Data Utilization. Chinese Journal of Health Informatics and Management 2020;17(3):275-278.
54. Jia QG. Practice of cultivating the ability of big data mining in graduates working for professional degree in medical oncology. Chin J Med Edu 2020;19(5):551-554. [doi: 10.3760/cma.j.cn116021-20191204-00125]

Abbreviations

EHR: electronic health record
EMR: electronic medical record
EMRAM: Electronic Medical Record Adoption Model
HIMSS: Healthcare Information and Management Systems Society
XWHCMU: Xuanwu Hospital, Capital Medical University

Edited by C Lovis; submitted 17.09.20; peer-reviewed by A Reimer, I Mircheva, Z Ren, N Kaur; comments to author 07.10.20; revised version received 01.12.20; accepted 07.06.21; published 03.08.21.

Please cite as:
Li R, Niu Y, Scott SR, Zhou C, Lan L, Liang Z, Li J
Using Electronic Medical Record Data for Research in a Healthcare Information and Management Systems Society (HIMSS) Analytics Electronic Medical Record Adoption Model (EMRAM) Stage 7 Hospital in Beijing: Cross-sectional Study
JMIR Med Inform 2021;9(8):e24405
URL: https://medinform.jmir.org/2021/8/e24405
doi:10.2196/24405
PMID:34342589

©Rui Li, Yue Niu, Sarah Robbins Scott, Chu Zhou, Lan Lan, Zhigang Liang, Jia Li. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 03.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Predicting Patients' Intention to Use a Personal Health Record Using an Adapted Unified Theory of Acceptance and Use of Technology Model: Secondary Data Analysis

Consuela Cheriece Yousef[1,2,3], MPH, PharmD, PhD; Teresa M Salgado[4], MPharm, PhD; Ali Farooq[5], DSc, MSc, MCS, MBA; Keisha Burnett[6], EdD, MS, SCT; Laura E McClelland[7], PhD; Abin Thomas[8], MSc, PhD; Ahmed O Alenazi[1,2,3], BSc, PharmD, CACP; Laila Carolina Abu Esba[2,3,9], BSc, MSc, PharmD; Aeshah AlAzmi[2,3,10], BSc, PharmD, SCSCP, FISQUA; Abrar Fahad Alhameed[2,3,11], PharmD, BCPS, BCIDP; Ahmed Hattan[2,3,9], PharmD; Sumaya Elgadi[2,12], MS, PharmD; Saleh Almekhloof[2,3,13], BSc, PharmD; Mohammed A AlShammary[2,3,14], BSPharm; Nazzal Abdullah Alanezi[2,3,15], BSHIM, MSHSA; Hani Solaiman Alhamdan[2,3,10], BSc, MSc; Sahal Khoshhal[2,3,11], PharmD; Jonathan P DeShazo[7], MPH, PhD

[1]Pharmaceutical Care Department, Ministry of National Guard-Health Affairs, Dammam, Saudi Arabia
[2]King Abdullah International Medical Research Center, Riyadh, Saudi Arabia
[3]King Saud bin Abdul-Aziz University for Health Sciences, Riyadh, Saudi Arabia
[4]Department of Pharmacotherapy & Outcome Science, School of Pharmacy, Virginia Commonwealth University, Richmond, VA, United States
[5]Department of Computing, University of Turku, Turku, Finland
[6]Department of Clinical Laboratory Sciences, Cytopathology Practice Program, University of Tennessee Health Science Center, Memphis, TN, United States
[7]Department of Health Administration, Virginia Commonwealth University, Richmond, VA, United States
[8]Department of Biostatistics and Bioinformatics, King Abdullah International Medical Research Center, Riyadh, Saudi Arabia
[9]Pharmaceutical Care Department, Ministry of National Guard-Health Affairs, Riyadh, Saudi Arabia
[10]Pharmaceutical Care Department, Ministry of National Guard-Health Affairs, Jeddah, Saudi Arabia
[11]Pharmaceutical Care Department, Ministry of National Guard-Health Affairs, Madinah, Saudi Arabia
[12]Department of Pharmacy Practice, College of Pharmacy, Princess Noura Bint Abdulrahman University, Riyadh, Saudi Arabia
[13]Pharmaceutical Care Department, Ministry of National Guard-Health Affairs, Al Ahsa, Saudi Arabia
[14]Primary Health Care, Prince Bader Housing Clinic, Riyadh, Saudi Arabia
[15]Qassim Primary Health Care Center, Ministry of National Guard-Health Affairs, Qassim, Saudi Arabia

Corresponding Author:
Consuela Cheriece Yousef, MPH, PharmD, PhD
Pharmaceutical Care Department
Ministry of National Guard-Health Affairs
PO Box 4616
Dammam
Saudi Arabia
Phone: 966 138532555 ext 1680
Email: consuela_73@hotmail.com

Abstract

Background: With the rise in the use of information and communication technologies in health care, patients have been encouraged to use eHealth tools such as personal health records (PHRs) for better health and well-being services. PHRs support patient-centered care and patient engagement. To support the achievement of the Kingdom of Saudi Arabia's Vision 2030 ambitions, the National Transformation program provides a framework to use PHRs in meeting the 3-fold aim for health care—increased access, reduced cost, and improved quality of care—and to provide patient- and person-centered care. However, there has been limited research on PHR uptake within the country.

XSL•FO
RenderX

Objective: Using the Unified Theory of Acceptance and Use of Technology (UTAUT) as the theoretical framework, this study aims at identifying predictors of patient intention to utilize the Ministry of National Guard-Health Affairs PHR (MNGHA Care) app.

Methods: Using secondary data from a cross-sectional survey, data measuring the intention to use the MNGHA Care app, along with its predictors, were collected from among adults (n=324) visiting Ministry of National Guard-Health Affairs facilities in Riyadh, Jeddah, Dammam, Madinah, Al Ahsa, and Qassim. The relationship of predictors (main theory constructs) and moderators (age, gender, and experience with health apps) with the dependent variable (intention to use MNGHA Care) was tested using hierarchical multiple regression.

Results: Of the eligible population, a total of 261 adult patients were included in the analysis. They had a mean age of 35.07 (SD 9.61) years, 50.6 % were male (n=132), 45.2% had university-level education (n=118), and 53.3% had at least 1 chronic medical condition (n=139). The model explained 48.9% of the variance in behavioral intention to use the PHR (P=.38). Performance expectancy, effort expectancy, and positive attitude were significantly associated with behavioral intention to use the PHR (P<.05). Prior experience with health apps moderated the relationship between social influence and behavioral intention to use the PHR (P=.04).

Conclusions: This study contributes to the existing literature on PHR adoption broadly as well as in the context of the Kingdom of Saudi Arabia. Understanding which factors are associated with patient adoption of PHRs can guide future development and support the country's aim of transforming the health care system. Similar to previous studies on PHR adoption, performance expectancy, effort expectancy, and positive attitude are important factors, and practical consideration should be given to support these areas.

(JMIR Med Inform 2021;9(8):e30214) doi:10.2196/30214

KEYWORDS

personal health record; patient portal; eHealth; Middle East; Saudi Arabia; Unified Theory of Acceptance and Use of Technology; prediction; intention; electronic health record; acceptance; model; framework; secondary analysis

Introduction

Background

The transformation of health care delivery has been a global phenomenon since the turn of the 21st century [1,2]. Health care delivery has evolved from a paternalistic "doctor knows best" model to one where individuals are encouraged to play an active role in their health [3]. As the prevalence of chronic diseases increases along with the rise in information and communication technologies, patients have been encouraged to accept more responsibility for their health and well-being by using eHealth tools [4,5].

Personal health records (PHRs) are eHealth tools that aim to increase patient engagement and empowerment by allowing individuals to keep track of their personal health information. PHRs have been defined as "an Internet-based set of tools that allows people to access and coordinate their lifelong health information and make appropriate parts of it available to those who need it" [6]. Nevertheless, PHR has no uniform definition, with numerous terms being used interchangeably in the literature, namely "patient web portal," "patient portal," "computerized patient portal," "patient-accessible electronic health record," "tethered PHR," and "electronic PHR." PHRs hold great potential in chronic disease management [7].

Health care organizations adopt PHRs to increase patient engagement to meet the 3-fold aim for health care: increased access, reduced cost, and improved quality of care [7-9]. Some of the proposed benefits from the use of PHRs are empowerment, continuity of care, education, patient-provider partnership, individual control, and engagement. Managing chronic diseases requires regular use of self-management skills such as identifying problems, finding solutions, using information sources, collaborating with health care providers, altering behavior, and assessing results [10].

Research Problem and Aim

In 2018, the Ministry of National Guard Health Affairs (MNG-HA) implemented its PHR, known as MNGHA Care. MNHGA Care features include checking laboratory results, scheduling appointments, requesting medical reports, requesting prescription refills, viewing radiology reports, and providing vaccination reminders. It allows patients to upload personal health information such as blood pressure, blood sugar measurements, weight, and exercise information. A self-assessment feature allows patients to enter information on pain control, performance status, and quality of life. Educational resources are also provided on the PHR. Two years prior to implementing the PHR, Al Sahan and Saddik [11] evaluated the knowledge and perceptions toward using a PHR among 454 patients and 9 technical staff from an MNG-HA hospital in Riyadh before implementation. Participants reported a high level of interest (very interested: 60.6%, interested: 25.2%) in a web-based PHR. Since the implementation, further research is needed on patient adoption.

The aim of this study was to identify a set of constructs that predict the intention to use the MNGHA Care PHR among patients, using the Unified Theory of Acceptance and Use of Technology (UTAUT) as a theoretical framework. Before a technology is adopted, a user must first intend to use the technology [12]. The benefits of increased accessibility, reduced costs, and better quality of health care with the PHR can only be achieved by understanding what motivates individuals to use this technology.

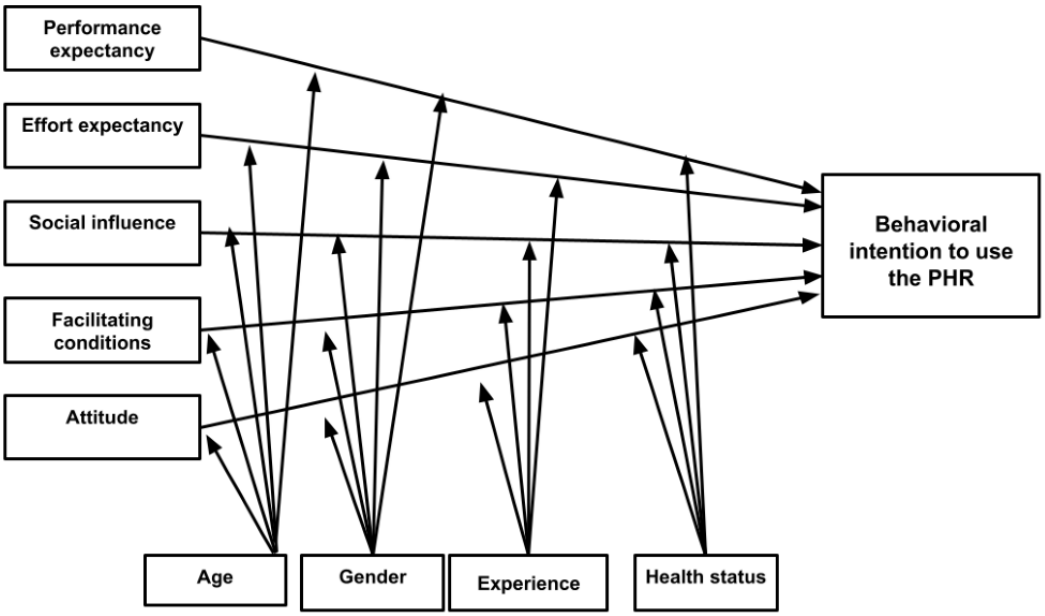
XSL•FO
RenderX

Theoretical Framework

While there are many models available to explain user acceptance, Venkatesh et al [12] developed the UTAUT to provide a comprehensive framework to explain acceptance and usage of information technology in organizations. It is a synthesis of 8 theoretical models, including Theory of Reasoned Action, Technology Acceptance Model, Motivational Model, Theory of Planned Behavior, Combined Technology Acceptance Model–Theory of Planned Behavior, Model of Personal Computer Utilization, Diffusion of Innovation Theory, and Social Cognitive Theory [12]. Venkatesh et al [12] evaluated the independent variables that influence behavioral intention and actual use of technology. The three independent constructs—performance expectancy, effort expectancy, and social influence—directly influence the behavioral intention to use technology. Facilitating conditions and behavioral intention act directly on actual use of technology. Gender, age, voluntariness, and experience are moderators in the framework. This study will adapt UTAUT to investigate the factors that influence patients' intention to use MNGHA Care.

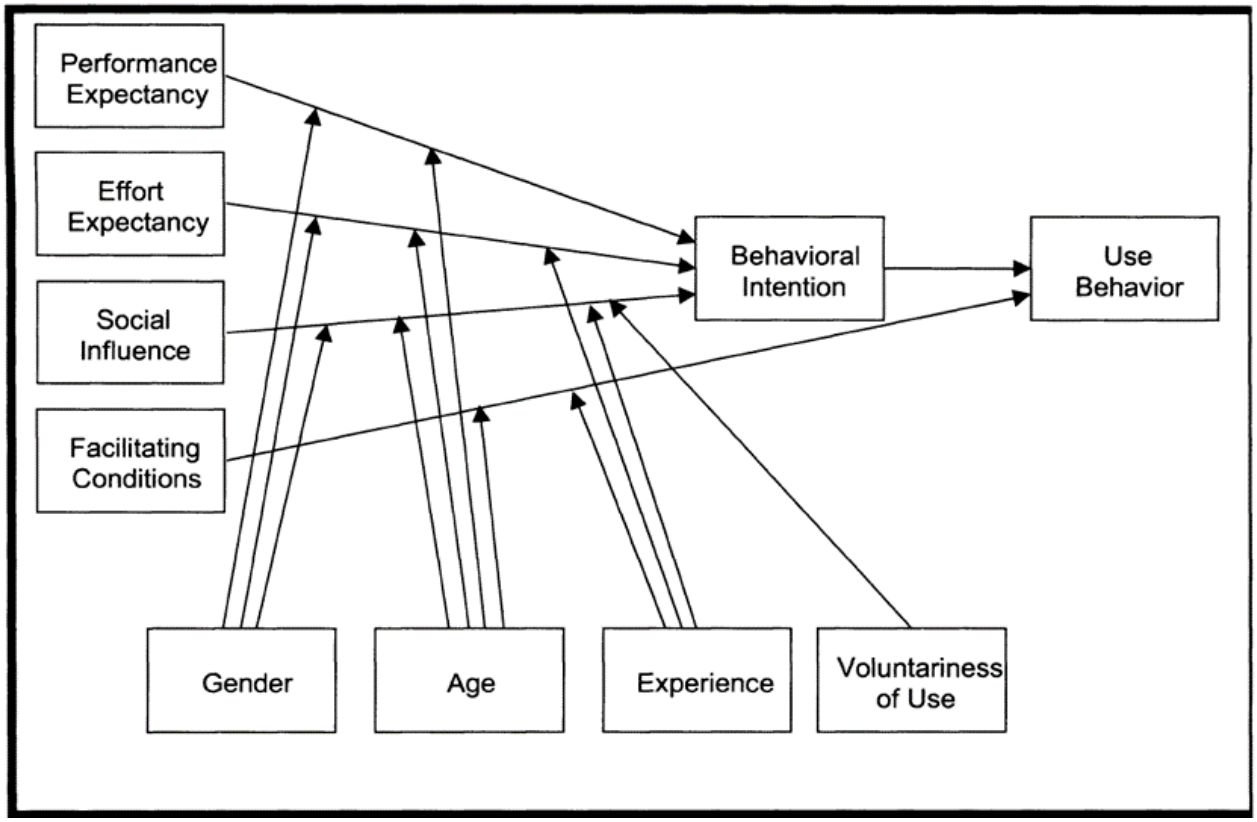
The adapted UTAUT model for this study is presented in Figure 1. Figure 2 shows the original UTAUT. There are 3 adaptations to the original model. First, the construct of attitude is added. In the critical review of the UTAUT model, Dwivedi et al [13] recommended revising the model to include the construct of attitude. Individual characteristics are not included in UTAUT [13]. However, studies have found individual traits to be important predictors of technology acceptance [14,15]. Secondly, the moderators of gender, age, experience, and voluntariness of use are used in the original UTAUT model. In the adapted model, the voluntariness of use is dropped as a moderator since PHR use is voluntary. Finally, health status is added to moderate the relationships among the predictors and the behavioral intention to use the PHR.

Figure 1. Adapted Unified Theory of Acceptance and Use of Technology model to predict patient intention to use the MNGHA Care PHR.



XSL•FO
RenderX

Figure 2. Original Unified Theory of Acceptance and Use of Technology [12].



The proposed differences between this research model and the original UTAUT model are shown in Table 1. Age and gender will moderate all relationships. Women and younger individuals are expected to have a stronger behavioral intention to use the PHR. Experience is operationalized as the prior use of health apps. Venkatesh et al [12] characterized experience as experience with the system being implemented. Experience using a health app would imply that the individual has the necessary computer and internet skills to use a PHR. Limited computer and internet experience has been identified as a barrier to PHR adoption [16]. Individuals with experience using health apps are expected to have a stronger behavioral intention to use the PHR. Finally, health status was selected as a moderator because it has been shown to be an important driver of PHR acceptance [8,17]. If resources and support are available, individuals with poorer health are more likely to use eHealth technologies [18]. Health status in this study will be based on self-reported health status. Patients with poorer health status are expected to have a stronger behavioral intention to use the PHR.

Table 1. Proposed differences between the original and adapted Unified Theory of Acceptance and Use of Technology model for patients.

Relationships	Original model moderators				Adapted model moderators			
	Gender	Age	Experience	Voluntariness	Age	Gender	Experience	Health status
Performance expectancy–behavioral intention	✓	✓			✓	✓		✓
Effort expectancy–behavioral intention	✓	✓	✓		✓	✓	✓	
Social influence–behavioral intention	✓	✓	✓	✓	✓	✓	✓	✓
Behavioral intention–actual usage								
Facilitating conditions–actual usage		✓	✓					
Facilitating conditions–behavioral intention					✓	✓	✓	✓
Attitude–behavioral intention					✓	✓	✓	✓

XSL•FO
RenderX

Methods

Study Design

Data for the study were obtained from a cross-sectional survey study [19] in which data were collected to examine health information–seeking behavior and PHR (MNG-HA Care) use among patients. Secondary data were used in the current study. Institutional Review Board approval (RD19/002/D) was obtained from King Abdullah International Medical Research Center and Virginia Commonwealth University (HM20020713).

Setting and Participants

MNG-HA is a large health care system that provides medical care to the National Guard's soldiers and their dependents in all regions across the Kingdom of Saudi Arabia. The target study population consisted of adults who visited outpatient facilities (primary or specialty care) in five major cities—Dammam, Riyadh, Jeddah, Madinah, and Qassim. In the original study, a total of 546 adults completed the survey.

For this secondary analysis, participants who answered all questions related to the use of the MNGHA Care PHR constituted the study sample (n=324). A minimum sample size of 270 was calculated for the analysis on the basis of the 10 times rule, which posits that the minimum sample size should be 10 times the number of predictors (27 in this case, including 5 independent variables, 4 moderators, and 18 interaction terms) [20].

Data Collection

As mentioned above, secondary data were used in this study. The original data were collected between December 2019 and February 2020. The survey we used was adapted from Hoogenbosch et al [21]'s study of a PHR using UTAUT, with minor modifications to existing items and additional items created to fit the objectives of the study. Responses to each question were provided on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). However, questions were limited to avoid respondent burden resulting in 1 or 2 items used for each construct.

Behavioral intention, the dependent variable, measures the strength of an individual's intention to perform a specific behavior; that is, to use the MNGHA Care PHR [22]. A 2-item scale was used to measure behavioral intention: "I will probably use MNGHA Care in the future" and "I intend to use MNGHA Care regularly." The reliability coefficient was Cronbach α=.76.

Performance expectancy reflects the degree to which an individual believes that using a technology will help attain significant rewards. Unlike Hoogenbosch et al [21], who used 3 items to measure this construct, we used the single item, "By using MNGHA Care, I feel more involved in my care."

Effort expectancy is the degree of ease associated with the use of technology—in this case, the PHR [12]. The single item, "Information in MNGHA Care is understandable," was used to measure effort expectancy, unlike Hoogenbosch et al [21], who used a 5-item scale.

Social influence refers to an individual's perception of how important people in their social circle are, using technology [12]. Consistent with Hoogenbosch et al [21], the following item was used to measure this construct: "My healthcare professional encouraged me to use MNGHA Care."

The construct of facilitating condition refers to organizational and technical infrastructure support technology use [12]. The single item, "Technical help is available when I do not know how to use MNGHA Care," was used to measure this construct instead of the 3 items used by Hoogenbosch et al [21].

Attitude relates to positive or negative feelings associated with using a technology [22] and was assessed with the self-constructed item "MNGHA Care is a valuable service."

Performance expectancy, effort expectancy, social influence, facilitating conditions, and attitude were independent variables.

Self-reported age, educational level, gender, health care facility, marital status, employment status, and monthly household income were recorded. Health care characteristics included the following: presence of a medical condition, number and type of medical conditions, self-reported health status, hospitalization in the past 6 months, and emergency department visits in the past 6 months. Health status was a categorical variable self-reported as excellent, very good, good, fair, or poor. Experience was a dichotomous variable defined as experience with health apps and assessed through the question: "Do you use health applications (apps) on your mobile phone?"

The moderators for the model were age, gender, experience, and health status.

Statistical Analysis

Descriptive statistics and hierarchical multiple regression were conducted using SPSS (version 25, IBM Corp) [23]. While structural equation modeling is a more robust statistical method for testing a theoretical model and allows for single-item measures [24], it was not used owing to concerns that the model would not yield good results since all constructs were a single item. Data were assessed for normality, linearity, homoscedasticity, and absence of multicollinearity. Normality was assessed using skewness and kurtosis and found to be within the required threshold of –1.96 to +1.96 [25]. A Kolmogorov–Smirnov test was also used to test for normality with nonstatistical significance (P>.05), indicating that the data were normally distributed. Independence of observations was tested using the Durbin–Watson test, which yielded a coefficient of 1.905. As a rule of thumb, values between 1.5 and 2.5 are considered normal [26]. Linearity was confirmed by the appearance of a linear representation of standardized residuals on a scatterplot. Multicollinearity was checked by examining correlations and variance inflation factor (VIF) between variables. A VIF above 10 is an indicator of multicollinearity [27]. "No VIF greater than 10 was identified, indicating a lack of multicollinearity.

Three-stage hierarchical multiple regression analysis was conducted with behavioral intention as the dependent variable. The independent variables were entered into the regression model in 3 sequential blocks with all assumptions of regression

XSL•FO
RenderX

met and outliers removed. The first block included the 5 independent variables of performance expectancy, effort expectancy, social influence, facilitating conditions, and attitude. The second block contained the moderator variables of age, gender, experience, health status, and independent variables. Experience was a categorical variable with 0 representing people with no experience using health apps and 1 representing people with experience using health apps. To test the moderating effects of gender, age, experience, and health status on the relationship of independent variables (performance expectancy, effort expectancy, social influence, facilitating conditions, and attitude) and behavioral intention to use the PHR, interaction terms were added to the regression model in block 3. For each block, the standardized regression coefficient (β) and the R^2 were calculated.

Results

Demographic and Health Care Characteristics

Of the 324 participants who completed the survey about MNGHA Care use, 261 comprised the final sample after outlier removal. The mean age of the participants was 35.07 (SD 9.61) years. Most users were male (n=132, 50.6%), from the Central region (n=110, 42.1%), married (n=208, 79.7%), and had a higher educational level (university graduate: n=118, 45.2%) and a monthly income of at least US $2666 (n=95, 36.4%). For health status, the majority of participants (n=178, 68.2%) had a medical condition with the following being the most common chronic conditions: asthma or chronic obstructive pulmonary disease (n=46, 17.6%), diabetes (n=38, 14.6%), and hypertension (n=32, 12.3%). Table 2 summarizes the demographic and health care characteristics of the respondents.

XSL•FO
RenderX

Table 2. Demographic and health care characteristics of the study participants (N=261).

Characteristic	Value
Demographic information	
Age (years), mean (SD)	35.07 (9.61)
Region of the country, n (%)	
Eastern	81(31.0)
Central	110 (42.1)
Western	70 (26.8)
Gender, n (%)	
Male	132 (50.6)
Female	129 (49.4)
Marital status, n (%)	
Married	208 (79.7)
Single	53 (20.3)
Education level, n (%)	
Elementary school or less	14 (5.4)
Middle school	17 (6.5)
High school	91 (34.9)
University	118 (45.2)
Postgraduate	20 (7.7)
Employment status, n (%)	
Employed	142 (54.4)
Retired	16 (6.1)
Student	17 (6.5)
Unemployed	84 (32.2)
Monthly household income, n (%)	
<5000 SAR (US $1333)	69 (26.4)
5000-9999 SAR (US $1333-2666)	84 (32.2)
>10,000 SAR (US $2666)	95 (36.4)
Health status characteristics	
Have a medical condition, n (%)	178 (68.2)
Number of medical conditions, n (%)	
None	83 (31.8)
1	139 (53.3)
≥2	39 (14.9)
Type of medical condition, n (%)	
Diabetes	38 (14.6)
Hypertension	32 (12.3)
Asthma or chronic obstructive pulmonary disease	46 (17.6)
Heart failure	9 (3.4)
Cancer	11 (4.2)
Sickle cell disease	7 (2.7)
Psychiatric condition	4 (1.5)
Other	78 (29.9)

XSL•FO
RenderX

Characteristic	Value
Self-reported health status, n (%)	
Excellent	121 (46.4)
Very good	95 (36.4)
Good	33 (12.6)
Fair	8 (3.1)
Poor	4 (1.5)
Hospitalized within the last 6 months, n (%)	54 (20.7)
Visited the emergency department within the last 6 months, n (%)	124 (47.5)

Hypothesized Relationships

The results of hierarchical multiple regression analysis are presented in Table 3. The first stage of the model revealed that performance expectancy, effort expectancy, social influence, facilitating conditions, and attitude contributed significantly to the regression model ($F_{5,255}$=38.874; P<.001) and accounted for 43.3% of the explained variance in patients' intention to use MNGHA Care. Effort expectancy and attitude were almost equally important predictors with standardized regression coefficients of 0.249 and 0.198, respectively.

In the second stage of the model, the variables age, gender, experience with health applications, and health status were entered along with the independent variables. These variables did not significantly contribute to the regression model with an additional explained variance of 0.8% in the R^2 value ($F_{4,251}$=0.950; P=.44).

In the third stage, the full model included the independent variables, moderating variables (age, gender, experience with health applications, and health status), and interaction terms. Adding the interaction terms to the model accounted for an additional 5.6% of explained variance and was not significant ($F_{20,231}$=1.075; P=.38). Figure 3 reflects the moderating effect of app experience on social influence in behavioral intention to use the PHR (β=–0.236; t_{231}=–2.036; P=0.04).

XSL•FO
RenderX

Table 3. Summary of the results of hierarchical regression analysis for variables predicting behavioral intention to use the personal health record among study participants (N=261).

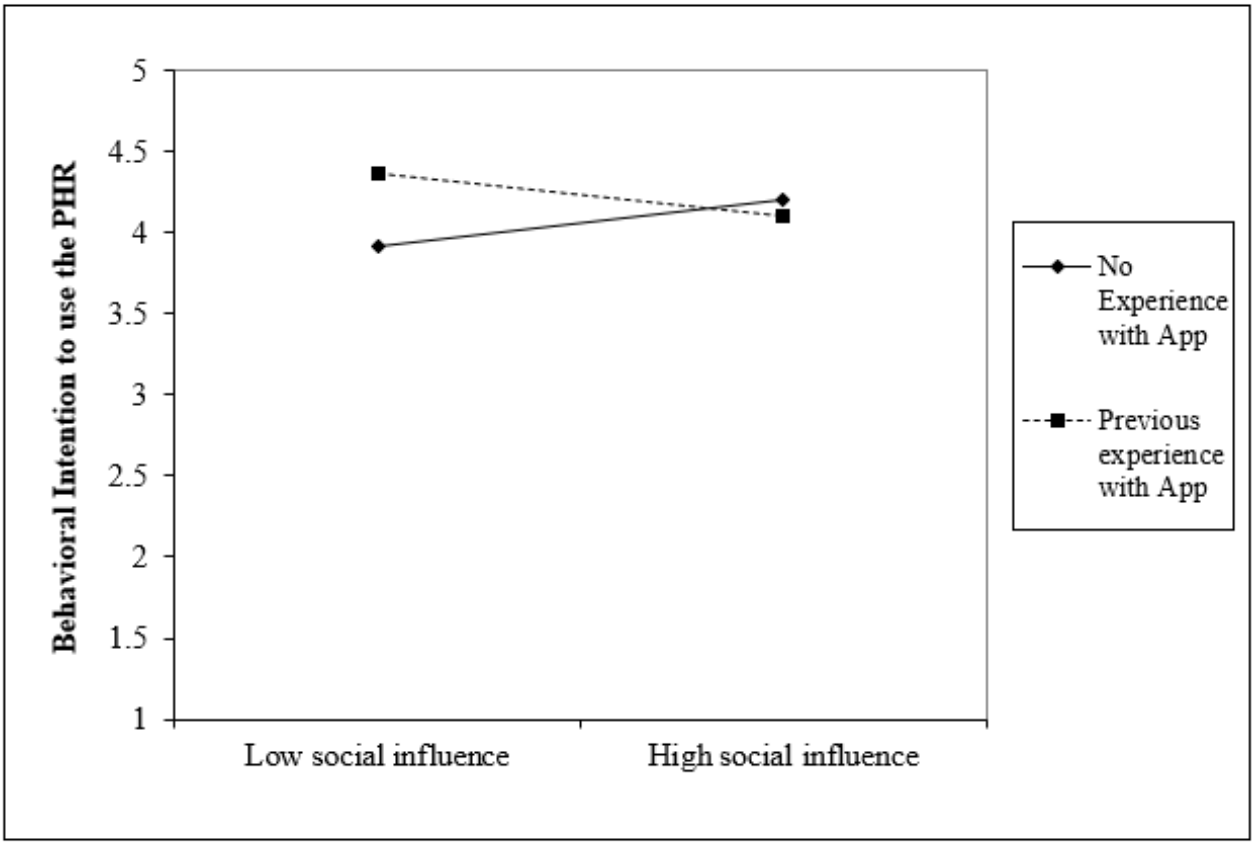
Variables	β (SE)	β	*t* test (*df*)	R^2
Block 1				0.433[b]
Performance expectancy	0.261[a] (0.054)	0.286[a]	4.847[a] (*255*)	
Effort expectancy	247[a] (0.057)	0.249[a]	4.338[a] (*255*)	
Social influence	011 (0.040)	0.017	0.267 (*255*)	
Facilitating conditions	062 (0.038)	0.100	1.618 (*255*)	
Attitude	0.174[a] (0.053)	0.198[a]	3.282[a] (*255*)	
Block 2				0.441
Gender	–0.008 (.055)	–0.007	–0.150 (*251*)	
Age	0.001 (.003)	0.023	0.446 (*251*)	
Experience	0.108 (.056)	0.095	1.934 (*251*)	
Health status	–0.003 (.030)	–0.004	–0.084 (*251*)	
Block 3				0.489
Performance expectancy * gender	0.184 (0.126)	0.143	1.456 (*231*)	
Performance expectancy * age	0.005 (0.009)	0.052	0.631 (*231*)	
Performance expectancy * experience	0.128 (0.129)	0.116	0.991 (*231*)	
Performance expectancy * health status	0.034 (0.083)	0.034	0.403 (*231*)	
Effort expectancy * gender	–0.131 (0.129)	–0.099	–1.013 (*231*)	
Effort expectancy * age	–0.006 (0.007)	–0.053	–0.875 (*231*)	
Effort expectancy * experience	0.027 (0.141)	0.022	0.191 (*231*)	
Effort expectancy * health status	0.001 (0.061)	0.001	0.018 (*231*)	
Social influence * gender	0.064 (0.088)	0.071	0.734 (*231*)	
Social influence * age	–0.005 (0.005)	–0.066	–0.912 (*231*)	
Social influence * experience	–0.182[a] (0.090)	–0.236[a]	–2.036[a] (*231*)	
Social influence * health status	–0.079 (0.054)	–0.107	–1.459 (*231*)	
Facilitating conditions * gender	0.016 (0.091)	0.020	0.178 (*231*)	
Facilitating conditions * age	–0.002 (0.004)	–0.038	–0.506 (*231*)	
Facilitating conditions * experience	0.074 (0.093)	0.096	0.794 (*231*)	
Facilitating conditions * health status	0.002 (0.055)	0.003	0.043 (*231*)	
Attitude * gender	–0.115 (0.124)	–0.098	–0.930 (*231*)	
Attitude * age	–0.008 (0.008)	–0.080	–1.015 (*231*)	
Attitude * experience	–0.219 (0.144)	–0.206	–1.518 (*231*)	
Attitude * health status	0.034 (0.082)	0.035	0.421 (*231*)	

[a]$P<.05$.

[b]$P<.001$.

XSL•FO
RenderX

Figure 3. Interaction between social influence and experience on behavioral intention (*P*=.04). PHR: personal health record.



Discussion

Principal Findings

This study attempted to identify predictors in the adoption of the MNGHA Care PHR among patients from a single, large, integrated health care organization in the Kingdom of Saudi Arabia, using an adapted UTAUT model. The structural model used in this study explained 48.9% of the variance in behavioral intention to use MNGHA Care. Performance expectancy, effort expectancy, and positive attitude were positive predictors of behavioral intention, confirming the construct of attitude has a significant impact on PHR adoption. The individual characteristics of age, gender, experience with health applications, and health status did not significantly influence behavioral intention. As depicted in Figure 3, higher social influence led to higher behavioral intention to use MNGHA Care in patients without previous experience using health apps. On the contrary, among patients who had experience using health applications, social influence negatively affected behavioral intention to use the app. There was a greater impact of experience with low social influence than with the high social influence.

Other studies have also shown performance expectancy and effort expectancy to be significantly and positively associated with PHR adoption [8,18,21,28-31]. This study supports the evidence that patients are more likely to use PHRs when they perceive them as useful and easy to use.

In this study, social influence and facilitating conditions were not associated with behavioral intention. This aligns with the findings of Tavares and Oliveira [18]. Although social influences such as interactions with health care providers have been identified as important in patients' adoption of PHRs, our findings did not find a significant impact [7,8,32]. Yousef et al [19], however, reported that health care providers (47.9%) or hospital staff (10.8%) were mainly responsible for recommending the use of MNGHA Care. Facilitating conditions likely did not have a significant impact as users found the organizational resources and technical help adequate.

Finally, a positive attitude toward the PHR was found to have a significant impact on behavioral intention. Attitude is a strong predictor of behavioral intention to use various types of technology and is the direct precedent of intention [22]. This is aligned with the findings of other studies on PHRs [28,33]. Since attitudes may be influenced by various factors (eg, peers, health care providers, and other health care staff), promoting the PHR can encourage positive attitudes, which can ultimately lead to PHR adoption.

Implications for Theory

This study contributes to the existing literature on PHRs and provides several implications for theory. First, it provides an understanding of the predictors of PHR adoption in general and, more specifically, within the context of the Middle East and the Kingdom of Saudi Arabia. PHRs have not been widely adopted, and there is limited data on predictors of PHR adoption in this region [34,35].

XSL•FO
RenderX

Second, it extends UTAUT with the construct of attitude and the moderators "experience with health applications" and "health status" in a health care setting. The results of this study provided further support for the constructs of performance expectancy, effort expectancy, and attitude to have significant and positive effects on PHR adoption, which is consistent with the literature. Alsahafi et al [34] was the first study to empirically examine predictors of PHR acceptance in the Kingdom of Saudi Arabia. In their study of the general Saudi adult population, they conducted a cross-sectional study and extended UTAUT with the construct of eHealth literacy. Similar to the findings of Alsahafi et al [34], this study found that performance expectancy and effort expectancy were positive predictors of behavioral intention. Contrary to our findings, social influence was found to be a positive predictor for behavioral intention to use a PHR in women. While gender, age, and internet experience were used as moderators, gender was the only variable with a significant moderating role in the aforementioned study. In contrast, our study found the experience with health apps to be the only significant moderator even though the moderating effect was small and accounted for 4.8% of the explained variance.

In the health care context, the integration of constructs from health behavior theories, such as perceived health threat and self-perception, may be useful [18,36]. Though UTAUT was developed to be a comprehensive framework to study technology acceptance, contextual considerations are required to explain PHR adoption behavior best.

Implications for Practice

The Kingdom of Saudi Arabia has prioritized the use of eHealth technologies such as PHRs in health care delivery [34,35,37-39]. To meet the goals of the National Transformation Program, health care organizations around the country will increasingly be called upon to leverage PHRs to efficiently deliver person- and patient-centered care. This study may help organizations better understand patient perceptions of the PHR and lead them to identify strategies to engage patients with the PHR to better manage their health and well-being.

This study found that performance expectancy, effort expectancy, and attitude significantly impact the adoption of PHRs. Tailored marketing strategies have been used to promote the advantages of PHRs and are a way for patients to see the benefits of using a PHR to manage their health [7]. The design and functionalities of the PHR can play an important role in patients' intention to use [7]. Designing a PHR with an easy-to-use, attractive interface with simple language will improve patients' perceptions of the ease of use and help prevent health disparities [40]. Attitude have been identified as a barrier to the use of PHRs in a number of studies [7]. Patients may have negative attitudes toward a PHR for a number of reasons, and this can contribute to their refusal to use PHRs. When health care providers educate and train patients on the features, functionalities, and benefits of the PHR, a positive attitude will develop and facilitate acceptance. However, for health care providers to play this role, they must be knowledgeable about the benefits and purpose of a PHR.

Limitations

There are several limitations to this study. First, this was secondary data analysis, and all constructs for the independent variables were single-item measures. This could have affected the reliability and validity of our findings. Most conceptual constructs are complex and multifaceted and, therefore, a single item may not be an "accurate, comprehensive, and reliable measurement" [41]. However, this was necessary to avoid the respondent burden. Second, a common method bias may be present since the independent variable and dependent variable were measured at a single point in time with only 1 data collection instrument. Finally, the generalizability may have been affected because the study was limited to 1 organization in the country.

Recommendations for Future Research

Because this study was subject to common method bias, future researchers should examine the independent and dependent variables at different time points and with at least 2 different instruments. We were unable to secure access to either the system logs or patient records, but a future study may incorporate these types of data to minimize this bias.

Examining theories in new contexts advances theories and increases external validity [13,42]. Selecting constructs that explain the behavioral intention relationship should be context-based. In this study, the model tested explained 48.9% of the variance in behavioral intention, suggesting the inclusion of attitude was relevant and reasonable. However, other predictors may have improved the model. Future studies may consider adding other constructs shown to be influential in PHR adoption or, more broadly, eHealth adoption. Alaiad et al [36] recommend including constructs recognized as inhibitors of technology adoption as well as adding constructs related to health-related behavior.

The construct of privacy and security should be investigated. Studies showed that privacy and security concerns have a significantly negative effect on behavioral intention to use a PHR [7,8,43-45]. As opposed to technology such as e-banking, PHRs may be accessible to a wide range of health care personnel [46] as well as family members. Patients have raised concerns about identity theft and the possibility of their leaked health information limiting employment opportunities [46]. This study is one of the few to evaluate the moderating effect of variables on the relationship between the independent variables and behavioral intention to use a PHR. Most PHR studies have not assessed moderating or mediating effects [8]. The only significant moderating effect observed was experience with health apps on the relationship between social influence and behavioral intention. Other variables acting as either mediators or moderators may help enrich our understanding of PHR adoption within this context. Abd-Alrazaq et al [47] developed the Abd-Alrazaq Model to examine mediating, moderating, and moderated mediating effects on patients' behavioral intention to use a PHR in England.

For the moderator of health status, a single self-reported health status item was used owing to its simplicity and to reduce the respondent burden; it has been found to be a valid and reliable

XSL•FO
RenderX

measure of health status in high-income countries [48]. However, operationalizing health status in another way may have provided alternative findings. Future studies should measure health status through another method.

Further, future studies should consider using more mixed-methods approaches. In the systematic review of PHR use by Abd-Alrazaq [8], 88% of the studies were quantitative. Mixed-methods studies are suitable to develop multiple perspectives and a comprehensive understanding of PHR adoption. A qualitative approach alongside quantitative methods will provide deeper insight into the patient's perspective.

Finally, more studies should evaluate the health care provider's perspective of PHR adoption. The focus on a more engaged patient has been a paradigm shift in medicine [49]. Therefore, understanding health care provider perspectives is fundamental to the successful implementation, adoption, and continued use of a PHR [49,50]. Negative or indifferent attitudes among health care providers have been identified as a barrier to patient adoption [7]. Fears of increased workload, threats to autonomy, or upsetting patients are some concerns [50]. Addressing these concerns can lead to health care provider endorsement and subsequent patient adoption.

Conclusions

The use of PHRs in the Kingdom of Saudi Arabia is relatively new and will continue to grow in line with Vision 2030 and the MNG-HA's aim to be a center of excellence through the effective use of technology in health care delivery. This study extended the UTAUT model by adding the construct of attitude along with age, gender, experience, and health status as moderators. Our findings show that performance expectancy and effort expectancy had a significant positive effect on behavioral intention. This study provides evidence that attitude had a significant positive effect on behavioral intention to use a PHR. Additionally, the impact of experience with health apps as a moderator of social influence was supported in our study. These results can help the organization further understand ways to encourage and support patients in adopting PHRs.

Acknowledgments

We would like to thank the patients who participated in this study.

Conflicts of Interest

None declared.

References

1. Eysenbach G. What is e-health? J Med Internet Res 2001;3(2):E20 [FREE Full text] [doi: 10.2196/jmir.3.2.e20] [Medline: 11720962]
2. Eysenbach G, Diepgen TL. The role of e-health and consumer health informatics for evidence-based patient choice in the 21st century. Clin Dermatol 2001;19(1):11-17. [doi: 10.1016/s0738-081x(00)00202-9] [Medline: 11369478]
3. Meier CA, Fitzgerald MC, Smith JM. eHealth: extending, enhancing, and evolving health care. Annu Rev Biomed Eng 2013;15:359-382. [doi: 10.1146/annurev-bioeng-071812-152350] [Medline: 23683088]
4. Ariaeinejad R, Archer N. Importance of mobile technology in successful adoption and sustainability of a chronic disease support system. Int J Soc Behav Educ Econ Bus Ind Eng 2014;8:875. [doi: 10.5281/zenodo.1091730]
5. Tenforde M, Jain A, Hickner J. The value of personal health records for chronic disease management: what do we know? Fam Med 2011 May;43(5):351-354 [FREE Full text] [Medline: 21557106]
6. The Personal Health Working Group. Connecting for Health: A Public-Private Collaborative. Markle Foundation. 2003 Jul 01. URL: https://www.markle.org/sites/default/files/final_phwg_report1.pdf [accessed 2021-08-03]
7. Zhao JY, Song B, Anand E, Schwartz D, Panesar M, Jackson GP, et al. Barriers, Facilitators, and Solutions to Optimal Patient Portal and Personal Health Record Use: A Systematic Review of the Literature. AMIA Annu Symp Proc 2017;2017:1913-1922 [FREE Full text] [Medline: 29854263]
8. Abd-Alrazaq AA, Bewick BM, Farragher T, Gardner P. Factors that affect the use of electronic personal health records among patients: A systematic review. Int J Med Inform 2019 Jun;126:164-175. [doi: 10.1016/j.ijmedinf.2019.03.014] [Medline: 31029258]
9. Wolfe A. Institute of Medicine Report: Crossing the Quality Chasm: A New Health Care System for the 21st Century. Policy Polit Nurs Pract 2016 Aug 13;2(3):233-235. [doi: 10.1177/152715440100200312]
10. O'Leary K, Vizer L, Eschler J, Ralston J, Pratt W. Understanding patients' health and technology attitudes for tailoring self-management interventions. AMIA Annu Symp Proc 2015;2015:991-1000 [FREE Full text] [Medline: 26958236]
11. Al-Sahan A, Saddik B. Perceived challenges for adopting the Personal Health Record (PHR) at Ministry of National Guard Health Affairs (MNGHA)- Riyadh. Online J Public Health Inform 2016;8(3):e205 [FREE Full text] [doi: 10.5210/ojphi.v8i3.6845] [Medline: 28210426]
12. Venkatesh V, Morris MG, Davis GB, Davis FD. User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]

XSL•FO
RenderX

13. Dwivedi Y, Rana N, Jeyaraj A, Clement M, Williams M. Re-examining the Unified Theory of Acceptance and Use of Technology (UTAUT): Towards a Revised Theoretical Model. Inf Syst Front 2017 Jun 8;21(3):719-734. [doi: 10.1007/s10796-017-9774-y]
14. Williams MD, Rana NP, Dwivedi YK. The unified theory of acceptance and use of technology (UTAUT): a literature review. Journal of Ent Info Management 2015 Apr 13;28(3):443-488. [doi: 10.1108/JEIM-09-2014-0088]
15. Rosen P. The effect of personal innovativeness in the domain of information technology on the acceptance and use of technology: A working paper. Oklahoma State University. URL: https://www.researchgate.net/profile/Peter-Rosen-2/publication/228868534_The_effect_of_personal_innovativeness_in_the_domain_of_information_technology_on_the_acceptance_and_use_of_technology/links/5409d24f0cf2f2b29a2cc559/The-effect-of-personal-innovativeness-in-the-domain-of-information-technology-on-the-acceptance-and-use-of-technology.pdf [accessed 2021-08-03]
16. Taha J, Czaja SJ, Sharit J, Morrow DG. Factors affecting usage of a personal health record (PHR) to manage health. Psychol Aging 2013 Dec;28(4):1124-1139 [FREE Full text] [doi: 10.1037/a0033911] [Medline: 24364414]
17. Najaftorkamam M, Ghapanchi A, Talaei-Khoei A. Analysis of Research in Adoption of Person-Centred Healthcare Systems: The Case of Online Personal Health Record. 2014 Presented at: 25th Australasian Conference on Information Systems (ACIS); December 8-10, 2014; Auckland URL: http://openrepository.aut.ac.nz/handle/10292/8104
18. Tavares J, Oliveira T. Electronic Health Record Patient Portal Adoption by Health Care Consumers: An Acceptance Model and Survey. J Med Internet Res 2016 Mar 02;18(3):e49 [FREE Full text] [doi: 10.2196/jmir.5069] [Medline: 26935646]
19. Yousef CC, Thomas A, Alenazi AO, Elgadi S, Abu Esba LC, AlAzmi A, et al. Adoption of a Personal Health Record in the Digital Age: Cross-Sectional Study. J Med Internet Res 2020 Oct 28;22(10):e22913 [FREE Full text] [doi: 10.2196/22913] [Medline: 32998854]
20. Hair J, Hult G, Ringle C, Sarstedt M. A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). Thousand Oaks, CA: Sage Publications; 2016.
21. Hoogenbosch B, Postma J, de Man-van Ginkel JM, Tiemessen NA, van Delden JJ, van Os-Medendorp H. Use and the Users of a Patient Portal: Cross-Sectional Study. J Med Internet Res 2018 Sep 17;20(9):e262 [FREE Full text] [doi: 10.2196/jmir.9418] [Medline: 30224334]
22. Davis FD, Bagozzi RP, Warshaw PR. User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. Manage Sci 1989 Aug;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
23. IBM SPSS software. IBM. 2017. URL: https://www.ibm.com/analytics/spss-statistics-software [accessed 2021-08-03]
24. Hair JF, Risher JJ, Sarstedt M, Ringle CM. When to use and how to report the results of PLS-SEM. Eur Bus Rev 2019 Jan 14;31(1):2-24. [doi: 10.1108/ebr-11-2018-0203]
25. George D, Mallery P. SPSS for Windows Step by Step: A Simple Guide and Reference. Boston, MA: Allyn & Bacon; 2010.
26. Tabachnick B, Fidell L. Using Multivariate Statistics. Boston, MA: Pearson; 2013.
27. Field A. Discovering Statistics Using IBM SPSS Statistics. Thousand Oaks, CA: Sage Publications; 2013.
28. Chung M, Ho C, Wen H. Predicting intentions of nurses to adopt patient personal health records: A structural equation modeling approach. Comput Methods Programs Biomed 2016 Nov;136:45-53. [doi: 10.1016/j.cmpb.2016.08.004] [Medline: 27686702]
29. Dontje K, Corser WD, Holzman G. Understanding Patient Perceptions of the Electronic Personal Health Record. J Nurse Pract 2014 Nov;10(10):824-828. [doi: 10.1016/j.nurpra.2014.09.009]
30. Emani S, Yamin CK, Peters E, Karson AS, Lipsitz SR, Wald JS, et al. Patient perceptions of a personal health record: a test of the diffusion of innovation model. J Med Internet Res 2012 Nov 05;14(6):e150 [FREE Full text] [doi: 10.2196/jmir.2278] [Medline: 23128775]
31. Hsieh H, Kuo Y, Wang S, Chuang B, Tsai C. A Study of Personal Health Record User's Behavioral Model Based on the PMT and UTAUT Integrative Perspective. Int J Environ Res Public Health 2016 Dec 23;14(1):8 [FREE Full text] [doi: 10.3390/ijerph14010008] [Medline: 28025557]
32. Vreugdenhil M, Ranke S, de Man Y, Haan M, Kool R. Patient and Health Care Provider Experiences With a Recently Introduced Patient Portal in an Academic Hospital in the Netherlands: Mixed Methods Study. J Med Internet Res 2019 Aug 20;21(8):13743 [FREE Full text] [doi: 10.2196/13743] [Medline: 31432782]
33. Khaneghah P, Miguel-Cruz A, Bentley P, Liu L, Stroulia E, Ferguson-Pell M. Users' Attitudes Towards Personal Health Records: A Cross-Sectional Pilot Study. Appl Clin Inform 2016;7(2):573-586 [FREE Full text] [doi: 10.4338/ACI-2015-12-RA-0180] [Medline: 27437062]
34. Alsahafi Y, Gay V, Khwaji A. Factors affecting the acceptance of integrated electronic personal health records in Saudi Arabia: The impact of e-health literacy. Health Inf Manag 2020 Nov 28:1833358320964899. [doi: 10.1177/1833358320964899] [Medline: 33249857]
35. Alanazi A, Anazi YA. The Challenges in Personal Health Record Adoption. J Healthc Manag 2019;64(2):104-109. [doi: 10.1097/JHM-D-17-00191] [Medline: 30845058]
36. Alaiad A, Alsharo M, Alnsour Y. The Determinants of M-Health Adoption in Developing Countries: An Empirical Investigation. Appl Clin Inform 2019 Oct;10(5):820-840 [FREE Full text] [doi: 10.1055/s-0039-1697906] [Medline: 31667819]

XSL•FO
RenderX

37. Alharbi M. The Status Quo of Health Information Technology and Health Information Management Efficiency in Saudi Arabia: A Narrative Review. Int J Health Res Innov 2018;6(1):11-23.
38. Alshahrani A, Stewart D, MacLure K. A systematic review of the adoption and acceptance of eHealth in Saudi Arabia: Views of multiple stakeholders. Int J Med Inform 2019 Aug;128:7-17. [doi: 10.1016/j.ijmedinf.2019.05.007] [Medline: 31160014]
39. Alsulame K, Khalifa M, Househ M. E-Health status in Saudi Arabia: A review of current literature. Health Policy Technol 2016 Jun;5(2):204-210. [doi: 10.1016/j.hlpt.2016.02.005]
40. Hoque R, Sorwar G. Understanding factors influencing the adoption of mHealth by the elderly: An extension of the UTAUT model. Int J Med Inform 2017 May;101:75-84. [doi: 10.1016/j.ijmedinf.2017.02.002] [Medline: 28347450]
41. Hassard J. Secondary Data Analysis: An Introduction. Birkbeck, University of London. URL: http://www.bbk.ac.uk/cswl/publications/conference-events-speaking-engagements/resources/Secondary [accessed 2021-08-03]
42. Venkatesh V, Sykes T, Zhang X. 'Just What the Doctor Ordered': A Revised UTAUT for EMR System Adoption and Use by Doctors. 2011 Presented at: 44th Hawaii International Conference on System Sciences; January 4-7, 2011; Kauai, HI. [doi: 10.1109/HICSS.2011.1]
43. Elsafty A, Elbouseery IM, Shaarawy A. Factors Affecting the Behavioral Intention to Use Standalone Electronic Personal Health Record Applications by Adults in Egypt. BMS 2020 Nov 22;6(4):14. [doi: 10.11114/bms.v6i4.5066]
44. Niazkhani Z, Toni E, Cheshmekaboodi M, Georgiou A, Pirnejad H. Barriers to patient, provider, and caregiver adoption and use of electronic personal health records in chronic care: a systematic review. BMC Med Inform Decis Mak 2020 Jul 08;20(1):153 [FREE Full text] [doi: 10.1186/s12911-020-01159-1] [Medline: 32641128]
45. Showell C. Barriers to the use of personal health records by patients: a structured review. PeerJ 2017;5:e3268 [FREE Full text] [doi: 10.7717/peerj.3268] [Medline: 28462058]
46. Pushpangadan S, Seckman C. Consumer Perspective on Personal Health Records: A Review of the Literature. Healthcare Information and Management Systems Society. 2015. URL: https://www.himss.org/resources/consumer-perspective-personal-health-records-review-literature [accessed 2021-08-03]
47. Abd-Alrazaq A, Alalwan A, McMillan B, Bewick B, Househ M, Al-Zyadat A. Patients' Adoption of Electronic Personal Health Records in England: Secondary Data Analysis. J Med Internet Res 2020 Oct 07;22(10):e17499 [FREE Full text] [doi: 10.2196/17499] [Medline: 33026353]
48. Cullati S, Mukhopadhyay S, Sieber S, Chakraborty A, Burton-Jeangros C. Is the single self-rated health item reliable in India? A construct validity study. BMJ Glob Health 2018;3(6):e000856 [FREE Full text] [doi: 10.1136/bmjgh-2018-000856] [Medline: 30483411]
49. Shah SD, Liebovitz D. It Takes Two to Tango: Engaging Patients and Providers With Portals. PM R 2017 May;9(5S):S85-S97. [doi: 10.1016/j.pmrj.2017.02.005] [Medline: 28527507]
50. Nazi KM. The personal health record paradox: health care professionals' perspectives and the information ecology of personal health record systems in organizational and clinical settings. J Med Internet Res 2013 Apr 04;15(4):e70 [FREE Full text] [doi: 10.2196/jmir.2443] [Medline: 23557596]

Abbreviations

MNG-HA: Ministry of National Guard Health Affairs
PHR: personal health record
UTAUT: Unified Theory of Acceptance and Use of Technology
VIF: variance inflation factor

Edited by C Lovis; submitted 05.05.21; peer-reviewed by J Tavares; comments to author 29.05.21; revised version received 29.05.21; accepted 25.07.21; published 17.08.21.

Please cite as:
Yousef CC, Salgado TM, Farooq A, Burnett K, McClelland LE, Thomas A, Alenazi AO, Abu Esba LC, AlAzmi A, Alhameed AF, Hattan A, Elgadi S, Almekhloof S, AlShammary MA, Alanezi NA, Alhamdan HS, Khoshhal S, DeShazo JP
Predicting Patients' Intention to Use a Personal Health Record Using an Adapted Unified Theory of Acceptance and Use of Technology Model: Secondary Data Analysis
JMIR Med Inform 2021;9(8):e30214
URL: https://medinform.jmir.org/2021/8/e30214
doi:10.2196/30214
PMID:34304150

XSL•FO
RenderX

©Consuela Cheriece Yousef, Teresa M Salgado, Ali Farooq, Keisha Burnett, Laura E McClelland, Abin Thomas, Ahmed O Alenazi, Laila Carolina Abu Esba, Aeshah AlAzmi, Abrar Fahad Alhameed, Ahmed Hattan, Sumaya Elgadi, Saleh Almekhloof, Mohammed A AlShammary, Nazzal Abdullah Alanezi, Hani Solaiman Alhamdan, Sahal Khoshhal, Jonathan P DeShazo. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 17.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Usage Patterns of Web-Based Stroke Calculators in Clinical Decision Support: Retrospective Analysis

Benjamin Kummer[1,2], MD; Lubaina Shakir[3], MSc; Rachel Kwon[4], MD; Joseph Habboushe[3,5], MD, MBA; Nathalie Jetté[1,6], MD, MSc

[1]Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[2]Clinical Informatics, Mount Sinai Health System, New York, NY, United States
[3]MD Aware LLC, New York, NY, United States
[4]Ro, New York, NY, United States
[5]Department of Emergency Medicine, Weill Cornell Medicine, New York, NY, United States
[6]Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Corresponding Author:
Benjamin Kummer, MD
Department of Neurology
Icahn School of Medicine at Mount Sinai
One Gustave Levy Pl
Box 1137
New York, NY, 10029
United States
Phone: 1 2122415050
Email: benjamin.kummer@mountsinai.org

Abstract

Background: Clinical scores are frequently used in the diagnosis and management of stroke. While medical calculators are increasingly important support tools for clinical decisions, the uptake and use of common medical calculators for stroke remain poorly characterized.

Objective: We aimed to describe use patterns in frequently used stroke-related medical calculators for clinical decisions from a web-based support system.

Methods: We conducted a retrospective study of calculators from MDCalc, a web-based and mobile app–based medical calculator platform based in the United States. We analyzed metadata tags from MDCalc's calculator use data to identify all calculators related to stroke. Using relative page views as a measure of calculator use, we determined the 5 most frequently used stroke-related calculators between January 2016 and December 2018. For all 5 calculators, we determined cumulative and quarterly use, mode of access (eg, app or web browser), and both US and international distributions of use. We compared cumulative use in the 2016-2018 period with use from January 2011 to December 2015.

Results: Over the study period, we identified 454 MDCalc calculators, of which 48 (10.6%) were related to stroke. Of these, the 5 most frequently used calculators were the CHA_2DS_2-VASc score for atrial fibrillation stroke risk calculator (5.5% of total and 32% of stroke-related page views), the Mean Arterial Pressure calculator (2.4% of total and 14.0% of stroke-related page views), the HAS-BLED score for major bleeding risk (1.9% of total and 11.4% of stroke-related page views), the National Institutes of Health Stroke Scale (NIHSS) score calculator (1.7% of total and 10.1% of stroke-related page views), and the $CHADS_2$ score for atrial fibrillation stroke risk calculator (1.4% of total and 8.1% of stroke-related page views). Web browser was the most common mode of access, accounting for 82.7%-91.2% of individual stroke calculator page views. Access originated most frequently from the most populated regions within the United States. Internationally, use originated mostly from English-language countries. The NIHSS score calculator demonstrated the greatest increase in page views (238.1% increase) between the first and last quarters of the study period.

Conclusions: The most frequently used stroke calculators were the CHA_2DS_2-VASc, Mean Arterial Pressure, HAS-BLED, NIHSS, and $CHADS_2$. These were mainly accessed by web browser, from English-speaking countries, and from highly populated areas. Further studies should investigate barriers to stroke calculator adoption and the effect of calculator use on the application of best practices in cerebrovascular disease.

XSL•FO
RenderX

(JMIR Med Inform 2021;9(8):e28266) doi:10.2196/28266

KEYWORDS

medical informatics; clinical informatics; mhealth; digital health; cerebrovascular disease; medical calculators; health information; health information technology; information technology; economic health; clinical health; electronic health records

Introduction

Since the introduction of the Health Information Technology for Economic and Clinical Health Act in 2009, hospital systems in the United States have seen a five-fold increase in electronic health record (EHR) system adoptions [1,2]. These increases in EHR adoption have been accompanied by an upsurge in the amount of clinical data contained in EHRs. Providers' increasingly challenging task of managing this growing amount of information may result in cognitive burdening [3]. Moreover, the manner in which many EHRs display large amounts of clinical information may not support optimal cognitive reasoning [4]. Providers that use EHRs may therefore experience a number of unwanted adverse effects, including reductions in situational awareness, increases in mental workload, and reduced cognitive performance [5].

Clinical decision support (CDS) systems endeavor to enhance health care delivery by providing clinician-facing and patient-facing information that can improve decision-making at key steps in the workflow [6]. CDS systems are common in modern EHRs and range from passive banners to modal alerting systems for clinical conditions and adverse drug interactions [7,8]. Given that they are capable of delivering variably complex and tailored clinical content at the point of care [9], CDS systems are also well-suited for reducing cognitive overload. Medical calculators are specialized CDS instruments that incorporate user-entered clinical parameters to compute the discrete output of various types of functions [6,10], including physiological equations, risk stratification scores, and disease-quantifying or disability-quantifying scales. While medical calculators are increasingly prevalent in the growing armamentarium of CDS solutions available to providers, few studies have investigated their use patterns and barriers to adoption [5,10,11].

Stroke is a leading cause of disability and mortality worldwide, imposing a heavy economic and public health burden [12,13]. Several clinical scoring systems that draw on clinical, demographic, and laboratory parameters to predict risk, determine disease severity, or grade disability are widely available for the evaluation and management of stroke [14-28]. While medical calculators lend themselves naturally to such use cases, there is a lack of studies describing the current state of medical calculator use in stroke and cerebrovascular disease. Considering this and the need to better understand the adoption and use of medical calculators, we sought to study the use patterns of frequently used stroke calculators from a widely used web platform.

Methods

We conducted a retrospective, descriptive study of medical calculators published by MDCalc (MD Aware LLC, New York, NY, USA), a free, web-based and mobile app–based CDS platform that is used by over 65% of US-based physicians monthly and millions of clinicians worldwide [29]. MDCalc's CDS tools consist of medical score calculator forms for over 200 clinical conditions that allow users to input clinical variables and visualize clinical score outputs, along with an interpretation of the output and an appraisal of the available evidence supporting the use for each score (Multimedia Appendix 1) [6,29].

We used MDCalc's analytics platform to identify all calculators that were accessed between January 1, 2016 and December 31, 2018. We extracted calculator names; number of cumulative, nonunique page views; mode of access (eg, mobile app or web page); page view ranks; and calculator metadata, including launch dates and structured disease area categories (ie, "tags"). Page view ranks were assigned for each calculator based on total page views over the study period, with the lowest rank corresponding to the highest number of page views. Each calculator's cumulative page views were expressed relative to total cumulative page views for the entire MDCalc platform over the study period.

We defined calculators related to stroke as any calculator that contained 1 or more stroke-related tag (ie, "ischemic stroke," "transient ischemic attack," "intracerebral hemorrhage," or "subarachnoid hemorrhage"). For the 5 calculators with the highest relative page views over the study period, we determined quarterly page views, page views stratified by mode of access (eg, web page, iOS mobile app, or Android mobile app), country, and US state. For each calculator, we additionally determined page views relative to all stroke-related calculators and calculated the rate of increase in relative page views between the first and last quarter of the study period. To describe the evolution in stroke-related calculator use and rankings in the 5 years prior to the start of the study period, we determined relative page views and ranks for the same 5 calculators between January 1, 2011 and December 31, 2015. We then compared these measurements to those for the 2016-2018 study period. We only included calculators that were published by MDCalc.

Results

Between January 1, 2016 and December 31, 2018, we identified 454 MDCalc calculators, of which 48 (10.6%) were related to stroke. By cumulative page view, the 5 most highly ranked stroke calculators were the CHA_2DS_2-VASc (congestive heart failure, hypertension, 75 years of age and older, diabetes mellitus, previous stroke or transient ischemic attack, vascular disease, 65 to 74 years of age, female) score for atrial fibrillation stroke risk calculator (5.5% of total MDCalc and 32% of stroke-related page views), the Mean Arterial Pressure (MAP) calculator (2.4% of total MDCalc and 14% of stroke-related page views), the HAS-BLED (hypertension, abnormal renal/liver

XSL•FO
RenderX

function, stroke, bleeding history or predisposition, labile international normalized ratio, elderly, drugs/alcohol concomitantly) score for major bleeding risk calculator (1.9% of total MDCalc and 11.4% of stroke-related page views), the National Institutes of Health Stroke Scale (NIHSS) score calculator (1.7% of total MDCalc and 10.1% of stroke-related page views), and the $CHADS_2$ (congestive heart failure, hypertension, 75 years of age or older, diabetes mellitus, and previous stroke or transient ischemic attack) score for atrial fibrillation stroke risk calculator (1.4% of total MDCalc and 8.1% of stroke-related page views; Table 1).

Table 1. Relative page views and ranks of the 5 most frequently used MDCalc stroke calculators, 2011-2018.

			2011-2015[a]			2016-2018[b]		
Calculator	Description[c]	Launch date	Proportion of all calculator page views, %[d,e]	Proportion of stroke calculator page views, %[d,f]	Rank[g]	Proportion of all calculator page views, %[d,e]	Proportion of stroke calculator page views, %[h]	Rank[g]
CHA_2DS_2-VASc[i]	Calculates stroke risk for patients with atrial fibrillation, possibly better than the $CHADS_2$[j] score	April 1, 2011	4.9	38.7	2	5.5	32	2
MAP[k]	Calculates mean arterial pressure	January 1, 2009	1.1	8.6	31	2.4	14	7
HAS-BLED[l]	Estimates risk of major bleeding for patients on anticoagulation to assess risk-benefit in atrial fibrillation care	April 1, 2011	2.2	17.4	12	1.9	11.4	9
NIHSS[m]	Calculates the NIH[n] Stroke Scale for quantifying stroke severity	January 1, 2009	1.0	7.7	33	1.7	10.1	15
$CHADS_2$	Estimates stroke risk in patients with atrial fibrillation	January 1, 2009	2.9	22.6	7	1.4	8.1	22

[a]The 2011-2015 period is from January 1, 2011 to December 31, 2015.

[b]The 2016-2018 period is from January 1, 2016 to December 31, 2018.

[c]Descriptions are as appears on each MDCalc calculator webpage.

[d]All page views exclude Android/iOS MDCalc app page views.

[e]Percentage is relative to page views for all MDCalc calculators available during specified period.

[f]Percentage is relative to page views for 22 stroke-related calculators available during specified period.

[g]Rank is assigned according to cumulative, nonunique MDCalc page views relative to all available MDCalc calculator page views for each specified period; lowest rank corresponds to the highest proportion of page views.

[h]Percentage is relative to page views for 48 stroke-related calculators available during specified period.

[i]CHA_2DS_2-VASc: congestive heart failure, hypertension, 75 years of age and older, diabetes mellitus, previous stroke or transient ischemic attack, vascular disease, 65 to 74 years of age, female.

[j]$CHADS_2$: congestive heart failure, hypertension, 75 years of age or older, diabetes mellitus, and previous stroke or transient ischemic attack.

[k]MAP: mean arterial pressure.

[l]HAS-BLED: hypertension, abnormal renal/liver function, stroke, bleeding history or predisposition, labile international normalized ratio, elderly, drugs/alcohol concomitantly.

[m]NIHSS: National Institutes of Health Stroke Scale.

[n]NIH: National Institutes of Health.

Native English-language countries accounted for the highest proportion of page views for all calculators. Among individual countries, the United States, followed by the United Kingdom, accounted for the highest proportion of page views for all calculators except for the $CHADS_2$ score, for which Canada accounted for the second-highest proportion of page views. Within the United States, the states of California, Texas, New York, Pennsylvania, and Florida accounted for the highest proportion of page views for all calculators except the MAP score, for which Washington, California, Oregon, Texas, and

XSL•FO
RenderX

New York accounted for the greatest share. Among individual states, the highest proportion of page views originated from California for the CHA_2DS_2-VASc, NIHSS, and $CHADS_2$ scores, whereas the highest number of page views originated from New York for the HAS-BLED score and Washington for the MAP score. Use patterns for the NIHSS calculator are shown in Table 2, which shows similar use patterns as for the CHA_2DS_2-VASc, HAS-BLED, and $CHADS_2$ score calculators. The MAP calculator use pattern is represented separately in Table 3.

Table 2. Growth in relative page views of the National Institutes of Health Stroke Scale score calculator by quarter and year.

Quarter (year)	Proportion of total page views, %
Q1 (2016)	4.2
Q2 (2016)	4.3
Q3 (2016)	6.6
Q4 (2016)	4.7
Q1 (2017)	6.7
Q2 (2017)	6.5
Q3 (2017)	7
Q4 (2017)	8.8
Q1 (2018)	10.8
Q2 (2018)	12.2
Q3 (2018)	13.9
Q4 (2018)	14.2

Table 3. Growth in relative page views of the Mean Arterial Pressure score calculator by quarter and year.

Quarter (year)	Proportion of total page views, %
Q1 (2016)	5.1
Q2 (2016)	5.4
Q3 (2016)	8.5
Q4 (2016)	7
Q1 (2017)	7.2
Q2 (2017)	7.5
Q3 (2017)	8
Q4 (2017)	8.5
Q1 (2018)	9.9
Q2 (2018)	10
Q3 (2018)	10.6
Q4 (2018)	12

All 5 calculators were predominantly accessed by web browser rather than by mobile apps. The proportion of access attributable to web browsers varied depending on the specific calculator. However, web browser access accounted for 82.7%-91.2% of frequently used stroke calculator page views, with the NIHSS and MAP calculators respectively representing the minimum and maximum in the range. The NIHSS calculator had the highest proportion of Android app page views (10.7%). Two calculators, the NIHSS and CHA2DS2-VASc, generated the highest and equal proportion of iOS app pageviews (6.6%) (data not shown). The NIHSS score calculator demonstrated the greatest increase in page views (238.1% increase) between the first and last quarters of the study period (Table 2).

All 5 calculators were released by MDCalc between January 2009 and April 2011. In chronological order, the $CHADS_2$ score and MAP calculators were released the earliest (January 1, 2009), followed by the NIHSS calculator (January 1, 2011) and the HAS-BLED and CHA_2DS_2-VASc score calculators (April 1, 2011). Over the study period, the CHA_2DS_2-VASc score calculator was ranked 2nd; MAP, 7th; HAS-BLED, 9th; NIHSS, 15th; and $CHADS_2$, 22nd. By contrast, between January 2011 and December 2016, the corresponding ranks for these calculators were 2nd, 31st, 12th, 33rd, and 7th, respectively (Table 1).

XSL•FO
RenderX

Discussion

Principal Findings

In this study, we found that the most frequently accessed calculators relating to stroke comprised 1 of 3 types: risk prediction tools for complications that were conditional on the presence of a specific disease state (eg, $CHADS_2$, HAS-BLED, and CHA_2DS_2-VASc scores), scales to quantify severity in ischemic stroke (eg, NIHSS), and calculators for computing physiologic parameters (eg, MAP). These calculators were among the most frequently used calculators on the MDCalc platform, as demonstrated by the CHA2DS2-VASc score calculator that ranked second by relative page views in both the 2011-2015 and 2016-2018 periods and by the increases in ranks observed in all stroke calculators during the 2016-2018 period. The majority of the calculators were accessed from the most highly populated US states [30] with the greatest number of licensed physicians [31]. While a number of page views did originate from outside the United States, most of these, nonetheless, originated from English-language countries.

Characteristics of Highly Used Stroke Calculators

English-Language Dominance and Association With High-Prevalence Conditions

Many drivers of stroke calculator use that we uncovered in our analysis may also be generalizable features of highly used calculators outside the field of stroke. One primary such driver may be the predominance of the English language, which is best exemplified by our findings that the highest rates of geographical calculator use originated in English-language countries. However, potential additional factors contributing to the predominance of English in calculator use include the widespread use of English in scientific and clinical communities worldwide [32], the fact that MDCalc has an English-only website [29] and was founded by 2 US emergency medicine physicians, and the platform's primarily word-of-mouth advertising strategy in English-language countries. A second potentially generalizable feature of highly used calculators is high disease prevalence. Our findings demonstrate that 3 of the 5 (60%) most highly used calculators related to atrial fibrillation, which is both highly prevalent in elderly patients [33] as well as patients with ischemic stroke [34]. As suggested by our findings, calculators addressing highly prevalent diseases may be likely to generate higher use.

Inclusion in Professional Society Guidelines

A third potentially generalizable feature of calculators is their inclusion of corresponding scores in professional society guidelines, as shown in our study by both CHA_2DS_2-VASc and HAS-BLED. The former score was incorporated into US and international professional society guidelines for the management of atrial fibrillation, including the European Society of Cardiovascular in 2012 and 2016 [35,36], the American Heart Association in 2014 [37], the National Institute for Health and Care Excellence United Kingdom guidelines in 2014 [38], and the Asia Pacific Heart Rhythm Society guidelines in 2017 [39]. Similarly, the HAS-BLED score was incorporated in European Society of Cardiovascular in 2012 and 2016 [35,36], the Canadian Cardiovascular Society in 2014 and 2018 [40,41], and the National Institute for Health and Care Excellence United Kingdom guidelines in 2014 [38]. Relatedly, evidence suggests that the predictive ability of the HAS-BLED score outperformed that of other hemorrhage risk scores [42], which may have also solidified this score's position in multiple society guidelines.

Updates to Widely Used Score Calculators

A fourth factor associated with high calculator popularity may be the use of calculators for clinical scores that constitute an update to an already existing high-profile clinical score. In our study, this is best exemplified by the CHA_2DS_2-VASc score, which was responsible for nearly one-third of stroke-related calculator page views between 2016 and 2018. This score was originally developed as a risk stroke prediction tool in atrial fibrillation that was improved compared with the existing $CHADS_2$ score by incorporating several additional thromboembolic risk factors [17]. Dating back to the original score's publication in Journal of the American Medical Association in 2001, practicing clinicians may have already been familiar with the concept of data-driven stroke risk prediction in atrial fibrillation by the time of the second score's publication in 2009. This familiarity, in turn, may have cemented widespread acceptance of the CHA_2DS_2-VASc score's viability as a clinical risk predictor.

Broad Applicability to Nonstroke Conditions

Applicability of calculators to multiple disease states may be additionally responsible for widespread use. For instance, we found that the second-most used cerebrovascular calculator was the MAP, which rose in relative page views between the 5-year period ending on December 31, 2015 and the end of the 3-year study period. Although MAP is often used to guide management of aneurysmal subarachnoid hemorrhage [43], our findings are likely attributable to the usefulness of MAP in diagnosing and managing several nonstroke states, such as sepsis, septic shock [44], and neurotrauma [45]. Indeed, in addition to subarachnoid hemorrhage, MDCalc metadata tags for the MAP calculator include both "sepsis" and "trauma." Considering that severe sepsis and septic shock have higher yearly incidence than subarachnoid hemorrhage [46,47], the usefulness of MAP in the management of sepsis, rather than subarachnoid hemorrhage, may have been a more likely explanation for the high use of the MAP calculator during the study period. MAP is also less commonly used than systolic and diastolic blood pressure to guide the management of acute ischemic stroke [48,49] and intracerebral hemorrhage [50], thereby further supporting the theory that noncerebrovascular use cases were likely to be the primary drivers of high MAP calculator page views.

Score Use in High-Profile Randomized Trials

Inclusion of scores in high-profile randomized trials may also translate to high use of calculators associated with these scores. While the NIHSS score is not the sole factor in selecting patients for tissue plasminogen activator in acute ischemic stroke [49], the NIHSS was included in the first randomized controlled trial of tissue plasminogen activator for acute ischemic stroke [51] and incorporated as an inclusion criterion for several large randomized controlled trials demonstrating the effectiveness of

XSL•FO
RenderX

mechanical thrombectomy for acute ischemic stroke due to anterior circulation large-vessel arterial occlusion [52-55], along with several confirmatory meta-analyses in 2015 and 2016 [56,57].

Factors other than guideline adoptions and validations for study publications may also explain the patterns we observed in our study, such as increased global use of medical calculators and increased popularity of the MDCalc service across all calculators. These factors remain difficult to measure. In addition, several health care institutions across the world already use internal calculator repositories for clinical care, which are variably integrated into institutional EHRs. While the worldwide extent of this practice remains poorly characterized, increasing prevalence of such repositories in the future is likely to reduce clinician reliance on and use of external calculators.

Duration of Calculator Availability

Calculators that are released earlier may also be more widely employed than more recently released calculators due to increased awareness or ongoing search engine optimization. In this study, incorporation into society guidelines may be the main factor explaining why CHA_2DS_2-VASc and HAS-BLED calculators were released the latest, yet demonstrated higher use than calculators that were released earlier, such as the MAP, $CHADS_2$ and NIHSS. However, the unmistakable presence of calculators such as the MAP and NIHSS among the 5 most highly used stroke calculators may be a result of their earlier release dates.

Accessibility via Web Browser

Finally, our findings in stroke calculators suggest that web-accessible calculators may be more widely used than those that are primarily mobile app–based. These results are interesting, given that smartphone ownership in the United States has significantly increased since the early 2010s [58] and smartphone-based and tablet-based calculators are uniquely well suited to clinicians' flexible and dynamic workflow requirements. However, MDCalc's introduction of mobile apps in March 2016 (iOS) [59] and April 2017 (Android) should also be considered when interpreting our results [60]. Moreover, a significant proportion of the predominant web access we observed in our results may have occurred through mobile web browsers, which are highly prevalent in mobile devices and function identically to those found in stationary (eg, laptop or desktop) computers. However, because this study could not differentiate these different types of web access or the context in which these calculators were used, our findings cannot allow us to make definitive conclusions regarding the optimal mode or setting for stroke calculator deployment.

Limitations

This study was limited by several factors. First, we restricted our analysis to calculators from a single platform. Because many other web-based CDS platforms are available for use, our results may not generalize to other platforms or to the entire community of medical professionals that actively use the 5 identified stroke-related scores in day-to-day practice. Second, because we used deidentified page view data for the study, we lacked user information that could permit a more detailed understanding of calculator use, such as discipline, medical speciality, level of training, as well as EHR, care setting, and disease states in which stroke-related calculators were used. For similar reasons, we have limited insight into whether MDCalc calculator use was potentially affected by alternative calculators embedded in care providers' EHRs. Third, we did not investigate the effects these calculators, as CDS tools, had on aspects of clinician decision-making, such as diagnostic speed and accuracy, as studied by Abedin and colleagues [61]. We also did not investigate the relationship between calculator use and adherence to best practices or meaningful clinical outcomes. Finally, our study period was restricted to 3 years, which may have provided limited insights on use patterns and impacts on clinical care, especially as smartphone and mobile app usage have only become more ubiquitous since 2018.

Conclusions

In this retrospective analysis, we demonstrated that the most commonly used stroke calculators were related to secondary stroke prevention in atrial fibrillation, blood pressure measurement, and computation of the NIHSS score. As medical calculators become increasingly important CDS tools, further studies should seek to understand optimal implementation and integration of these calculators into EHR systems and clinical care pathways. This can be achieved by incorporating a broader spectrum of calculator platforms, including platforms for user specialty and training and analyses of the behavior of clinicians during calculator use at the point of care. Additionally, considering our findings that stroke calculators were predominantly adopted in English-speaking countries and highly populated areas, further studies should aim to investigate barriers to adoption and whether translation of calculators into non-English languages may potentially improve calculator adoption.

Authors' Contributions

BK conceptualized the study, drafted the manuscript, analyzed and interpreted study data, and revised the manuscript for intellectual content. LS obtained and analyzed data and revised the manuscript for intellectual content. RK conceptualized the study, obtained and analyzed data, and revised the manuscript for intellectual content. JH analyzed data and revised the manuscript for intellectual content. NJ revised the manuscript for intellectual content.

Conflicts of Interest

JH is the cofounder and owner of MD Aware LLC. LS is a full-time employee of MD Aware LLC. RK is a full-time of employee of Ro. BK serves on the advisory board of and owns equity in Syntrillo LLC. NJ is the Bludhorn Professor of International

XSL•FO
RenderX

Medicine at the Icahn School of Medicine at Mount Sinai. She receives grant funding paid to her institution for grants unrelated to this work from NINDS (NIH U24NS107201, NIH IU54NS100064) and PCORI. She receives an honorarium for her work as an Associate Editor of Epilepsia.

Multimedia Appendix 1
Example of an MDCalc medical calculator webpage (CHA₂DS₂-VASc Score for Atrial Fibrillation Stroke Risk).
[PNG File , 76 KB - medinform_v9i8e28266_app1.png]

References

1. Gold M, Mc L. Assessing HITECH Implementation and Lessons: 5 Years Later. Milbank Q 2016 Sep;94(3):654-687 [FREE Full text] [doi: 10.1111/1468-0009.12214] [Medline: 27620687]
2. Adler-Milstein J, Holmgren A, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. J Am Med Inform Assoc 2017 Nov 01;24(6):1142-1148 [FREE Full text] [doi: 10.1093/jamia/ocx080] [Medline: 29016973]
3. Farri O, Pieckiewicz D, Rahman A, Adam T, Pakhomov S, Melton G. A qualitative analysis of ehr clinical document synthesis by clinicians. 2012 Presented at: Proceedings of the American Medical Informatics Association Annual Symposium; . PMC3540510; 2012; Chicago, Illinois.
4. Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records. J Am Med Inform Assoc 2015 Sep;22(5):938-947 [FREE Full text] [doi: 10.1093/jamia/ocv032] [Medline: 25882031]
5. Aakre C, Dziadzko M, Keegan M, Herasevich V. Automating Clinical Score Calculation within the Electronic Health Record. Appl Clin Inform 2017 Dec 21;08(02):369-380. [doi: 10.4338/aci-2016-09-ra-0149]
6. Dorner S, Yun B, Kwon R, Habboushe J, Raja A. Characteristics of frequently used clinical decision support tools. Physician Leadership Journal 2018 Nov 11;5(6):62-66.
7. Bubp J, Park M, Kapusnik-Uner J, Dang T, Matuszewski K, Ly D, et al. Successful deployment of drug-disease interaction clinical decision support across multiple Kaiser Permanente regions. J Am Med Inform Assoc 2019 Oct 01;26(10):905-910 [FREE Full text] [doi: 10.1093/jamia/ocz020] [Medline: 30986823]
8. Devarakonda MV, Mehta N, Tsou C, Liang JJ, Nowacki AS, Jelovsek JE. Automated problem list generation and physicians perspective from a pilot study. Int J Med Inform 2017 Sep;105:121-129 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.05.015] [Medline: 28750905]
9. Habboushe J, Altman C, Lip GYH. Time trends in use of the CHADS and CHA DS VASc scores, and the geographical and specialty uptake of these scores from a popular online clinical decision tool and medical reference. Int J Clin Pract 2019 Feb 30;73(2):e13280. [doi: 10.1111/ijcp.13280] [Medline: 30281876]
10. Green T, Whitt S, Belden J, Erdelez S, Shyu C. Medical calculators: Prevalence, and barriers to use. Comput Methods Programs Biomed 2019 Oct;179:105002 [FREE Full text] [doi: 10.1016/j.cmpb.2019.105002] [Medline: 31443857]
11. Green T, Shyu C. Developing a Taxonomy of Online Medical Calculators for Assessing Automatability and Clinical Efficiency Improvements. Stud Health Technol Inform 2019 Aug 21;264:601-605. [doi: 10.3233/SHTI190293] [Medline: 31437994]
12. Tong X, Yang Q, Ritchey MD, George MG, Jackson SL, Gillespie C, et al. The Burden of Cerebrovascular Disease in the United States. Prev Chronic Dis 2019 Apr 25;16:180411 [FREE Full text] [doi: 10.5888/pcd16.180411] [Medline: 31022369]
13. Feigin V, Krishnamurthi R, Parmar P, Norrving B, Mensah GA, Bennett DA, et al. Update on the Global Burden of Ischemic and Hemorrhagic Stroke in 1990-2013: The GBD 2013 Study. Neuroepidemiology 2015;45(3):161-176 [FREE Full text] [doi: 10.1159/000441085] [Medline: 26505981]
14. Flint A, Cullen S, Faigeles B, Rao V. Predicting long-term outcome after endovascular stroke treatment: the totaled health risks in vascular events score. AJNR Am J Neuroradiol 2010 Aug;31(7):1192-1196 [FREE Full text] [doi: 10.3174/ajnr.A2050] [Medline: 20223889]
15. Ntaios G, Faouzi M, Ferrari J, Lang W, Vemmos K, Michel P. An integer-based score to predict functional outcome in acute ischemic stroke: the ASTRAL score. Neurology 2012 Jun 12;78(24):1916-1922. [doi: 10.1212/WNL.0b013e318259e221] [Medline: 22649218]
16. Singer D, Chang Y, Borowsky L, Fang MC, Pomernacki NK, Udaltsova N, et al. A new risk scheme to predict ischemic stroke and other thromboembolism in atrial fibrillation: the ATRIA study stroke risk score. J Am Heart Assoc 2013 Jun 21;2(3):e000250 [FREE Full text] [doi: 10.1161/JAHA.113.000250] [Medline: 23782923]
17. Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns HJGM. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest 2010 Feb;137(2):263-272. [doi: 10.1378/chest.09-1584] [Medline: 19762550]
18. Gage B, Waterman A, Shannon W, Boechler M, Rich M, Radford M. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. JAMA 2001 Jun 13;285(22):2864-2870. [doi: 10.1001/jama.285.22.2864] [Medline: 11401607]

XSL•FO
RenderX

19. Fisher CM, Kistler JP, Davis JM. Relation of cerebral vasospasm to subarachnoid hemorrhage visualized by computerized tomographic scanning. Neurosurgery 1980 Jan;6(1):1-9. [doi: 10.1227/00006123-198001000-00001] [Medline: 7354892]
20. Johnston SC, Rothwell PM, Nguyen-Huynh MN, Giles MF, Elkins JS, Bernstein AL, et al. Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack. The Lancet 2007 Jan;369(9558):283-292. [doi: 10.1016/s0140-6736(07)60150-0]
21. Hemphill J, Bonovich DC, Besmertis L, Manley GT, Johnston SC. The ICH score: a simple, reliable grading scale for intracerebral hemorrhage. Stroke 2001 Apr;32(4):891-897. [doi: 10.1161/01.str.32.4.891] [Medline: 11283388]
22. Delgado Almandoz JE, Schaefer P, Goldstein J, Rosand J, Lev MH, González RG, et al. Practical scoring system for the identification of patients with intracerebral hemorrhage at highest risk of harboring an underlying vascular etiology: the Secondary Intracerebral Hemorrhage Score. AJNR Am J Neuroradiol 2010 Oct;31(9):1653-1660 [FREE Full text] [doi: 10.3174/ajnr.A2156] [Medline: 20581068]
23. Strbian D, Engelter S, Michel P, Meretoja A, Sekoranja L, Ahlhelm FJ, et al. Symptomatic intracranial hemorrhage after stroke thrombolysis: the SEDAN score. Ann Neurol 2012 May;71(5):634-641. [doi: 10.1002/ana.23546] [Medline: 22522478]
24. Myint PK, Clark AB, Kwok CS, Davis J, Durairaj R, Dixit AK, et al. The SOAR (Stroke subtype, Oxford Community Stroke Project classification, Age, prestroke modified Rankin) score strongly predicts early outcomes in acute stroke. Int J Stroke 2014 Apr 09;9(3):278-283. [doi: 10.1111/ijs.12088] [Medline: 23834262]
25. Pisters R, Lane DA, Nieuwlaat R, de Vos CB, Crijns HJGM, Lip GYH. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey. Chest 2010 Nov;138(5):1093-1100. [doi: 10.1378/chest.10-0134] [Medline: 20299623]
26. Hunt W, Hess R. Surgical risk as related to time of intervention in the repair of intracranial aneurysms. J Neurosurg 1968 Jan;28(1):14-20. [doi: 10.3171/jns.1968.28.1.0014] [Medline: 5635959]
27. Kernan WN, Ovbiagele B, Black HR, Bravata DM, Chimowitz MI, Ezekowitz MD, et al. Guidelines for the prevention of stroke in patients with stroke and transient ischemic attack: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2014 Jul;45(7):2160-2236. [doi: 10.1161/STR.0000000000000024] [Medline: 24788967]
28. American Heart Association. Secondary stroke prevention checklist. Secondary stroke prevention checklist. 2019. URL: https://www.stroke.org/-/media/stroke-files/stroke-resource-center/recovery/patient-focused/secondary-stroke-prevention-checklist.pdf?la=en [accessed 2020-08-05]
29. MDCalc. MDCalc - Medical calculators, equations, scores, and guidelines. MDCalc website. 2019. URL: https://www.mdcalc.com [accessed 2019-09-14]
30. United States Census Bureau, Population Division. Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2018 (nst-est2018-01). United States Census 2018 National and State Population Estimates. 2019. URL: http://www2.census.gov/programs-surveys/popest/tables/2010-2018/national/totals/na-est2018-01.xlsx [accessed 2019-11-19]
31. Young A, Chaudhry H, Pei X, Arnhart K, Dugan M, Steingard S. Federation of State Medical Boards Census of licensed physicians in the united states, 2018. J Med Reg 2019;105(2):7-23 [FREE Full text] [doi: 10.30770/2572-1852-105.2.7]
32. Roca A, Boum Y, Wachsmuth I. Plaidoyer contre l'exclusion des francophones dans la recherche en santé mondiale. The Lancet Global Health 2019 Jun;7(6):e701-e702. [doi: 10.1016/s2214-109x(19)30175-5]
33. Go A, Hylek E, Phillips K, Chang Y, Henault LE, Selby JV, et al. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study. JAMA 2001 May 09;285(18):2370-2375. [doi: 10.1001/jama.285.18.2370] [Medline: 11343485]
34. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. Stroke 1991 Aug;22(8):983-988. [doi: 10.1161/01.str.22.8.983] [Medline: 1866765]
35. Camm A, Lip G, De Caterina R, Savelieva I, Atar D, Hohnloser SH, et al. 2012 focused update of the ESC Guidelines for the management of atrial fibrillation: an update of the 2010 ESC Guidelines for the management of atrial fibrillation. Developed with the special contribution of the European Heart Rhythm Association. Eur Heart J 2012 Nov;33(21):2719-2747. [Medline: 22922413]
36. Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B, et al. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. Eur Heart J 2016 Oct 07;37(38):2893-2962. [doi: 10.1093/eurheartj/ehw210] [Medline: 27567408]
37. January CT, Wann LS, Alpert JS, Calkins H, Cigarroa JE, Cleveland JC, American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. J Am Coll Cardiol 2014 Dec 02;64(21):e1-76 [FREE Full text] [doi: 10.1016/j.jacc.2014.03.022] [Medline: 24685669]
38. Jones C, Pollit V, Fitzmaurice D, Cowan C, Guideline Development Group. The management of atrial fibrillation: summary of updated NICE guidance. BMJ 2014 Jun 19;348:g3655. [doi: 10.1136/bmj.g3655] [Medline: 24948694]

XSL•FO
RenderX

39. Chiang C, Okumura K, Zhang S, Chao T, Siu C, Wei Lim T, et al. 2017 consensus of the Asia Pacific Heart Rhythm Society on stroke prevention in atrial fibrillation. J Arrhythm 2017 Aug;33(4):345-367 [FREE Full text] [doi: 10.1016/j.joa.2017.05.004] [Medline: 28765771]
40. Verma A, Cairns JA, Mitchell LB, Macle L, Stiell IG, Gladstone D, CCS Atrial Fibrillation Guidelines Committee. 2014 focused update of the Canadian Cardiovascular Society Guidelines for the Management of atrial fibrillation. Can J Cardiol 2014 Oct;30(10):1114-1130. [doi: 10.1016/j.cjca.2014.08.001] [Medline: 25262857]
41. Andrade JG, Verma A, Mitchell LB, Parkash R, Leblanc K, Atzema C, et al. 2018 Focused Update of the Canadian Cardiovascular Society Guidelines for the Management of Atrial Fibrillation. Can J Cardiol 2018 Nov;34(11):1371-1392. [doi: 10.1016/j.cjca.2018.08.026] [Medline: 30404743]
42. Apostolakis S, Lane D, Guo Y, Buller H, Lip G. Performance of the HEMORR(2)HAGES, ATRIA, and HAS-BLED bleeding risk-prediction scores in patients with atrial fibrillation undergoing anticoagulation: the AMADEUS (evaluating the use of SR34006 compared to warfarin or acenocoumarol in patients with atrial fibrillation) study. J Am Coll Cardiol 2012 Aug 28;60(9):861-867 [FREE Full text] [doi: 10.1016/j.jacc.2012.06.019] [Medline: 22858389]
43. Diringer M, Bleck T, Claude Hemphill J, Menon D, Shutter L, Vespa P, et al. Critical care management of patients following aneurysmal subarachnoid hemorrhage: recommendations from the Neurocritical Care Society's Multidisciplinary Consensus Conference. Neurocrit Care 2011 Sep;15(2):211-240. [doi: 10.1007/s12028-011-9605-9] [Medline: 21773873]
44. Dellinger RP, Levy MM, Rhodes A, Annane D, Gerlach H, Opal SM, Surviving Sepsis Campaign Guidelines Committee including the Pediatric Subgroup. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. Crit Care Med 2013 Feb;41(2):580-637. [doi: 10.1097/CCM.0b013e31827e83af] [Medline: 23353941]
45. Long B, Koyfman A. Secondary Gains: Advances in Neurotrauma Management. Emerg Med Clin North Am 2018 Feb;36(1):107-133. [doi: 10.1016/j.emc.2017.08.007] [Medline: 29132572]
46. Mouncey PR, Osborn TM, Power GS, Harrison DA, Sadique MZ, Grieve RD, et al. Trial of Early, Goal-Directed Resuscitation for Septic Shock. N Engl J Med 2015 Apr 02;372(14):1301-1311. [doi: 10.1056/nejmoa1500896]
47. de Rooij NK, Linn F, van der Plas JA, Algra A, Rinkel G. Incidence of subarachnoid haemorrhage: a systematic review with emphasis on region, age, gender and time trends. J Neurol Neurosurg Psychiatry 2007 Dec;78(12):1365-1372 [FREE Full text] [doi: 10.1136/jnnp.2007.117655] [Medline: 17470467]
48. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, et al. 2018 Guidelines for the Early Management of Patients With Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. Stroke 2018 Mar;49(3):e46-e110. [doi: 10.1161/STR.0000000000000158] [Medline: 29367334]
49. Jauch EC, Saver JL, Adams HP, Bruno A, Connors JJB, Demaerschalk BM, et al. Guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2013 Mar;44(3):870-947 [FREE Full text] [doi: 10.1161/STR.0b013e318284056a] [Medline: 23370205]
50. Hemphill JC, Greenberg SM, Anderson CS, Becker K, Bendok BR, Cushman M, et al. Guidelines for the Management of Spontaneous Intracerebral Hemorrhage. Stroke 2015 May 28;46(7):2032-2060. [doi: 10.1161/str.0000000000000069]
51. National Institute of Neurological DisordersStroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. N Engl J Med 1995 Dec 14;333(24):1581-1587. [doi: 10.1056/NEJM199512143332401] [Medline: 7477192]
52. Berkhemer OA, Fransen PS, Beumer D, van den Berg LA, Lingsma HF, Yoo AJ, et al. A randomized trial of intraarterial treatment for acute ischemic stroke. N Engl J Med 2015 Jan 01;372(1):11-20. [doi: 10.1056/NEJMoa1411587] [Medline: 25517348]
53. Campbell BCV, Mitchell PJ, Kleinig TJ, Dewey HM, Churilov L, Yassi N, et al. Endovascular therapy for ischemic stroke with perfusion-imaging selection. N Engl J Med 2015 Mar 12;372(11):1009-1018. [doi: 10.1056/NEJMoa1414792] [Medline: 25671797]
54. Molina C, Chamorro A, Rovira A, de Miquel A, Serena J, Roman LS, et al. REVASCAT: a randomized trial of revascularization with SOLITAIRE FR device vs. best medical therapy in the treatment of acute stroke due to anterior circulation large vessel occlusion presenting within eight-hours of symptom onset. Int J Stroke 2015 Jun;10(4):619-626. [doi: 10.1111/ijs.12157] [Medline: 24206399]
55. Saver JL, Goyal M, Bonafe A, Diener H, Levy EI, Pereira VM, et al. Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke. N Engl J Med 2015 Jun 11;372(24):2285-2295. [doi: 10.1056/NEJMoa1415061] [Medline: 25882376]
56. Badhiwala J, Nassiri F, Alhazzani W, Selim MH, Farrokhyar F, Spears J, et al. Endovascular Thrombectomy for Acute Ischemic Stroke: A Meta-analysis. JAMA 2015 Nov 03;314(17):1832-1843. [doi: 10.1001/jama.2015.13767] [Medline: 26529161]
57. Goyal M, Menon BK, van Zwam WH, Dippel DWJ, Mitchell PJ, Demchuk AM, et al. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. Lancet 2016 Apr 23;387(10029):1723-1731. [doi: 10.1016/S0140-6736(16)00163-X] [Medline: 26898852]

XSL•FO
RenderX

58. Demographics of mobile phone ownership and adoption in the united states. Pew Research Center Mobile Fact Sheet. 2019. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2020-09-21]
59. Maurer D. MDCalc app, the best online medical calculator, is now an app. iMedicalApps. URL: https://www.imedicalapps.com/2016/03/mdcalc-medical-calculator-app/ [accessed 2021-02-15]
60. Husain I. MDCalc is finally available for Android. MD Tech Tips. URL: https://www.imedicalapps.com/2017/04/md-tech-tips-mdcalc/ [accessed 2021-02-15]
61. Abedin Z, Hoerner R, Kawamoto K. Evaluation of a FHIR-based Clinical Decision Support Tool for Calculating CHA2Ds2-VAsc scores. In: Proceedings of the Circulation Cardiovascular Quality and Outcomes Scientific Sessions. Dallas, Texas: American heart Association; 2019 Presented at: Cardiovascular Quality and Outcomes Scientific Sessions; 2019; Arlington, Virginia p. 5-6.

Abbreviations

CDS: clinical decision support.
$CHADS_2$: congestive heart failure, hypertension, 75 years of age or older, diabetes mellitus, and previous stroke or transient ischemic attack.
CHA_2DS_2-VASc: congestive heart failure, hypertension, 75 years of age and older, diabetes mellitus, previous stroke or transient ischemic attack, vascular disease, 65 to 74 years of age, female.
EHR: electronic health record.
HAS-BLED: hypertension, abnormal renal/liver function, stroke, bleeding history or predisposition, labile international normalized ratio, elderly, drugs/alcohol concomitantly.
MAP: mean arterial pressure.
NIHSS: National Institutes of Health Stroke Scale.

Edited by C Lovis; submitted 10.03.21; peer-reviewed by A Nowacki, L Chirchir; comments to author 14.04.21; revised version received 24.05.21; accepted 05.06.21; published 02.08.21.

Please cite as:
Kummer B, Shakir L, Kwon R, Habboushe J, Jetté N
Usage Patterns of Web-Based Stroke Calculators in Clinical Decision Support: Retrospective Analysis
JMIR Med Inform 2021;9(8):e28266
URL: https://medinform.jmir.org/2021/8/e28266
doi:10.2196/28266
PMID:34338647

©Benjamin Kummer, Lubaina Shakir, Rachel Kwon, Joseph Habboushe, Nathalie Jetté. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 02.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Quality of Hospital Electronic Health Record (EHR) Data Based on the International Consortium for Health Outcomes Measurement (ICHOM) in Heart Failure: Pilot Data Quality Assessment Study

Hannelore Aerts[1,2], PhD; Dipak Kalra[1,2], MD, PhD; Carlos Sáez[3], PhD; Juan Manuel Ramírez-Anguita[4], PhD; Miguel-Angel Mayer[4], MD, PhD, MPH; Juan M Garcia-Gomez[3], PhD; Marta Durà-Hernández[3], MSc; Geert Thienpont[2,5], BSc; Pascal Coorevits[1], PhD

[1]Medical Informatics and Statistics Unit, Department of Public Health and Primary Care, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium

[2]The European Institute for Innovation through Health Data (i~HD), Ghent, Belgium

[3]Biomedical Data Science Lab, Instituto Universitario de Tecnologías de la Información y Comunicaciones, Universitat Politècnica de València, Valencia, Spain

[4]Research Programme on Biomedical Informatics, Hospital del Mar Medical Research Institute and Universitat Pompeu Fabra, Barcelona, Spain

[5]Research in Advanced Medical Informatics and Telematics (RAMIT), Ghent, Belgium

Corresponding Author:
Miguel-Angel Mayer, MD, PhD, MPH
Research Programme on Biomedical Informatics
Hospital del Mar Medical Research Institute and Universitat Pompeu Fabra
C/ Dr Aiguader 88
Barcelona, 08003
Spain
Phone: 34 933 160 539
Email: miguelangel.mayer@upf.edu

Abstract

Background: There is increasing recognition that health care providers need to focus attention, and be judged against, the impact they have on the health outcomes experienced by patients. The measurement of health outcomes as a routine part of clinical documentation is probably the only scalable way of collecting outcomes evidence, since secondary data collection is expensive and error-prone. However, there is uncertainty about whether routinely collected clinical data within electronic health record (EHR) systems includes the data most relevant to measuring and comparing outcomes and if those items are collected to a good enough data quality to be relied upon for outcomes assessment, since several studies have pointed out significant issues regarding EHR data availability and quality.

Objective: In this paper, we first describe a practical approach to data quality assessment of health outcomes, based on a literature review of existing frameworks for quality assessment of health data and multistakeholder consultation. Adopting this approach, we performed a pilot study on a subset of 21 International Consortium for Health Outcomes Measurement (ICHOM) outcomes data items from patients with congestive heart failure.

Methods: All available registries compatible with the diagnosis of heart failure within an EHR data repository of a general hospital (142,345 visits and 12,503 patients) were extracted and mapped to the ICHOM format. We focused our pilot assessment on 5 commonly used data quality dimensions: completeness, correctness, consistency, uniqueness, and temporal stability.

Results: We found high scores (>95%) for the consistency, completeness, and uniqueness dimensions. Temporal stability analyses showed some changes over time in the reported use of medication to treat heart failure, as well as in the recording of past medical conditions. Finally, the investigation of data correctness suggested several issues concerning the characterization of missing data values. Many of these issues appear to be introduced while mapping the IMASIS-2 relational database contents to the ICHOM format, as the latter requires a level of detail that is not explicitly available in the coded data of an EHR.

Conclusions: Overall, results of this pilot study revealed good data quality for the subset of heart failure outcomes collected at the Hospital del Mar. Nevertheless, some important data errors were identified that were caused by fundamentally different data

XSL•FO
RenderX

collection practices in routine clinical care versus research, for which the ICHOM standard set was originally developed. To truly examine to what extent hospitals today are able to routinely collect the evidence of their success in achieving good health outcomes, future research would benefit from performing more extensive data quality assessments, including all data items from the ICHOM standards set and across multiple hospitals.

(JMIR Med Inform 2021;9(8):e27842) doi:10.2196/27842

KEYWORDS

data quality; electronic health records; heart failure; value-based health insurance; patient outcome assessment

Introduction

Increasing quantities of health data are being collected across care organizations, creating a powerful opportunity to learn from these data how to improve patient care and accelerate research. The earliest call to action and formalized approach for using health data to assess quality of care was probably the Donabedian model of quality [1]. He categorized the assessment of health care quality under structure (how services are organized and resourced), process (how care is delivered and what care activities are undertaken), and outcome (what health impact it has). Over the decades, it has proved much easier to develop and implement audits of structure or process, but formalized assessments of outcome appear to be more challenging because it is harder to define what we mean by outcomes and how best to measure them [2]. A formalized approach to measuring health outcomes was proposed by Porter and Teisberg [3], within their model of the assessment of "value" in a seminal publication in 2006. Within this value equation, outcomes were defined as "the outcomes that matter to patients and the costs to achieve those outcomes" [3]. This "Value-Based Health Care" model has grown into a portfolio of health outcomes standards for measuring value, developed and promoted by the International Consortium for Health Outcomes Measurement (ICHOM). These health outcomes standards, formalized as indicators to be collected, quantified, and compared between health care providers, have stimulated a global interest in benchmarking and comparing health outcomes [4].

All these models hinge upon the essential ability to measure health, health care, and its outcomes. Health data are therefore a vital ingredient. To enable accurate measurement, data have to be captured and represented to a high quality. Unreliable data, such as incomplete, incorrect, or missing data entries, will inevitably lead to biased analyses, resulting in misdirected efforts to improve quality or false research interpretations.

Yet, several studies have pointed out significant issues regarding availability and quality of electronic health record (EHR) data [5-10]. For example, the "Electronic Health Records for Clinical Research" project, funded by the Innovative Medicine Initiative, clearly demonstrated that many variables, among which even fundamental ones such as patient weight, are frequently not present within EHR systems [8]. Incorrect or absent recording of patient weights, though, can lead to medication dosage errors. Hirata and colleagues [11] examined the frequency and consequences of weight errors that occurred across 79,000 emergency department encounters of children under the age of 5 years. They revealed that, although weight errors were relatively rare (0.63%), a large proportion of weight errors led to subsequent medication-dosing errors (34%). An earlier study by Selbst and colleagues [12] also investigated the consequences of medication errors in a paediatric emergency department. They found that almost half of patients required additional monitoring (30%), examination (6%), or treatment (12%) after medication errors resulting from weight errors. To obtain reliable outcome measures from routinely collected EHR data, Sáez et al [10] developed a national, standardized, data quality–assessed, integrated data repository on maternal-child care. During this process, they found that variability in data quality across hospital sites could lead to imprecise comparison of measurements. Moreover, data quality indices, the efficiency of research processes, and the reliability of subsequent results have been found to improve if patient records are assessed for data quality [13,14]. Hence, quality assessment of source health data is crucial to identify and mitigate data quality problems for proper data use and reuse.

In this paper, we first describe our practical approach to quality assessment of health outcomes data. Adopting this methodology, we performed a pilot study on a subset of ICHOM outcomes data collected during routine clinical care of patients with congestive heart failure (CHF) in a general hospital, given the high prevalence and margin for outcomes improvement in heart failure [15]. Assessing data quality of outcomes data obtained during routine clinical care is of great interest since ICHOM indicators are currently collected through dedicated data collection into specialist outcome measurement systems, which results in useful data but is not a scalable process. The complexity of the analysis and in selecting the diagnosis for more than one condition, as well as the comorbidities associated with each disease, the different treatments received in each case, and all the variables used in the analysis, make it very difficult to conduct a system-wide quality assessment including several diseases and to interpret the results of a multiple disease analysis.

Methods

Data Quality Assessment

Research into data quality has gained attention since the seminal work by Wang and Strong [16], who proposed a comprehensive "fit-for-use" data quality assessment framework using data quality dimensions. Since then, several studies have aimed to define data quality dimensions and methodologies to describe and measure the complex multidimensional aspects of data quality [14,17-20]. Across studies, little agreement exists about the exact definition and meaning of data quality dimensions. Despite differences in terminology, though, many of the

XSL•FO
RenderX

proposed dimensions and solutions aim to address conceptually similar data quality features [14].

Following a review of existing literature, the data quality task force of the European Institute for Innovation through Health Data (i~HD) [21] identified 9 frameworks for quality assessment of health data [5,14,19,22-27]. From these frameworks, 9 data quality dimensions were selected during a series of workshops with clinical care, clinical research, and information and communication technology leads from 70 European hospitals: completeness, consistency, correctness, uniqueness, stability, timeliness, trustworthiness, contextualization, and representativeness. The selected data quality dimensions were deemed most important to assess the quality of health data if these data are to be useful for patient care, organizational learning (quality improvement, such as the assessment of health outcomes), and research (big data research and case finding for clinical trial recruitment). Multimedia Appendix 1 provides an overview of the selected data quality dimensions, together with their original terminology; the completeness, consistency, correctness, uniqueness, and stability dimensions were the most commonly used in the data quality literature, and for this reason, we selected them for the quality assessment in this study [14,20]. For instance, trustworthiness and timeliness are based on some types of metadata that are not usually available or accessible in EHR. Although sometimes the first 3 can overlap in their definitions or be contained within each other, we prefer making them orthogonal. For instance, a patient observation is incomplete if it is not registered, inconsistent if it does not comply with formatting requirements, or incorrect if it is unlikely to be true for a specific patient. For example, multiple normal kidney blood test results for a patient on dialysis would be consistent, though incorrect. Uniqueness, in turn, assesses whether duplications are present among patient records, for example as a result of an incomplete merging of patient records between hospital departments.

Further, stability relates to the probabilistic concordance of data among different data sources such as hospitals, physicians, or devices or over time [28]. For example, variability among centers has been found in liver offer acceptance rates for pediatric patients and cannot be explained by donor and recipient factors [29]. In some cases, standardization of procedures and analyses can reduce levels of variability. However, sometimes differences among centers persist even when using standard procedures, for instance, between diffusion tensor magnetic resonance imaging findings obtained at different acquisition centers using a standard protocol [30]. Likewise, when data are collected over time, temporal changes can occur due to several reasons, including changes in clinical practice or coding scheme used in the EHR [31].

Next, timeliness describes how promptly information is processed or how current recorded information is, for instance, to evaluate whether a current medication list within an EHR system is up to date or if there is a delay in updating this from a pharmacy subsystem. Trustworthiness relates to the availability of registry governance metadata and the data owner's reputation. For example, it must be possible for someone accessing a health data item or clinical document to confidently know when and where it was captured, by whom, and if it has been modified since the original entry. Further, contextualization relates to whether the data are annotated with their acquisition context, which can be crucial for correct interpretation of the results, for instance, whether blood glucose laboratory results were obtained while the patient was fasting. Finally, representativeness captures whether a dataset is representative for the population from which it is supposed to be drawn, in order to allow valid inference.

Pilot Assessment

Dataset

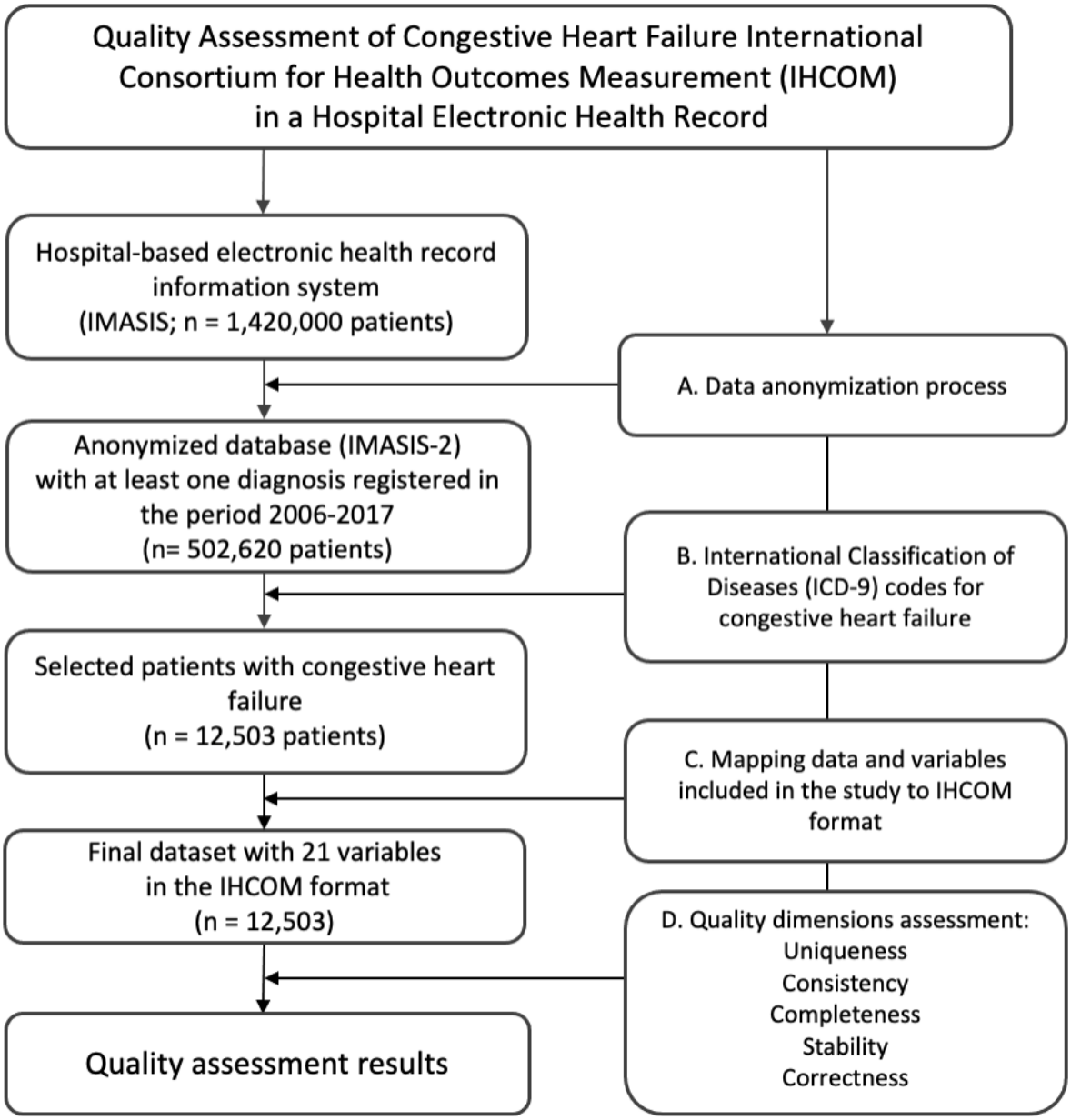
For this pilot assessment, we used data from the Parc Salut Mar Barcelona, a complete health care services organization with its information system database (IMASIS) as EHR. IMASIS includes and shares clinical information from 2 general hospitals, 1 mental health care center, 1 social health care center, and 5 emergency rooms in the Barcelona city area (Spain). IMASIS contains clinical information from approximately 1.5 million patients who have used the services of this health care system since 1989, across different settings such as admissions, outpatient consultations, emergency room visits, and major ambulatory surgery appointments. IMASIS-2 is the anonymized relational database of IMASIS that was created during the European Medical Information Framework (EMIF) project [32] and is the data source used for research purposes. It contains structured data related to diagnosis, procedures, drug administration, and laboratory tests and clinical annotations in a free-text format. Since natural language processing falls beyond the scope of this project, we only used structured data. The study protocol was approved by the Ethics Committee of Parc Salut Mar (num. 2016/6935/I), under the research activities related to ischemic heart disease carried out during the EMIF project funded by the Innovative Medicines Initiative.

As a case study, data from patients diagnosed with CHF were used. Heart failure is a chronic condition, severely impacting people's quality of life. With a prevalence of over 23 million worldwide, it poses a significant public health problem [33]. Collecting meaningful data on the health status of heart failure patients is therefore an important step to ensure better quality care and as a result, better quality of life for these patients.

All patients (n=502,620) who attended the hospital at least once between January 1, 2006 and November 7, 2017 and who had at least one diagnosis entry of CHF were extracted from the IMASIS-2 database. Specifically, the selection of patients was based on the following diagnosis codes of the International Classification of Diseases ninth edition (ICD-9): 428, 428.0, 428.1, 428.2, 428.20, 428.21, 428.22, 428.23, 428.3, 428.30, 428.31, 428.32, 428.33, 428.4, 428.40, 428.41, 428.42, 428.43, 428.9. In total, the dataset included 142,345 patient visit records describing the medical history of 12,503 different patients who had one or more of these diagnoses. Figure 1 provides a flow diagram of the different steps that were performed to obtain our study dataset. The main steps followed in the study were (1) a data anonymization process, (2) selection of the ICD-9 codes to select patients with CHF, (3) mapping data and variables included in the study to the IHCOM standard format, and (4) quality dimensions analysis.

XSL•FO
RenderX

Figure 1. Overview of the procedure to identify the patients to generate the study dataset.



The ICHOM heart failure outcomes standard set [13] was chosen as the most appropriate source of outcome indicators to target. Of the total of 72 ICHOM data items, a subset of 21 variables was selected as being most likely to be routinely collected within the hospital for patients with CHF and to be indicative of the overall quality of data collected for this type of patient. In addition, these variables allowed us to have complete information for the main characteristics of patients including age and sex as well as relevant comorbidities, such as hypertension or diabetes mellitus, and some of the most frequent treatments received for CHF, such as beta blockers, diuretics, and digoxin. The 21 variables were organized in 6 areas: identifiers, demographic factors, baseline health status, treatment variables, burden of care, and mortality. In addition, a visit identifier was included to distinguish different patient visit records. An overview of all variables included in the pilot assessment can be found in Table 1. In addition, Multimedia Appendix 2 shows the ICD-9 codes used to identify baseline health status variables, and Multimedia Appendix 3 shows the Anatomical Therapeutic Chemical classification system codes of the World Health Organization [34] to retrieve patients' medication usage.

XSL•FO
RenderX

Table 1. Overview of International Consortium for Health Outcomes Measurement (ICHOM) variables used in the pilot assessment.

Item	Definition	Response options
Identifiers		
Patient ID	Patient's medical record number	According to institution
Visit ID	Unique visit record identifier	Not included in the ICHOM standard set
Demographic factors		
Age	Date of birth	DD/MM/YYYY
Sex	Sex at birth	1=Male, 2=Female
Baseline health status		
Atrial fibrillation	Ever diagnosed with atrial fibrillation	0=No, 1=Yes, 999=Unknown
Prior myocardial infarction	Ever diagnosed with myocardial infarction	0=No, 1=Yes, 999=Unknown
Hypertension	History of hypertension	0=No, 1=Yes, 999=Unknown
Diabetes mellitus	Ever diagnosed with diabetes mellitus	0=No, 1=Yes, 999=Unknown
Echocardiogram performed	Echocardiogram performed to assess ejection fraction	0=No, 1=Yes, 999=Unknown
Height	Height (cm)	Numeric value of height in the metric system
Weight	Weight (kg)	Numeric value of weight in the metric system
Alcohol use	Consumption of >1 alcoholic drink a day	0=No, 1=Yes, 999=Unknown
Smoking status	Current smoking status	0=No, 1=Yes, 999=Unknown
Treatment variables		
Beta blocker	Beta blockers currently prescribed for heart failure	0=No, 1=Yes, 999=Unknown
Calcium channel blocker	Calcium channel blockers currently prescribed for heart failure	0=No, 1=Yes, 999=Unknown
Digoxin	Digoxin currently prescribed for heart failure	0=No, 1=Yes, 999=Unknown
Diuretics	Diuretics currently prescribed for heart failure	0=No, 1=Yes, 999=Unknown
Burden of care		
Date of arrival	Date of admittance	DD/MM/YYYY
Date of discharge	Date of discharge	DD/MM/YYYY
Hospital admissions	Number of hospitalizations in last 12 months due to heart failure	Numerical value or 999=Unknown
Hospital appointments	Number of hospital appointments in last 12 months due to heart failure	Numerical value or 999=Unknown
Mortality		
Date of death	Date patient was declared dead	DD/MM/YYYY or 999=Unknown

Anonymized data on patients, visits, diagnosis, procedures, drug administration events, laboratory tests and patient measures were collected from the relational database IMASIS-2 where all these fact tables are connected to the patient table via the patient identifier. In addition, visit, diagnosis, and procedures are connected to each other via the visit identifiers, whereas drugs, laboratory, and patient measures are connected to all domains via date matching. Specific queries requesting data from each of these tables yielded the "Temporary datasets" that were subjected to several transformation steps and to a successive left outer join merging process in which patient and visit identifiers were set as the initial left dataset. As a result, data were organized in a "visit-centered" fashion (every row contains all data related to a visit), thus providing the final dataset according to the ICHOM format.

Data Quality Dimensions

To evaluate the quality of heart failure patient data collected during routine clinical care, a subset of 5 data quality dimensions was selected: completeness, correctness, consistency, uniqueness, and stability. These dimensions are most commonly used in the data quality literature and were deemed most interesting to assess given the nature of the data.

First, for *uniqueness*, we measured the frequency with which partially duplicated patient records occur. Second, for *consistency*, we assessed data compliance with their expected data type (percentage of fields of a different type than defined),

XSL•FO
RenderX

value range (percentage of fields out of the expected range), and basic multivariate rules (percentage of data not fulfilling rules; for example, patient's arrival date should be before or equal to their date of discharge) [10]. Next, for *completeness*, we measured the proportion of complete fields per variable. Further, for *stability*, we qualitatively evaluated the temporal stability of recorded past medical conditions and usage of different types of medications. To this end, we computed, per month, how many patient visit records mentioned a history of a particular medical condition or usage of a specific medication out of the total number of patient visit records that month. We then visualized trends for each of these data items by plotting the respective relative frequencies over time. Finally, we inferred data correctness from the data, either by combining information across variables or by investigating data from the same patient over time. Specifically, plausibility of height and weight was examined by computing patients' BMIs. Further, we investigated the temporal order of past medical conditions, assuming that once a hospital visit record indicates that a patient has a history of atrial fibrillation, hypertension, diabetes, or myocardial infarction, the history of this diagnosis or event should be mentioned in all subsequent visit records. Based on this assumption, for assessment purposes, some deviations from this temporal order (ie, "history" followed by "no history") point to data errors in the extracted dataset.

Tools

We conducted the data quality assessment using R, version 3.6.1 [35]. For the temporal stability analyses, we used the EHRtemporalVariability R package [36].

Results

Uniqueness

Of a total of 142,345 patient visit records, 1.2% had identical visit identifiers even though values for one or more data items had different inputs (Uniqueness result 1=98.8%). In turn, 2.8% of all patient visit records had at least another record with a different visit identifier registered the same day and identical clinical data (Uniqueness result 2=97.2%). In IMASIS-2, visits and clinical data are connected via date matching. Therefore, for 1 patient attending 2 visits in the same day, both visits are connected to the same data. This amounts to an average score of 98% for uniqueness.

Consistency

Consistency by type and by multivariate rules both yielded a score of 100%; all values were in the right format, and no errors in relationships between dates were found. As a third consistency check, we examined whether numerical and date values fell within prespecified ranges and whether categorical variables had values that complied with predefined response options. An average score of 91.21% was obtained for consistency by range, resulting from errors in 3 variables. In particular, 85% of values for height and weight were "0." Since weight and height values of zero do not have a physical meaning, we hypothesized that these data points were missing data values. Indeed, zero entries are not even permitted in the structured data fields of height and weight. Rather, these zero values were introduced during data extraction from the IMASIS-2 database to indicate missingness, since only numeric values are accepted for height and weight according to the ICHOM Heart Failure data dictionary (summarized in Table 1). In addition, a small number of out-of-range data points were identified for height (n=54) and weight (n=20). Further, 16 visit records had arrival dates before January 1, 2006. Across the 3 domains of consistency, this yields an average score of 97.07%.

Completeness

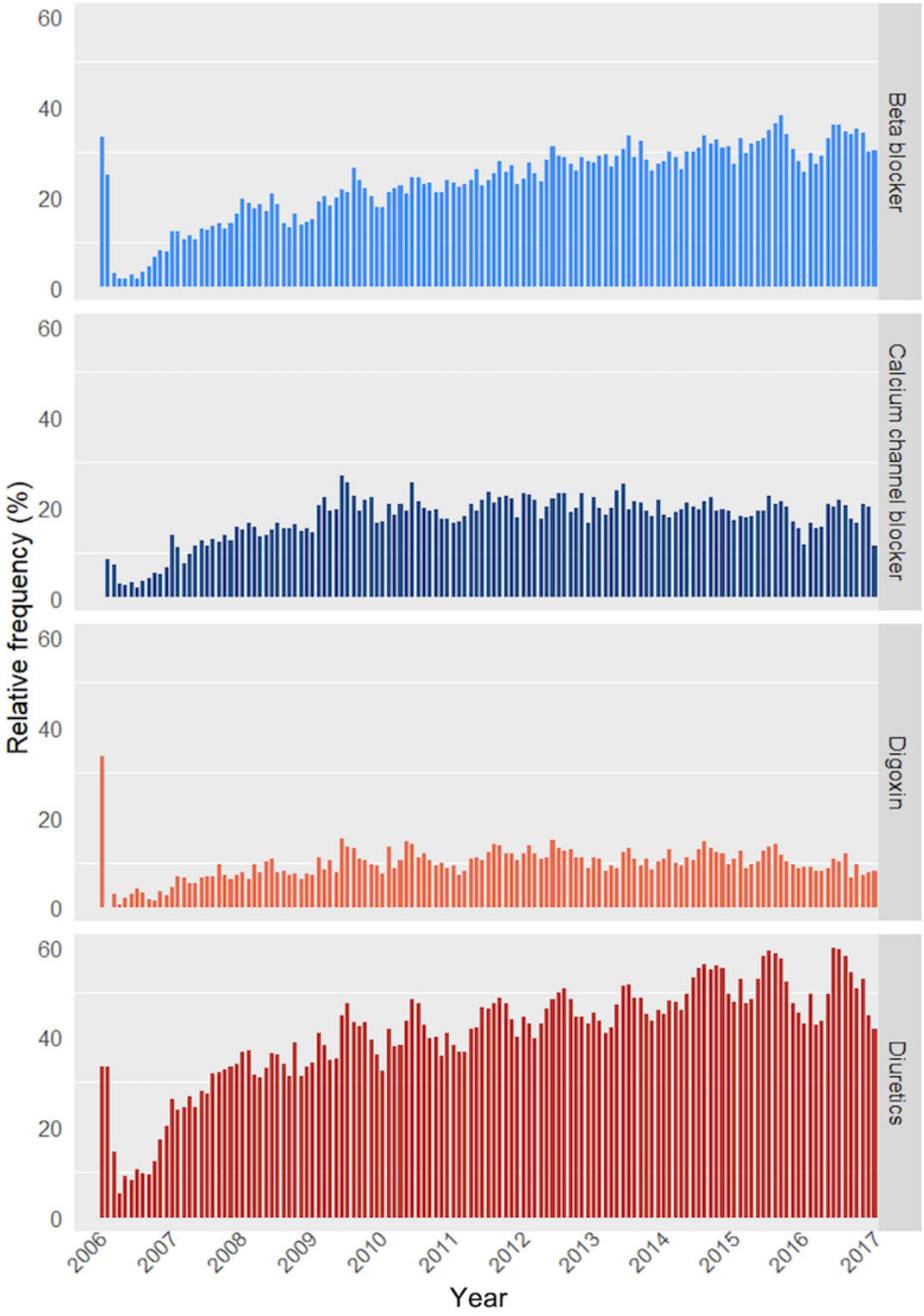
Assessing completeness of the dataset by column revealed that all included variables were completely documented, except for date of death, which was only recorded in 37.14% of all patient visits. This incompleteness is valid, though, since date of death is only provided when the patient died during the visit. Excluding this valid incompleteness result, an average score of 100% was obtained for completeness.

Stability

Two categories of data items were assessed for temporal variability: medication usage and past medical conditions. As illustrated in Figure 2, the results showed a gradual increase over time in the recorded usage of different types of medication to treat heart failure, especially of beta blockers and diuretics. Further, we found an abrupt change in the documentation pattern of past medical conditions in 2011, with drastically reduced frequencies of reported past medical conditions (Figure 3). Of note, only a small number of patient visit records (<10) was available for each month in the first half of 2016, explaining the absent or divergent results.

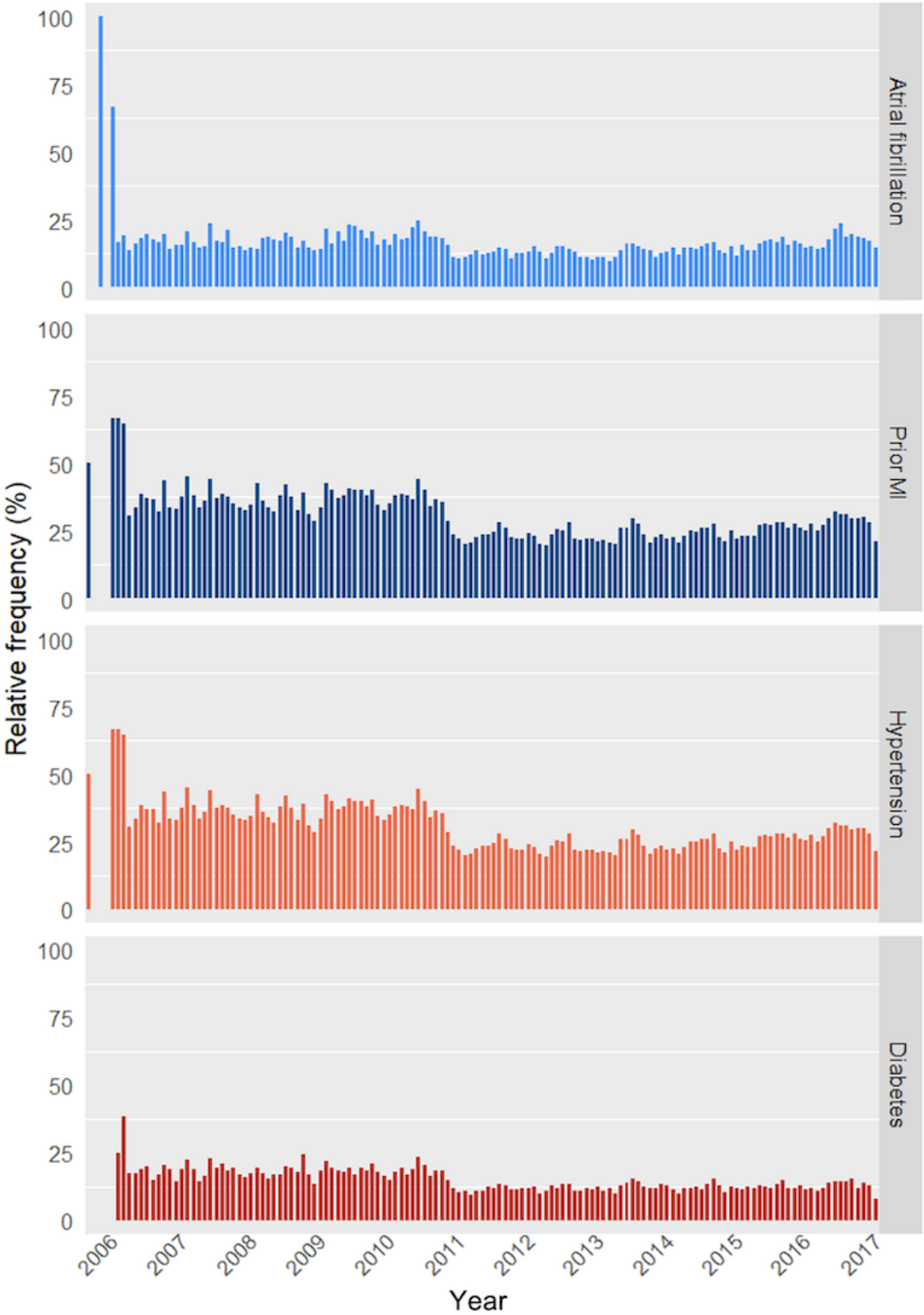
XSL•FO
RenderX

Figure 2. Percentage of patients with a record of specific drug usage per month, relative to the total number of patient admissions within that month, plotted over time.



XSL•FO
RenderX

Figure 3. Percentage of patients with a record of a specific past medical condition per month, relative to the total number of patient admissions within that month, plotted over time. MI=myocardial infarction.



Correctness

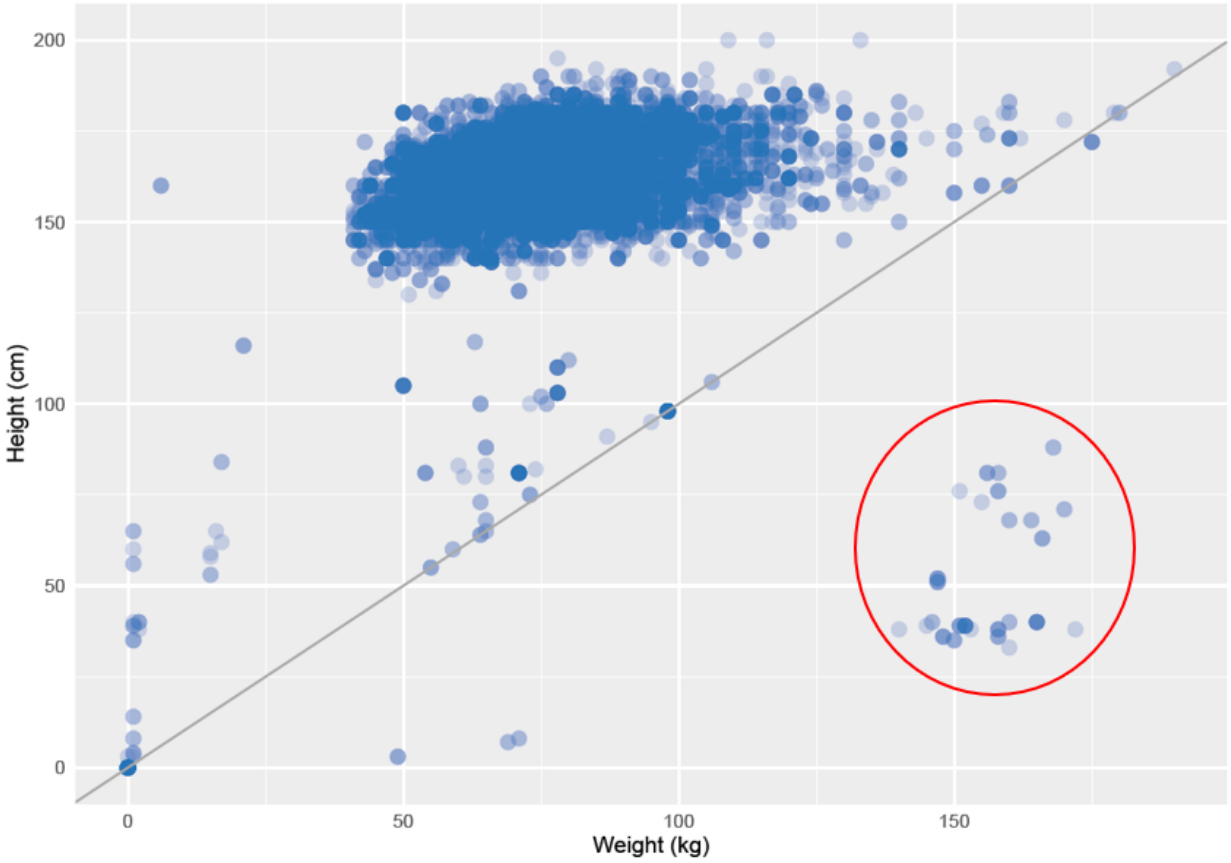
After performing basic descriptive analyses, results of which are summarized in Multimedia Appendix 4, 2 sets of variables were subjected to closer inspection. First, correctness of height and weight values was evaluated based on their bivariate distribution, as shown in Figure 4. All data points that fall below the main diagonal, implying that the patient's weight (in kg) is larger than his or her height (in cm), are very unlikely to be true. A subset of these data errors, highlighted by the red circle, were hypothesized to result from value inversion between height and weight recordings. To formally assess implausible height and weight values, we computed the patients' BMIs. Results showed that 16 patients had a suspiciously low BMI (<10 kg/m^2), and 180 patients had an implausibly high BMI (>70 kg/m^2). Hence,

XSL•FO
RenderX

a total of 196 probable errors were identified, corresponding to 0.13% of all patient visit records.

Further, we investigated the temporal order of past medical conditions. Results showed a substantial number of deviations. Specifically, 6.33% of all patient visit records mentioned that the patient did not have a history of atrial fibrillation, while earlier records indicated the patient had previously been diagnosed with atrial fibrillation. Similarly, for history of hypertension, diabetes mellitus, and myocardial infarction, error rates of 12.11%, 6.12%, and 12.11%, respectively, were obtained. These deviations in temporal order were introduced while mapping the IMASIS-2 relational database contents to the ICHOM format, as the latter requires a level of detail that is not explicitly available in the coded data of an EHR. In particular, diagnoses or events already recorded in a previous visit and not mentioned in a subsequent visit are not consistently recorded in EHR systems during routine clinical care, in contrast to data collected for research purposes. It is therefore practically impossible to distinguish true negatives from missing data when extracting data from the EHR. As a result, a substantial proportion of patient history data items that were negative in the dataset actually represent missing data values. Taken together, this amounts to a total score of 93.84% for correctness.

Figure 4. Bivariate distribution of height and weight values, with the red circle highlighting the data points where height and weight values were hypothesized to have been inverted.



Discussion

Data Quality Assessment Results and Suggestions for Improvement

Overall, this pilot assessment revealed high scores on each of the dimensions used to investigate the quality of heart failure patients' data. Nevertheless, several data quality issues were identified, based on which we propose a set of improvement strategies.

Regarding consistency, results of our data quality assessment showed that a substantial number of negative values in the dataset — indicating the absence of a particular data item — actually represented missing data. Consequently, some variable distributions seem to be biased. For example, according to the data, only a minority of patients currently smoked or had a past medical condition such as hypertension (see Multimedia Appendix 4), which is rather implausible for a population of patients with heart failure. This is an intrinsic issue associated with structured data sources in the framework of EHR databases. That is, when a code is not found in the EHR, it is practically impossible to distinguish whether the code is negative (ie, examination has confirmed the absence of a particular condition) or missing (ie, no examination has taken place, or examination confirmed the presence of a particular condition but is not recorded in a structured format) for a given patient. We are aware that good clinical practice does not mandate the measurement of every data item at each patient visit (eg, disease history), since these items usually are present as additional information in a typical EHR environment. Nevertheless, this

XSL•FO
RenderX

differs fundamentally from data collection practices in the context of research activities such as outcomes assessment, for which the ICHOM standard set was originally developed. When performing analytical and research activities, it would therefore be very useful to introduce mechanisms or tools that allow differentiation of data missingness from true negatives and to determine the duration of each condition and disease, regardless of whether they are mentioned in each visit.

Further, the uniqueness analyses revealed some partially duplicated patient visit records. First, duplications in visit identifiers were found, while clinical data showed different inputs. Data management staff at the Hospital del Mar clarified that this happened whenever different height and weight measurements were registered during a single visit. If a slight difference between values is observed, partial row duplicates are generated when merging data in the final dataset. Second, duplicated rows with different visit identifiers have arisen because of the data organization in IMASIS-2, where some clinical data are connected to visit IDs via date matching. As a result, all clinical data collected during different patient visits on the same day are connected to different visit IDs depending on the department or hospital service where these patients visit even on the same day. To reduce future data quality issues of this kind, we suggest a data reorganization including a 2-level visit structure. First, a more general level would describe a period in which one or different visits occur and is connected to clinical data obtained within this period. Second, a more specific level would then describe every distinct visit together with a corresponding diagnosis and procedure information obtained during the particular visit. This 2-level visit organization would contribute to the elimination of partial replicates, thus positively impacting the uniqueness aspect of data quality. This strategy has been previously adopted by the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) standard [37] with the aim of easing mappings from ambiguous visit-connected schemas.

When analyzing and interpreting completeness, it is essential to take into account the type of information that is registered based on the characteristics of the database, for instance, in this case a hospital-based EHR in which information and variables related to death and data for death are only registered when this situation occurs during admission. For instance, the link among different registries and databases such as primary care, hospital, and mortality registries is essential to contribute to the completeness of this type of information.

Temporal stability analyses revealed an abrupt change in the documentation pattern of past medical conditions in 2011, with drastically reduced frequencies of reported past medical conditions. For instance, the introduction of a new automated coding system in the emergency department EHR system accompanied an increase in the number of registries and codifications in this department and therefore in the system. Although we assume this evolution in the recording of past medical conditions had a positive impact on direct patient care, decision support and alert algorithms can be impacted by changes in diagnostic coding practice and should therefore be considered. In addition, these changes will affect the reuse of data for research and quality monitoring such as outcomes tracking. In this sense, quality assessment is an essential tool to detect the effects of changes in EHR systems introduced over time, which would contribute to a better understanding of the updates in the content and structure of these types of databases. Finally, regarding the important point related to the potential impact of changes or upgrades in EHR system and diagnostic coding practices due to common changes in the way diseases are coded or for instance the necessity to included new diseases, we recommend preparing carefully for this type of situation.

In relation to correctness, many data items are often recorded in free text rather than structured data fields, making it difficult to extract this information for research and analysis purposes. We therefore advise to maximally include data items in form format or specific fields or sections in the EHR. In addition, when using form formats, we recommend the use of alarms for avoiding missing values as well as for inputting out-of-range data. Alternatively, natural language processing techniques applied to free-text clinical annotation fields can be used to enrich structured sources.

Lessons Learned

The process of assessing the quality of outcomes data obtained during routine clinical care is of great value and allows us the opportunity to learn several relevant aspects in the management and evaluation of clinical information in EHR environments. The most relevant lessons learned were (1) the evaluation requires having considerable knowledge of the EHR (data available, how the data were collected, or who collected it) to fully understand its structure and different staff needs; (2) it is critical that the metrics are feasible, valid, and meaningful for a specific EHR system and its quality evaluation and should be understood and used accordingly; (3) once the quality of the data is assessed, it is important to monitor it regularly, and the value of an external data quality assessment by an independent organization should be considered. In addition, high-quality data enhance the validity and reliability of study findings and thinking of using EHR systems for purposes other than health care such as research. Finally, it is interesting to consider that EHR models would need to be expanded and redesigned in content and structure, and a data quality assessment can assist in doing these tasks.

Limitations and Future Directions

In interpreting the results of this study, some important limitations should be taken into consideration. First, although the selection of a subset of ICHOM outcome variables for the data quality assessment was made in agreement among all the members of the study assessment based on the most likely routinely collected data within their EHR for patients with CHF, it is possible that the use of more variables or other variables could affect the results of the quality assessment. For this reason, whether the data quality results from this pilot assessment are generalizable to the complete ICHOM standard set has yet to be investigated. Similarly, we selected 5 of 9 available data quality dimensions, as these were thought to be most relevant given the nature of the data. It is possible that the use of all 9 dimensions would show a more complete analysis of this type of data and therefore would offer additional recommendations for improvement. Further, data quality assessment was

XSL•FO
RenderX

performed on a data extract from the IMASIS-2 dataset after mapping the data items to the ICHOM outcomes format, which might have introduced additional errors. We therefore recommend future studies to examine the data quality of the EHR variables directly, in the hospital's own response format, or to perform an additional data quality assessment of the mapping procedure.

In sum, future research would benefit from performing more thorough data quality assessments, across multiple hospitals, to truly examine to what extent hospitals today are able to routinely collect the evidence of their success in achieving good health outcomes. The European Federation of Pharmaceutical Industries and Associations (EFPIA) is currently leading such a project together with i~HD. In particular, the goal of this project is to assess the availability and quality of routinely collected patient data to underpin a future scale-up of value-based care models in which ICHOM outcomes indicators serve as the measures of value delivered by health care provider organizations. For this project, data from patients with heart failure are also being examined, now using the complete set of ICHOM outcomes indicators and performing assessments across 10 European hospitals. The promotion of data quality is essential to advance learning health systems, patient empowerment, and clinical research, and the results of this larger project will provide interesting insights on the generalizability of this pilot project's findings.

Acknowledgments

MAM and JMRA had support from the Innovative Medicines Initiative Joint Under-taking under EMIF grant agreement no. 115372, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies. The funders were not involved in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Authors' Contributions

MAM and JMRA selected the variables to be included in the analysis and provided the data for analysis. HA performed data quality analyses, interpreted the results, and wrote the manuscript. CS and MDH performed data quality analyses. Baseline data quality assessment scripts in R were provided by CS, MDH, and JMGG. All authors interpreted the data quality analyses results, contributed to the writing of the manuscript, performed critical revisions of the manuscript, and approved the final version for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mapping of data quality dimensions.
[DOCX File , 27 KB - medinform_v9i8e27842_app1.docx]

Multimedia Appendix 2
ICD-9 classification codes used for the evaluation of baseline health status variables.
[DOCX File , 13 KB - medinform_v9i8e27842_app2.docx]

Multimedia Appendix 3
Anatomical Therapeutic Chemical classification system (ATC/DDD) codes of the World Health Organization used to retrieve patients' medication usage.
[DOCX File , 13 KB - medinform_v9i8e27842_app3.docx]

Multimedia Appendix 4
Results of descriptive analyses.
[DOCX File , 14 KB - medinform_v9i8e27842_app4.docx]

References

1. Donabedian A. Evaluating the Quality of Medical Care. The Milbank Memorial Fund Quarterly 1966 Jul;44(3):166. [doi: 10.2307/3348969]
2. O'Connor DP, Brinker M. Challenges in outcome measurement: clinical research perspective. Clin Orthop Relat Res 2013 Nov;471(11):3496-3503 [FREE Full text] [doi: 10.1007/s11999-013-3194-1] [Medline: 23884806]
3. Porter ME, Teisberg EO. Redefining Health Care: Creating Value-Based Competition on Results. Cambridge, MA: Harvard Business Review Press; 2006.

XSL•FO
RenderX

4. Kelley TA. International Consortium for Health Outcomes Measurement (ICHOM). Trials 2015 Nov 24;16(S3):1. [doi: 10.1186/1745-6215-16-s3-o4]
5. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. Summit Transl Bioinform 2010 Mar 01;2010:1-5 [FREE Full text] [Medline: 21347133]
6. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev 2010 Oct 11;67(5):503-527. [doi: 10.1177/1077558709359007] [Medline: 20150441]
7. Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. The AP-HP experience. Comput Methods Programs Biomed 2019 Nov;181:104804. [doi: 10.1016/j.cmpb.2018.10.016] [Medline: 30497872]
8. Doods J, Botteri F, Dugas M, Fritz F, EHR4CR WP7. A European inventory of common electronic health record data elements for clinical trial feasibility. Trials 2014 Jan 10;15:18 [FREE Full text] [doi: 10.1186/1745-6215-15-18] [Medline: 24410735]
9. Weir CR, Hurdle JF, Felgar MA, Hoffman JM, Roth B, Nebeker JR. Direct text entry in electronic progress notes. An evaluation of input errors. Methods Inf Med 2003;42(1):61-67. [Medline: 12695797]
10. Sáez C, Moner D, García-De-León-Chocano R, Muñoz-Soler V, García-De-León-González R, Maldonado JA, et al. A Standardized and Data Quality Assessed Maternal-Child Care Integrated Data Repository for Research and Monitoring of Best Practices: A Pilot Project in Spain. Stud Health Technol Inform 2017;235:539-543. [Medline: 28423851]
11. Hirata K, Kang A, Ramirez GV, Kimata C, Yamamoto LG. Pediatric Weight Errors and Resultant Medication Dosing Errors in the Emergency Department. Pediatr Emerg Care 2019 Sep;35(9):637-642. [doi: 10.1097/PEC.0000000000001277] [Medline: 28976456]
12. Selbst SM, Fein JA, Osterhoudt K, Ho W. Medication errors in a pediatric emergency department. Pediatr Emerg Care 1999 Feb;15(1):1-4. [doi: 10.1097/00006565-199902000-00001] [Medline: 10069301]
13. Burns DJ, Arora J, Okunade O, Beltrame JF, Bernardez-Pereira S, Crespo-Leiro MG, et al. International Consortium for Health Outcomes Measurement (ICHOM): Standardized Patient-Centered Outcomes Measurement Set for Heart Failure Patients. JACC Heart Fail 2020 Mar;8(3):212-222 [FREE Full text] [doi: 10.1016/j.jchf.2019.09.007] [Medline: 31838032]
14. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 01;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
15. Savarese G, Lund LH. Global Public Health Burden of Heart Failure. Card Fail Rev 2017 Apr;3(1):7-11 [FREE Full text] [doi: 10.15420/cfr.2016:25:2] [Medline: 28785469]
16. Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 2015 Dec 11;12(4):5-33. [doi: 10.1080/07421222.1996.11518099]
17. Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM Comput. Surv 2009 Jul;41(3):1-52. [doi: 10.1145/1541880.1541883]
18. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. A Data Quality Ontology for the Secondary Use of EHR Data. AMIA Annu Symp Proc 2015;2015:1937-1946 [FREE Full text] [Medline: 26958293]
19. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care 2012 Jul;50 Suppl:S21-S29 [FREE Full text] [doi: 10.1097/MLR.0b013e318257dd67] [Medline: 22692254]
20. Liaw S, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, et al. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. Int J Med Inform 2013 Jan;82(1):10-24. [doi: 10.1016/j.ijmedinf.2012.10.001] [Medline: 23122633]
21. Kalra D, Stroetmann V, Sundgren M, Dupont D, Schlünder I, Thienpont G, et al. The European Institute for Innovation through Health Data. Learn Health Syst 2017 Jan 25;1(1):e10008 [FREE Full text] [doi: 10.1002/lrh2.10008] [Medline: 31245550]
22. Zozus M, Hammond W, Green B, Kahn M, Richesson R, Rusincovitch S, et al. Assessing Data Quality for Healthcare Systems Data Used in Clinical Research (Version 10). 2014. URL: https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Assessing-data-quality_V1%200.pdf [accessed 2021-07-10]
23. Davoudi S, Dooling J, Glondys B, Jones T, Kadlec L, Overgaard S, et al. Data Quality Management Model (2015 Update) - Retired. The American Health Information Management Association. 2015. URL: http://library.ahima.org/PB/DataQualityModel#.XW6r-pNKjab [accessed 2021-07-10]
24. Sáez C, Martínez-Miranda J, Robles M, García-Gómez JM. Organizing data quality assessment of shifting biomedical data. Stud Health Technol Inform 2012;180:721-725. [Medline: 22874286]
25. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC) 2016 Sep 11;4(1):1244 [FREE Full text] [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
26. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. Eur J Cancer 2009 Mar;45(5):747-755. [doi: 10.1016/j.ejca.2008.11.032] [Medline: 19117750]

XSL•FO
RenderX

27. Sariyar M, Borg A, Heidinger O, Pommerening K. A practical framework for data management processes and their evaluation in population-based medical registries. Inform Health Soc Care 2013 Mar;38(2):104-119. [doi: 10.3109/17538157.2012.735731] [Medline: 23323639]
28. Sáez C, Zurriaga O, Pérez-Panadés J, Melchor I, Robles M, García-Gómez JM. Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories. J Am Med Inform Assoc 2016 Nov;23(6):1085-1095. [doi: 10.1093/jamia/ocw010] [Medline: 27107447]
29. Mitchell E, Loomes KM, Squires RH, Goldberg D. Variability in acceptance of organ offers by pediatric transplant centers and its impact on wait-list mortality. Liver Transpl 2018 Jun 06;24(6):803-809. [doi: 10.1002/lt.25048] [Medline: 29506323]
30. Pagani E, Hirsch JG, Pouwels PJ, Horsfield MA, Perego E, Gass A, et al. Intercenter differences in diffusion tensor MRI acquisition. J Magn Reson Imaging 2010 Jun;31(6):1458-1468. [doi: 10.1002/jmri.22186] [Medline: 20512899]
31. Sáez C, García-Gómez JM. Kinematics of Big Biomedical Data to characterize temporal variability and seasonality of data repositories: Functional Data Analysis of data temporal evolution over non-parametric statistical manifolds. Int J Med Inform 2018 Nov;119:109-124. [doi: 10.1016/j.ijmedinf.2018.09.015] [Medline: 30342679]
32. Lovestone S, EMIF Consortium. The European medical information framework: A novel ecosystem for sharing healthcare data across Europe. Learn Health Syst 2020 Apr;4(2):e10214 [FREE Full text] [doi: 10.1002/lrh2.10214] [Medline: 32313838]
33. Roger VL. Epidemiology of Heart Failure. Circ Res 2013 Aug 30;113(6):646-659. [doi: 10.1161/circresaha.113.300268]
34. Guidelines for ATC classification and DDD assignment, 2011. WHO Collaborating Centre for Drug Statistics Methodology. 2010. URL: https://www.whocc.no/filearchive/publications/2011guidelines.pdf [accessed 2021-07-10]
35. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2017. URL: https://www.r-project.org [accessed 2021-07-10]
36. Sáez C, Gutiérrez-Sacristán A, Kohane I, García-Gómez JM, Avillach P. EHRtemporalVariability: delineating temporal data-set shifts in electronic health records. Gigascience 2020 Aug 01;9(8):1 [FREE Full text] [doi: 10.1093/gigascience/giaa079] [Medline: 32729900]
37. The Book of OHDSI. Observational Health Data Sciences and Informatics. 2021 Jan 11. URL: http://book.ohdsi.org [accessed 2021-07-10]

Abbreviations

AI: artificial intelligence
CDM: Common Data Model
CHF: congestive heart failure
EFPIA: European Federation of Pharmaceutical Industries and Associations
EHR: electronic health record
EMIF: European Medical Information Framework
ICD-9: International Classification of Diseases ninth edition
ICHOM: International Consortium for Health Outcomes Measurement
i~HD: European Institute for Innovation through Health Data
OMOP: Observational Medical Outcomes Partnership

Edited by C Lovis; submitted 09.02.21; peer-reviewed by G Myreteg, S Beerten; comments to author 21.03.21; revised version received 30.05.21; accepted 05.06.21; published 04.08.21.

Please cite as:
Aerts H, Kalra D, Sáez C, Ramírez-Anguita JM, Mayer MA, Garcia-Gomez JM, Durà-Hernández M, Thienpont G, Coorevits P Quality of Hospital Electronic Health Record (EHR) Data Based on the International Consortium for Health Outcomes Measurement (ICHOM) in Heart Failure: Pilot Data Quality Assessment Study
JMIR Med Inform 2021;9(8):e27842
URL: https://medinform.jmir.org/2021/8/e27842
doi:10.2196/27842
PMID:34346902

©Hannelore Aerts, Dipak Kalra, Carlos Sáez, Juan Manuel Ramírez-Anguita, Miguel-Angel Mayer, Juan M Garcia-Gomez, Marta Durà-Hernández, Geert Thienpont, Pascal Coorevits. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 04.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete

XSL•FO
RenderX

bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Communicating the Implementation of Open Notes to Health Care Professionals: Mixed Methods Study

Karin Jonnergård[1], PhD; Lena Petersson[2], PhD; Gudbjörg Erlingsdóttir[2], PhD

[1]Department of Business Administration, Lund University, Lund, Sweden

[2]Department of Design Sciences, Lund University, Lund, Sweden

Corresponding Author:
Lena Petersson, PhD
Department of Design Sciences
Lund University
Box 118
Lund, SE-221 00
Sweden
Phone: 46 46 222 05 33
Email: Lena.Petersson@design.lth.se

Abstract

Background: The literature on how to communicate reform in organizations has mainly focused on levels of hierarchy and has largely ignored the variety of professions that may be found within an organization. In this study, we focus on the relationship between media type and professional responses.

Objective: The objective of this study was to investigate whether and how belonging to a profession influences the choice of communication media and the perception of information when a technical innovation is implemented in a health care setting.

Methods: This study followed a mixed methods design based on observations and participant studies, as well as a survey of professionals in psychiatric health care in Sweden. The $\chi 2$ test was used to detect differences in perceptions between professional groups.

Results: The use of available communication media differed among professions. These differences seem to be related to the status attached to each profession. The sense-making of the information appears to be similar among the professions, but is based on their traditional professional norms rather than on reflection on the reform at hand.

Conclusions: When communicating about the implementation of a new technology, the choice of media and the message need to be attuned to the employees in both hierarchical and professional terms. This also applies to situations where professional employees are only indirectly affected by the implementation. A differentiated communication strategy is preferred over a downward cascade of information.

(JMIR Med Inform 2021;9(8):e22391) doi:10.2196/22391

KEYWORDS

implementation; health care; electronic health records; communication strategy; eHealth; telemedicine; PAEHRs; Open Notes; professions; EHR

Introduction

Background

Since 2017, all residents of Sweden have been able to access their nonpsychiatric electronic health records (EHRs) through the internet, and thus have been able to read clinical notes. This "Open Notes" policy was first introduced in November 2012 by Region Uppsala, followed by Region Skåne in March 2014. In both regions, psychiatry was exempt because patient digital access was considered to be too sensitive. However, in 2015, Region Skåne included adult psychiatry in the service. This development is in line with the reasoning of the OpenNotes Project in the United States, which states that patients in psychiatric care should not be treated differently than other groups of patients in terms of their online access to EHRs [1-3].

Implementing new technical systems such as Open Notes in the public sector is often depicted as a complex process [4,5]. This is partly because activities in the public sector are affected by political and operational considerations, and are performed by both managers and street-level bureaucrats. Consequently, both

XSL•FO
RenderX

an administrative and a professional hierarchy [6,7] organize activities. This complexity needs to be considered [8-11] for successful implementation of new technology, particularly with regard to how the implementation is communicated to employees [12,13]. When a reform directly affects professionals, this is obvious; however, when Open Notes was implemented in adult psychiatry in Region Skåne [14,15], the aim was to empower patients and the new policy was not expected to influence the work of health care professionals [16]. Because the professionals are required to input notes into the system and thus cannot opt out from the service, the implementation of Open Notes was regarded as a service related to their work tool, the EHR, which would enhance the transparency of their work [17], but was not supposed to influence the way they used the work tool. Health care professionals would only indirectly be affected by the implementation, which makes it interesting to reflect on how the implementation was communicated to them.

General reviews by King et al [18] and Cresswell and Sheikh [19] discuss technical, social, and organizational obstacles in the implementation of electronic health (eHealth) technology. In addition to issues related to the technology per se, they identify the great importance of acceptance of the technology by different professional groups active in health care [8,20-23]. Careful choices need to be made about how the implementation of different devices is communicated to professional groups [24]. In principle, this also applies to the communication of information about reforms that will affect the professionals only indirectly.

Media selection theory [25] has identified factors that are important for successful communication when reforms are implemented. These include features of the sender and the organization, the characteristics of the communication media and the messages, the receiver, and consideration of the receiver's expected reaction. From a "perception management perspective," the aim is to manage the receivers' perceptions or sense-making of the information [24,26]. All information is subject to interpretation and its reception depends on the senders' and receivers' framing of the information [24,27], as well as the importance assigned to the matter [26,28]. However, it is also assumed that the receivers not only receive the information but also absorb it.

The perspectives of media selection theory and perception management are complementary. Media selection theories focus on the ability of an information channel to contribute rich information. It is commonly argued that channels that provide richer information should be available to managers at higher levels of the organization, especially in regard to equivocal tasks [29]. In addition, the need for coordination through communication is seen to increase along with the ambiguity of the implemented reform [30]. Research from a perception management perspective focuses on investigating the mode of implementation (hard/mixed/soft) [13] and the different stages in the diffusion of perceptions during implementation [26]. The perception management literature differs from media selection theory with respect to its emphasis that successful management of perception is context-dependent and may change over time. However, both perspectives assume the existence of an organizational hierarchy and task-oriented groups.

Professional groups are typically task-oriented. There is often a status hierarchy between different professions based on their different claims of jurisdiction [31] and on whether they are perceived as "full" professions or semiprofessions [32]. This status hierarchy is not always reflected in the formal organizational hierarchy. Thus, the theoretical assumptions described above may not always apply when implementing reforms in other contexts such as for health care organizations where the employees are professionals. Most of the research in both media selection theory and perception management perspective has dealt with formal hierarchical relationships within organizations without regard to the status of the professionals involved. Two related issues thus arise in relation to implementing an eHealth reform: (1) What media are best used to inform and change the perceptions of professional groups during the implementation phase? and (2) What aspects do the professionals perceive as important for the implementation? Far too little attention has been paid to these two issues. They may be mentioned, but have scarcely been investigated from a perception management perspective. Moreover, most research has focused on the implementation of reforms that are assumed to directly affect professionals. To our knowledge, no research has focused on a situation where professionals are indirectly affected by reforms.

The overall aim of this study was to investigate these two issues in relation to the implementation of Open Notes in adult psychiatry. Three research questions were formulated based on the results of a survey of professionals' perceptions of the implementation process:

> RQ1: Which strategies were used in the information and communication activities directed to professionals before the implementation of Open Notes?
>
> RQ2: Do different professional groups demonstrate different patterns regarding the communication media through which they absorbed information?
>
> RQ3: Which aspects did different professional groups perceive and prioritize as important in the communication?

This study builds on data from a previous mixed method study [33], focusing on (1) the strategies underlying the information and communication activities connected to the implementation of Open Notes in adult psychiatry in Region Skåne, (2) the channels through which the professionals received and absorbed information about the implementation, and (3) how essential they considered the information to be.

Communication in the Processes of Implementation

A common approach to specify how information about technological innovations is diffused is to differentiate between dissemination and implementation. Dissemination refers to "active and planned efforts to persuade target groups to adopt an innovation" [34], whereas implementation refers to presenting information about the technical device's functions and the way it will be integrated in the organization. According to Fidler and Johnson [35], implementation implies hierarchical power to implement the innovation, whereas dissemination does not. Considering Open Notes, the service is optional for the patients.

XSL•FO
RenderX

The diffusion of the information to the patients can be characterized as dissemination. By contrast, the introduction of Open Notes to health care personnel can be viewed as a type of "indirect implementation," as the reform did not aim to affect the work of the professionals, although it would increase the transparency between the patients and personnel. Open Notes was implemented in a typically top-down manner [36]: the decision was taken at a policy level and the implementation was seen as more of an administrative than a political process.

This type of implementation requires clear goals and change agents sympathetic to the goals. It presupposes that an organization has clear hierarchical relations and unifying organizational cultures. However, these presumptions are seldom valid in health care organizations, which are characterized as arenas for politicians, administrators, and professionals [6]. In such contexts, different groups may have different goals, and the hierarchical order may be blurred.

Markus and Pfeffer [37] mention three conditions that may obstruct implementation in organizations: (1) if the power distribution implicit in the reform does not match the existing power distribution in the organization, (2) if the language and symbols of the reform do not correspond to the dominant organizational paradigm and culture, and (3) if the goals and technology do not align with widely held goals and technology. Here, it should be noted that the hierarchy in professional groups is often assumed to be based on knowledge, with professionals being governed by professional norms and culture, and salient technology is regarded as directly connected to the work of the profession [38]. Both organizational and professional features could thus obstruct an indirect implementation such as Open Notes [9].

Rogers [10] argues that the complexity of a technological innovation is an important factor in the acceptance of reforms. If any part of the technology does not agree with the values of the adopters, or if the benefits are low or difficult to observe, the reform is likely to meet with resistance. In a case study on the implementation of diagnosis-related groups in Finnish health care, Lehtonen [11] drew conclusions in line with Rogers' [10] assumptions. However, Lehtonen [11] also noted that early communication with clinical personnel and their involvement eased the implementation, as did freedom of choice regarding the degree to which the reform would be applied.

Communication between change management representatives and employees is crucial in any planned change process [12]. Mikkelsen et al [13] argue that the style of communication influences the motivation of the employees. Hard regulation from upper levels of the hierarchy may crowd out intrinsic motivation, while softer regulation may encourage employees' own motivation and give them more positive perceptions of change. In addition, the views of what information should be shared and how it should be communicated may differ between management and employees. The greater the distance between the two groups, the less direct and rich the information the subordinate group receives will be. Employees have to rely on different levels of management to convey information to them. However, employee engagement and cooperation are key to success in the implementation process [39,40]. This is particularly true for organizations where employees are professionals and the implementation involves new technology [23,41].

A top-down implementation strategy implies programmatic change communication [42] (ie, a "telling and selling approach"). Russ [42] compares this approach to a "downward cascade of information about the change." The advantage of programmatic change communication is the ability to disseminate quality information from the top of the organization to everyone, which gives the impression of equal and fair information. Programmatic change communication can lower the uncertainty surrounding a reform as well as the resistance to the innovation; it is also known to have the appeal of "high communication efficiency" [42].

Studies have shown that programmatic change communication is associated with problems such as alienation of employees, information overload, and growing cynicism regarding both the reform and top management [12,42,43]. Given its similarities to the hard regulation of innovation [13], programmatic change communication may be expected to influence the personnel's intrinsic motivation, even though some research shows that this may not hamper the implementation as such [24]. However, the effects of programmatic change seem to be partly dependent on the channel of communication used. Common communication channels include general information meetings, posted information, and emailed information. In an early review of the field, Lewis [44] found that general information meetings and small informal discussions were the most commonly used channels for disseminating information about organizational changes. These results were confirmed by subsequent research [12,42]. Similarly, Ohemeng et al [26] found that workshops, seminars, training, one-on-one communication, and unit meetings were the main channels in attempts to manage perception.

As Open Notes was an indirect implementation, it is likely that programmatic change communication influenced both the way the employees perceived the information and the value they assigned to different information channels. Given the emphasis in earlier studies on "equal and fair" information, an important issue in the health care context is whether different professional groups perceived the communication to be in line with their professional norms and values.

Methods

Design

This study used a mixed methods approach with a sequential design [33]. There is thus a chronological link between the qualitative and quantitative data included in the study. First, we attended meetings, studied documents, and made observations to investigate the strategies behind the decisions about what information and communication professionals were deemed to need before implementation, and from which media they could access it. Thereafter, we designed a baseline survey and sent it to the professionals to investigate the actual use of media by different professional groups and how they perceived the reform.

XSL•FO
RenderX

Empirical Material

Participation in Meetings

A multiprofessional working group was established in the autumn of 2013 in the division of psychiatry in Region Skåne. The group consisted of one professional from each of the four geographical administrative areas for adult psychiatry, one professional from child and youth psychiatry, one professional from forensic psychiatry, representatives of the communication department, technical developers, and representatives of patient organizations. The head physician led the working group and reported to management at the division of psychiatry. The working group held regular meetings to discuss and make decisions on the introduction, information, implementation, and development of Open Notes in the division of psychiatry. One of the authors of this paper attended and took field notes during 20 meetings from spring 2015 onward. These notes include summaries of important discussions and reflections on topics discussed at each meeting. The notes were used to define the strategies used as well as the perceptions of the reform.

Document Study

In spring 2015, representatives of the working group carried out a risk analysis to identify risks before implementing Open Notes in adult psychiatry. The risk analysis was a source of information when creating the questionnaire for the baseline survey and the strategies for information and communication.

Observation of Education Events

One of the authors attended eight educational events that were held for professionals in adult psychiatry before the implementation. Our aim was to gain knowledge about the content of the education and about the questions raised by professionals in the division of psychiatry in Region Skåne. Field notes, focusing on important questions and discussions, were used to document the eight observations.

Baseline Survey

The baseline survey used in this study is based on the survey developed and implemented by the OpenNotes Project in the United States [45]. The original English version of the survey was translated and adapted to fit the Swedish context. The survey includes items concerning Open Notes and the work environment of the professionals. It was tested on two representative members of the working group and was then sent to all individuals employed in adult psychiatry in the region (N=3017). Four reminders were sent. As the survey closed 3 days before patients gained online access to their EHRs, all of the material in the baseline study was collected before the implementation.

The survey data reported in this article include demographic data on the participants' professions and the results from two of the fixed-choice questions: one about the communication process and one about the implementation process. The results from one open-ended question about the information campaign are also included. These three questions were developed for the Swedish version of the survey. In the first question, health professionals were asked to report where they had received information about the reform. In conjunction with this fixed-choice question, there was an open-ended question in which the respondents could elaborate on how they perceived the information campaign. In the third question, respondents were asked to choose 5 out of 11 aspects they perceived as important for the implementation. They were then asked to rank these 5 aspects on a scale of 1 to 5, with 1 being the least important and 5 the most important. The responses to the open-ended question were subsequently categorized depending on whether the responses dealt with the content of the information provided or the form of the information.

Ethics

The authors followed the guidelines on research ethics issued by the Swedish Research Council [46]. This study did not deal with any sensitive information, and according to Swedish regulations did not require ethical approval. Potential survey respondents were provided with information about the survey and its purpose in a prenotification email and a cover letter. The information stated that participation was voluntary and that withdrawal at any time without explanation was permitted, and further explained the confidentiality of the treatment and presentation of data.

Data Analysis

The empirical material from the document studies, working group meetings, and educational events were analyzed and are presented as a narrative description in the Results section. The survey material was coded in Excel and imported into SPSS Statistics 23. The χ^2 test was used to test differences between each profession and the rest of the professionals. All reported P values are two-sided. $P<.01$ was considered statistically significant.

Results

Demographics of the Survey Respondents

The response rate to the baseline survey was 28.87% (871/3017). The questionnaire was distributed to professionals in both permanent and temporary positions, which may have influenced the response rate negatively. Table 1 presents the demographics of the respondents and the entire population.

As the survey is a population study, it was important to investigate whether the 853 respondents were representative of the full population. The distribution of the different professions corresponds well with the overall percentage of professionals in each profession in the region. The survey population was compared with demographic information on all professionals in the field of adult psychiatry in Region Skåne. The comparison showed that the response rate was consistent for medical secretaries, a few percentage points lower for nurses and assistant nurses, and slightly higher for the other professional groups. All deviations were less than 10% (Table 1).

XSL•FO
RenderX

Table 1. Demographics of the respondents.

Characteristic	Survey respondents (N=871), n (%)	Population of the region (%)
Professional affiliation[a]		
Doctor	133 (15.6)	11
Medical secretary	76 (8.9)	8
Psychologist	91 (10.7)	16[b]
Nurse	228 (26.7)	28
Assistant nurse	182 (21.3)	29
Sociotherapist[c]	90 (10.6)	__[d]
Other	53 (6.2)	N/A[e]
Gender[f]		
Male	223 (26.2)	__[g]
Female	628 (73.8)	__[g]

[a]853 of the 871 respondents answered the question about their professional affiliation.

[b]Social workers, occupational therapists, physical therapists, and psychologists are included in the same group for the total region.

[c]Includes social workers, occupational therapists, and physiotherapists.

[d]Included in the psychologist category.

[e]N/A: not applicable.

[f]851 of the 871 respondents answered the question about their gender.

[g]No information available.

Communication Channels

This section deals with the first research question, which concerns the strategies underlying the information and communication activities prior to implementation. As mentioned in the Methods section, a multiprofessional working group was established in 2013. The group comprised professionals from different parts of the region and was intended to be representative of the professions as well as the different geographic areas.

The working group had regular meetings to discuss, plan, and make decisions on the introduction and implementation of Open Notes in adult psychiatry. Educational events were also planned. As this was the first psychiatric setting in Sweden to implement Open Notes, many questions had to be addressed before the service could be implemented. The working group decided that a risk analysis was needed.

A new group consisting of employees from different professions and a few members of the working group was asked to carry out the analysis. The risk analysis was performed at the beginning of 2015. The aim was to identify risks to patient safety in connection with the implementation and use of Open Notes in psychiatry, and to identify possible risks for patients' relatives and professionals. Another aim was to identify the benefits of Open Notes for patients and health care. The risk analysis group report mentioned the need for information to be given to professionals to safeguard patients. The analysis suggested that this information should be available on the intranet, where a site was developed and continuously updated. However, the working group realized that this was not sufficient since they became aware that some professionals felt that they had not received any information about the implementation. Consequently, the working group decided that more communication channels were needed. It was considered crucial to use all suitable communication channels to make professionals aware of the change and ensure that they understood the planned implementation. The choice of media was based on previous experience (ie, "how we used to do it") rather than on the specific characteristics of Open Notes.

Before the implementation, the following communication and information activities were carried out: (i) information was posted on Region Skåne's intranet about the implementation, (ii) information emails were sent to professionals by managers, (iii) information/education in two identical 1.5-hour sessions (one morning and one afternoon session) was provided in each of the four geographic areas in the spring of 2015, and (iv) information/communication was provided at workplace meetings.

Information/communication was also available at professional staff meetings arranged by unions. The first two activities were based on a push strategy and a one-way transmission model of communication, whereas the last two and the union meetings enabled opportunities for sense-making through dialog and feedback [26].

The working group considered the information/education events the most important change communication effort because they enabled more symmetrical communication. These events were used to inform health professionals about the implementation and give them opportunities to raise questions, participate, and become involved in their workplace. In other words, the working group aimed to change the professionals' perceptions by

XSL•FO
RenderX

applying both rich information strategies [47], which provided a base for interactions and collective interpretations [22], and less rich information strategies.

In total, approximately 300 professionals attended the eight information/educational events in late April and early May 2015. The project manager for the implementation of Open Notes in Region Skåne was responsible for each event, together with a local representative from the working group. Thus, different individuals were responsible for the information at different geographic locations, which resulted in slightly different presentations of Open Notes at each event. The events consisted of a video with general information about Open Notes, followed by a PowerPoint presentation about the decision-making process prior to implementation, the advantages of Open Notes, and the identified risks. Both the video and the PowerPoint introduced the reform rather superficially. There was also a demonstration of what the interface would look like for patients. There were opportunities to ask questions and discuss the implementation.

The professionals' questions were mainly about the technical prerequisites, the new routines with confidentiality checks when an entry was written in the health record, how information from relatives should be handled safely, and the need for more information about the implementation. As neither the full technical prerequisites nor the implementation date were clear when the events took place, it was not possible to answer some of the questions that the professionals considered important. This presented a major communication challenge.

In summary, patient security was the main focus of the working group, and the information given to professionals was in accordance with this focus. The media used were routinely chosen and the opportunity for "richer" information was limited by the state of the development of the technology at the time of the information/educational events.

Use of Different Communication Channels Among Professionals

This section deals with the second research question, which concerns the media used to inform and change the perceptions of professional groups prior to the implementation. In particular, we focus on how well management was able to communicate information to professionals in adult psychiatry using these media.

As different media were available, the focus was on the media the professions normally used. Table 2 presents the results, showing that the respondents received and absorbed information through a variety of channels. It is important to note that the questionnaire allowed the respondents to choose multiple responses; therefore, the percentages in the total number of responses column in Table 2 add up to more than 100%. There were 1750 responses to this question. Responses from those who did not state their profession were excluded. The results show that media that allowed for dialog and rich information predominated. Overall, 49% of the respondents stated that they received information at a workplace meeting, 25% from informal conversations with colleagues, and 14% at an education event held in the spring of 2015. The unidirectional channel of the intranet was the medium of information for 40% of the professionals, and 16% received information through mass media. The classification of email under dialog media depends on whether the receiver perceived it possible to respond by asking questions; 38% indicated that they received information through email. It is noteworthy that 7% of the professionals claimed that they had not received any information.

Considering the differences among individual professions, doctors stood out as obtaining information through professional meetings and informal conversations substantially more than the rest of the respondents, and significantly less through workplace meetings. In other words, their communication largely took place through rich channels with the ability to shape perceptions. By contrast, the medical secretaries informed themselves through the intranet significantly more than the total number of respondents (ie, they used unidirectional, less rich channels). Psychologists and sociotherapists received information through workplace meetings to a significantly higher degree. In addition, psychologists gained information through informal conversations, whereas sociotherapists gained information through an education event. Assistant nurses took part in education events to a significantly lower degree and used email significantly more often when compared to the total number of respondents. The nurses, assistant nurses, and sociotherapists gained significantly less information through meetings with fellow professionals than the total number of respondents.

In conjunction with the fixed-choice question, there was an open-ended question where the respondents could elaborate on how they perceived the information campaign. Among the 871 professionals, 92 (10.6%) responded with free-form text to the question, "Do you have any further comments on the information surrounding Open Notes?" The free-form text answers dealt either with the content of the information or the way the information was transferred. First, there were requests for a different type of educational event with more detailed information about such matters as the technical prerequisites for the Open Notes system and legal issues surrounding the new transparency of the contents of health records. There were also requests for clearer and more substantial content beyond information about the fact that Open Notes was going to be implemented in adult psychiatry. Second, there were comments about the information process, with a desire for more dialog and two-way communication for the professionals before and during the implementation process. Some also wished that the educational events had been more frequent and held in more locations.

XSL•FO
RenderX

Table 2. Responses to the statement "I have received information about online patient access to their electronic health records in adult psychiatry through...(you can choose several responses to this statement)" (N=1750).[a,b]

Information source	Doctor (N=132), n (%)	Medical secretary (N=73), n (%)	Psychologist (N=91), n (%)	Nurse (N=224), n (%)	Assistant nurse (N=180), n (%)	Sociotherapist (N=89), n (%)	Total responses, n (%)
Workplace meeting	38 (28.8)[c]	36 (49.3)	60 (65.9)[c]	120 (53.6)	80 (44.4)	57 (64.0)[c]	414 (49)
Intranet	45 (34.1)	38 (52.1)[c]	35 (38.5)	89 (39.7)	63 (35.0)	42 (47.2)	342 (40)
Email	59 (44.7)	29 (39.7)	27 (29.7)	77 (34.4)	84 (46.7)[c]	26 (29.2)	320 (38)
Informal conversation	47 (35.6)[c]	19 (26.0)	32 (35.2)[c]	54 (24.1)	32 (17.8)	18 (20.2)	211 (25)
Mass media	29 (22.0)	6 (8.2)	13 (14.3)	44 (19.6)	28 (15.6)	8 (8.9)	134 (16)
Education event	17 (12.9)	16 (21.9)	12 (13.2)	25 (11.2)	15 (8.3)[c]	22 (24.7)[c]	122 (14)
Professional meeting	66 (50.0)[c]	5 (6.8)	7 (7.7)	10 (4.5)[c]	3 (1.7)[c]	1 (1.1)[c]	110 (13)
Social media	4 (3.0)	4 (5.5)	1 (1.1)	11 (4.9)	9 (5.0)	2 (2.2)	35 (4)
No information	8 (6.1)	5 (6.8)	7 (7.7)	18 (8.0)	19 (10.6)	2 (2.2)	62 (7)

[a]Note to interpret the percentages in this table: As an example, 28.8% of doctors stated that they received information from workplace meetings, 34.1% of doctors replied intranet, 44.7% of doctors replied email, and so on.

[b]Since multiple responses were possible, the percentages are above 100%.

[c]$P<.01$ compared with all other professional groups.

Importance of Different Aspects of the Implementation Process

The third research question was related to the aspects the professionals perceived as important for the implementation. Ohemeng et al [26] describe this as "the third step…where the stakeholders attempt to make sense by trying to figure out the meaning of the proposed vision, and revising their understanding." Eleven aspects of the implementation were listed, and the professionals were asked to rank the five most important aspects on a scale of 1 to 5 (with 1 as the least important and 5 as the most important). Table 3 summarizes the total number of respondents who mentioned an aspect, the total ranking scores, and the mean. Table 4 shows the results for the different professions in terms of the percentage of professionals who mentioned the aspect, the mean of the ranking scores given by the professionals who mentioned the aspect, and the importance rank the professionals assigned to the aspect.

Overall, the most frequently chosen aspect was "Evaluation of the Open Notes service," although "Patient safety" had the highest total score and also the highest mean value (see Table 3). It is interesting to note that the aspect receiving the lowest score was "A support line for professionals." This aspect also had the lowest total score and the lowest mean value. However, the differences between the professional groups were small (see Table 4). Reviewing the aspects ranked as the five highest (according to their means) shows that all professional groups included "Patient safety," "Information to professionals," and "Professionals' participation in the process." The medical secretaries diverged the most from the general picture in that they ranked "Information to professionals" as the most important, whereas "Information to patients" was outside of the five highest means for this group. Again, note that the differences were small and nonsignificant. Both the medical secretaries and the doctors included "System reliability" among the first five aspects. However, the largest difference is that the importance of "Patient safety" only ranked in the third priority for the medical secretaries, whereas it was ranked of primary importance for all other groups. A tentative conclusion would be that perceptions differ between health care personnel and administrative personnel.

XSL•FO
RenderX

Table 3. Total scores of importance of different aspects of the implementation.

Aspects	Answers (n)	Total score	Mean
Information to management	201	532	2.65
Information to professionals	399	1334	3.34
Education for professionals	421	1383	3.29
Professionals' participation in the process	329	1025	3.12
A support line for professionals	156	388	2.49
Information to patients	477	1526	3.20
Patient safety	503	1814	3.61
A support line for patients	246	663	2.70
System reliability	283	823	2.91
System fitness for use and clarity	316	801	2.53
Evaluation of the Open Notes service	536	1445	2.70

XSL•FO
RenderX

Table 4. Responses to the statement "Choose five aspects you perceive as important for the Open Notes implementation. Rank the most important 5, the next most important 4, etc, down to 1."

Aspects	Doctors	Medical secretaries	Psychologists	Nurses	Assistant nurses	Sociotherapists, etc
Information to management						
%	19	29	23	20	30	20
Mean	2.36	2.86	2.52	2.76	2.82	2.11
Importance	__[a]	—	—	—	—	—
Information to professionals						
%	33	49	53	50	47	50
Mean	3.36	3.68	3.15	3.46	3.43	2.82
Importance	2	1	4	2	3	—
Education for professionals						
%	47	55	50	49	47	54
Mean	2.95	3.60	3.33	3.23	3.45	3.27
Importance	—	2	2	4	2	3
Professionals' participation in the process						
%	39	29	45	41	37	38
Mean	3.19	2.91	3.29	3.01	3.01	3.21
Importance	3	4	3	5	5	4
A support line for professionals						
%	23	17	15	16	19	18
Mean	2.61	2.38	2.71	2.58	2.47	2.31
Importance	—	—	—	—	—	—
Information to patients						
%	47	58	20	61	51	58
Mean	3.06	2.73	3.14	3.27	3.27	3.38
Importance	5	—	5	3	4	2
Patient safety						
%	59	53	63	59	59	62
Mean	3.63	3.40	3.88	3.52	3.48	3.77
Importance	1	3	1	1	1	1
A support line for patients						
%	39	25	20	28	28	31
Mean	2.67	2.84	2.44	2.86	2.80	2.43
Importance	—	—	—	—	—	—
System reliability						
%	35	26	32	35	25	39
Mean	3.15	2.80	2.52	2.94	2.76	2.49
Importance	4	5	—	—	—	—
System fitness for use and clarity						
%	41	25	39	37	35	39
Mean	2.69	2.63	2.34	2.58	2.47	2.43
Importance	—	—	—	—	—	—
Evaluation of the Open Notes service						

XSL•FO
RenderX

Aspects	Doctors	Medical secretaries	Psychologists	Nurses	Assistant nurses	Sociotherapists, etc
%	74	46	70	60	55	66
Mean	2.91	2.74	2.66	2.48	2.62	2.92
Importance	—	—	—	—	—	5

[a]Not ranked in the top 5.

Discussion

Principal Findings

The Open Notes reform was novel both as a technological innovation and because the receivers were defined as the patients rather than the organization or its professionals. The focus on patients and patient security therefore dominates the implementation [48]. The risk analysis performed by the working group amplified the importance of the patient. Informing the professionals was perceived as a step in the strategy to ensure patient safety, implying that other issues such as the professionals' work situation were of secondary concern when formulating the communication strategy. The communication strategies selected by the working group consisted both of richer media such as education and meetings, and of less rich media such as email and intranet pages. As mentioned above, the implementation was a "telling and selling" process focused on giving information, even though the education and meeting may also be viewed as a way to change perceptions.

Given the focus on the patient as a receiver of the reform, the emphasis on giving information rather than changing perceptions is not surprising. The communication strategies were uniform; that is, all media sources were intended to inform all groups of professionals. Where perceptions might be changed (eg, at education events), the information was provided through a video and PowerPoint presentation with general content. However, the presentations varied depending on the person leading the discussion, who was responsible for presenting the information, as well as on the stage of technological development of the Open Notes service at that time. It appears that the choice of communication strategies was based on the perception that the reform was unambiguous [30] and would have low technical complexity [10] in the eyes of the professionals. Because the working group did not view the reform as opposed to the values of the adapter [10], there was no perceived need to distinguish between professional groups.

These features of the communication strategy are not surprising given that the effect of the reform on health professionals was viewed as indirect; they were viewed as intermediaries rather than receivers of the reform. This supports our finding that although the implementers did not ignore the professionals' need for information and the need to change their perceptions, these concerns were regarded as background concerns rather than as key issues.

The pattern of reception of the information indicates that dialog media in workplace meetings were the most common modes of absorbing information, whereas intranet and email were the second most commonly used media. However, when considering the pattern across professions, a scale of media use connected to social status and workplace organization becomes visible. The scale is bookended by the doctors and the medical secretaries. The latter group relied mainly on nondialogic media, whereas the doctors primarily used dialogic information channels and mostly gained their information at professional meetings (ie, through their peers) or through informal conversations. The use of media by other professions was distributed between these two groups in a way that largely reflects the traditional order of professional status. However, work organization also appears to have an impact. For example, to obtain information at meetings, one has to participate, and for professional meetings to be important, a strong union is needed. Participation in workplace meetings is easier to achieve because the professionals must attend and because there is a sense of belonging to the workplace. However, doctors often obtained their information at internal professional meetings. As the union organized these meetings, management influence over the information given was low. This raises the issue of whose perception of the reform is diffused, and how this affects the implementation.

The use of nondialogic media was rather high for all personnel groups. This is not surprising as computers are a standard work tool in the health care sector in Sweden. The routine use of computers means that media distributed by computer are easily accessible, making obtaining new information part of the everyday routine of accessing information at the workplace. Previous research emphasizes that nondialogic media imply less rich information. It is likely that email is often one-way communication and is perceived as a hard regulation for diffusing information [13,24]. It follows that information transmitted by email shapes perceptions to a lesser degree and infuses a sense of incapacity.

Despite all of the different communication channels used in the communication campaign, the free-form text answers revealed that information had not reached all professionals or, at least, they had not all received the necessary information. It is also noteworthy that 7% of the survey respondents stated that they had received no information at all, despite all efforts made by management. According to the data (see Table 2), assistant nurses were the largest group in this category, with 11% responding that they had received no information. They were also the largest group receiving information by email, whereas primary sources of information for the other professional groups were either workplace or professional meetings. The reason for this finding is outside the scope of this study, but it indicates that either work organization or professional status is important when using information channels. In addition, the answers to the open-ended question indicate a need for more dialogic media and more substantive information. This indicates the importance

XSL•FO
RenderX

of using rich media when innovations are perceived as complex by the personnel (even if not by the implementers).

The medical secretaries stood out in regard to the aspects perceived as important in the implementation process. In contrast to all other personnel groups who ranked "Patient safety" as the highest on average, the medical secretaries showed higher average rankings for both "Information to professionals" and "Education for professionals." However, these differences were small. Overall, one can discern a tendency to emphasize aspects related to patients and personnel rather than technical concerns. This indicates that the social and organizational aspects of implementing Open Notes are the salient issues for the personnel in psychiatric care. The sense-making showed primary consideration for the patient, followed by the professionals. This emphasis on the patient is not surprising given that client care is the normative basis for most professions [31,38]. It is likely that the value of "patient safety" promoted by the working group was already embedded in the professionals' norm system. The reaction from the respondents is thus not surprising, but this finding enriches research from a perception management perspective by introducing consideration of the likely effect of the professional norms of the receiver on their sense-making of information. These results also suggest that when formulating a message, attention should be paid to the values and social aspects that are important to the receiver rather than to technical information such as system features. Rich information in this context does not simply imply "a lot of information" but rather information that agrees with or affects with the interpretative frames connected to the professions.

From the perspective of perception management, the implementation of Open Notes can be viewed as a hard regulation. The health care personnel had no option but to accept the implementation in the form decided by management. When the professionals "made sense" of the implementation, it was consequently not the technical issues that were in the forefront. Instead, aspects related to patient safety and in-depth information for the professions were salient. This can be interpreted as perception management having succeeded in "selling" the solution, but awakened concerns connected to professional norms and values while doing so.

Limitations

This study has several limitations. First, the response rate to the web questionnaire was only 28.86%. One explanation may be that this was a full population study and some employees were not working during the time when it was possible to answer the survey. Nevertheless, the group distribution among the respondents corresponds well with the percentage of employees in each profession, which indicates that we have good representation of all professional groups. As Open Notes was implemented later in adult psychiatry than in other areas of health care in the region, it is also possible that some of the respondents gained knowledge about the reform from other sources. The items in our survey do not cover this.

The majority of respondents reported that they had gained information from more than one medium. Since the respondents specified more than one medium in the survey, we do not know whether these media complemented or substituted for each other. The study would have been improved if we had also asked which media the responders found to be the most important. This would had given us a firmer base to discuss the importance of media type when interpreting reforms.

Conclusions

This paper makes several contributions. The first is the empirical evidence that different groups of professionals absorb information through different channels when informing themselves about reforms. Our working hypothesis was that health care organizations have "double hierarchies," and that these have to be considered when communicating implementations. The results of this study largely confirm this hypothesis. A main conclusion of the study is that professional association matters both for the choice of information media and for evaluating aspects of the message that comes from a higher level of an organization. Those in groups that are considered "full professions" with high professional status prefer to be informed among their peers, whereas semiprofessionals find other ways to become informed. This difference may be because full professions often have a more stable professional identity and more opportunities to meet fellow professionals. This observation adds to the importance in media selection theory of not only considering the hierarchical levels of the organization but also the different status of the professionals in the organization. We also discerned minor differences between professional groups regarding which issues they perceived to be important in implementing reforms. From the point of view of communications practitioners, this finding implies that communication strategies may be more successful when they combine common information with strategies directed toward different professionals for communicating indirect implementations.

A second contribution is the finding that professional status alone does not determine the choice of information channel. The pattern of media use described is many-sided. Most professional groups mention email or intranet as one source of information. These are channels that are routinely used for diffusing information in an organization. There are also indications that the work organization is important, such as participation in workplace meetings or educational events. Thus, a communication strategy has to consider professional diversity and workplace organization, as well as existing information paths to fully reach out to the receivers.

The third contribution is defining the implementation of Open Notes as an example of an indirect implementation. This feature influenced the perception of the management. From the point of view of the working group, health professionals, however significant, were perceived largely as "tools" to achieve the primary aim of the reform: patient empowerment without risking patient safety. This perception influenced both the channels of communication and the information content. The information to professionals was presented in a routine manner and was sometimes incomplete, possibly because management did not believe that the reform would affect health professionals. However, any implementation may involve new roles for various parties regardless of whether they are directly or indirectly affected. In this case, for example, the professionals will be

XSL•FO
RenderX

meeting more "empowered" patients. An alternative approach is to pay attention to the effect of the implemented reforms on both those directly affected and those indirectly affected. It is likely that implementation of reforms that indirectly affect professions will be more common in the future, as the discourse of patient empowerment is taken up in other areas. However, as a rather new phenomenon, more research is needed both about how indirect implementation of reforms may affect professionals and about the kind of information that is important to ease the implementation.

In conclusion, we have compared the communication strategy regarding choice of media and the most important aspect of the reform as perceived by the receivers. Our main conclusion is that there is a link between the management's (ie, work group's) perception of the main receiver of the reform (here, the patients), and the communication strategy used in the health organization. By contrast, the reception of the information seems to depend on the mix of professions in the organization and their professional norms, as well as on the work organization and routine paths used to disseminate information. However, research on indirect implementation is still in its early stages. More research is needed before these relationships are fully understood. Finding good strategies for providing information to different professional groups will be valuable when communicating with those indirectly affected by an implementation.

How various aspects of communication interact is an important issue for future research. However, the complexity of the use of media revealed in this study indicates that, in general, a multimedia approach may more easily succeed than a single-medium approach.

Acknowledgments

The research presented in this paper is funded by AFA Insurance in Sweden via the project "eHealth Services' Impact on the Working Environment of Health Professionals" (EPSA).

Conflicts of Interest

None declared.

References

1. Kahn MW, Bell SK, Walker J, Delbanco T. A piece of my mind. Let's show patients their mental health records. JAMA 2014 Apr 02;311(13):1291-1292. [doi: 10.1001/jama.2014.1824] [Medline: 24691603]
2. Walker J, Kahn MW, Delbanco T. Transparency in the delivery of mental health care--reply. JAMA 2014 Aug 13;312(6):650-651. [doi: 10.1001/jama.2014.7610] [Medline: 25117139]
3. Dobscha SK, Denneson LM, Jacobson LE, Williams HB, Cromer R, Woods S. VA mental health clinician experiences and attitudes toward OpenNotes. Gen Hosp Psychiatry 2016;38:89-93. [doi: 10.1016/j.genhosppsych.2015.08.001] [Medline: 26380876]
4. Goldfinch S. Pessimism, computer failure, and information systems development in the public sector. Public Admin Rev 2007 Sep;67(5):917-929. [doi: 10.1111/j.1540-6210.2007.00778.x]
5. Stewart J, O'Donnell M. Implementing change in a public agency. Intl Jnl Public Sec Manag 2007 Apr 10;20(3):239-251. [doi: 10.1108/09513550710740634]
6. Brunsson N, Sahlin-Andersson K. Constructing organizations: the example of public sector reform. Organiz Stud 2016 Jun 30;21(4):721-746. [doi: 10.1177/0170840600214003]
7. Lipsky M. Street-level bureaucracy: dilemmas of the individual in public services. New York: Russell Sage Foundation; 1980.
8. Cucciniello M, Lapsley I, Nasi G, Pagliari C. Understanding key factors affecting electronic medical record implementation: a sociotechnical approach. BMC Health Serv Res 2015 Jul 17;15:268 [FREE Full text] [doi: 10.1186/s12913-015-0928-7] [Medline: 26184405]
9. Moullin JC, Sabater-Hernández D, Fernandez-Llimos F, Benrimoj SI. A systematic review of implementation frameworks of innovations in healthcare and resulting generic implementation framework. Health Res Policy Syst 2015 Mar 14;13:16 [FREE Full text] [doi: 10.1186/s12961-015-0005-z] [Medline: 25885055]
10. Rogers E. Diffusion of Innovations. New York: Free Press; 2003.
11. Lehtonen T. DRG-based prospective pricing and case-mix accounting—Exploring the mechanisms of successful implementation. Manag Account Res 2007 Sep;18(3):367-395. [doi: 10.1016/j.mar.2006.12.002]
12. Lewis LK. Employee perspectives on implementation communication as predictors of perceptions of success and resistance. West J Commun 2006 Feb 14;70(1):23-46. [doi: 10.1080/10570310500506631]
13. Mikkelsen MF, Jacobsen CB, Andersen LB. Managing employee motivation: exploring the connections between managers' enforcement actions, employee perceptions, and employee intrinsic motivation. Int Public Manag J 2015 Jul 16;20(2):183-205. [doi: 10.1080/10967494.2015.1043166]
14. Petersson L, Erlingsdóttir G. Open Notes in Swedish psychiatric care (part 1): survey among psychiatric care professionals. JMIR Ment Health 2018 Mar 02;5(1):e11 [FREE Full text] [doi: 10.2196/mental.9140] [Medline: 29396386]

XSL•FO
RenderX

15. Petersson L, Erlingsdóttir G. Open Notes in Swedish psychiatric care (part 2): survey among psychiatric care professionals. JMIR Ment Health 2018 Jun 21;5(2):e10521 [FREE Full text] [doi: 10.2196/10521] [Medline: 29929946]
16. Erlingsdóttir G, Lindholm C. When patient empowerment encounters professional autonomy: The conflict and negotiation process of inscribing an eHealth service. Scand J Public Admin 2015;19(2):27-48 [FREE Full text]
17. Erlingsdóttir G, Petersson L, Jonnergård K. A theoretical twist on the transparency of Open Notes: qualitative analysis of health care professionals' free-text answers. J Med Internet Res 2019 Sep 25;21(9):e14347 [FREE Full text] [doi: 10.2196/14347] [Medline: 31573905]
18. King G, O'Donnell C, Boddy D, Smith F, Heaney D, Mair FS. Boundaries and e-health implementation in health and social care. BMC Med Inform Decis Mak 2012 Sep 07;12:100 [FREE Full text] [doi: 10.1186/1472-6947-12-100] [Medline: 22958223]
19. Cresswell K, Sheikh A. Organizational issues in the implementation and adoption of health information technology innovations: an interpretative review. Int J Med Inform 2013 May;82(5):e73-e86. [doi: 10.1016/j.ijmedinf.2012.10.007] [Medline: 23146626]
20. Murray E, Burns J, May C, Finch T, O'Donnell C, Wallace P, et al. Why is it difficult to implement e-health initiatives? A qualitative study. Implement Sci 2011 Jan 19;6:6 [FREE Full text] [doi: 10.1186/1748-5908-6-6] [Medline: 21244714]
21. Boonstra A, Broekhuis M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. BMC Health Serv Res 2010 Aug 06;10:231 [FREE Full text] [doi: 10.1186/1472-6963-10-231] [Medline: 20691097]
22. Barrett AK, Stephens KK. The pivotal role of change appropriation in the implementation of health care technology. Manag Commun Quart 2016 Dec 18;31(2):163-193. [doi: 10.1177/0893318916682872]
23. Eriksson-Zetterquist U, Lindberg K, Styhre A. When the good times are over: professionals encountering new technology. Hum Relat 2009 Jul 28;62(8):1145-1170. [doi: 10.1177/0018726709334879]
24. Grøn CH. Perceptions unfolded: managerial implementation in perception formation. Int J Public Sec Manag 2018 Jun 06;31(6):710-725. [doi: 10.1108/ijpsm-09-2017-0237]
25. Timmerman CE. Media selection during the implementation of planned organizational change. Manag Commun Quart 2016 Aug 17;16(3):301-340. [doi: 10.1177/0893318902238894]
26. Ohemeng FLK, Amoako Asiedu E, Obuobisa-Darko T. Giving sense and changing perceptions in the implementation of the performance management system in public sector organisations in developing countries. Int J Public Sec Manag 2018 Apr 09;31(3):372-392. [doi: 10.1108/ijpsm-05-2017-0136]
27. Deline MB. Framing resistance: identifying frames that guide resistance interpretations at work. Manag Commun Quart 2018 Sep 14;33(1):39-67. [doi: 10.1177/0893318918793731]
28. Martine T, Cooren F, Bénel A, Zacklad M. What does really matter in technology adoption and use? A CCO approach. Manag Commun Quart 2015 Dec 02;30(2):164-187. [doi: 10.1177/0893318915619012]
29. Donabedian B, McKinnon SM, Bruns WJ. Task characteristics, managerial socialization, and media selection. Manag Commun Quart 2016 Aug 15;11(3):372-400. [doi: 10.1177/0893318998113002]
30. Donabedian B. Optimization and its alternative in media choice: a model of reliance on social-influence processes. Inf Society 2006 Jun;22(3):121-135. [doi: 10.1080/01972240600677771]
31. Abbott A. The system of professions: an essay on the division of expert labor. Chicago: The University of Chicago Press; 1988.
32. Brante T. The professional landscape: the historical development of professions in Sweden. Prof Prof 2013 Dec 12;3(2):558. [doi: 10.7577/pp.558]
33. Teddlie C, Tashakkori A. Foundations of mixed methods research: integrating quantitative and qualitative approaches in the social and behavioral sciences. Los Angeles: SAGE; 2009.
34. Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O. Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Q 2004;82(4):581-629 [FREE Full text] [doi: 10.1111/j.0887-378X.2004.00325.x] [Medline: 15595944]
35. Fidler LA, Johnson JD. Communication and Innovation Implementation. Acad Manag Rev 1984 Oct;9(4):704-711. [doi: 10.5465/amr.1984.4277422]
36. Matland R. Synthesizing the implementation literature: The ambiguity-conflict model of policy implementation. J Public Admin Res Theory 1995;5(2):145-174. [doi: 10.1093/oxfordjournals.jpart.a037242]
37. Markus M, Pfeffer J. Power and the design and implementation of accounting and control systems. Account Organiz Soc 1983;8(2-3):205-218. [doi: 10.1016/0361-3682(83)90028-4]
38. Freidson E. Professionalism: the third logic. Chicago: University of Chicago; 2001.
39. Lewis L. Organizational change: creating change through strategic communication. West Sussex: Wiley-Blackwell; 2011.
40. Klein SM. A management communication strategy for change. J Org Change Manag 1996 Apr;9(2):32-46. [doi: 10.1108/09534819610113720]
41. Constantinides P, Barrett M. Negotiating ICT development and use: The case of a telemedicine system in the healthcare region of Crete. Inf Organ 2006 Jan;16(1):27-55. [doi: 10.1016/j.infoandorg.2005.07.001]

XSL•FO
RenderX

42. Russ TL. Communicating change: a review and critical analysis of programmatic and participatory implementation approaches. J Change Manag 2008 Dec;8(3-4):199-211. [doi: 10.1080/14697010802594604]
43. Qian Y, Daniels TD. A communication model of employee cynicism toward organizational change. Corp Comm 2008 Aug 06;13(3):319-332. [doi: 10.1108/13563280810893689]
44. Lewis LK. Disseminating information and soliciting input during planned organizational change. Manag Commun Quart 2016 Nov 06;13(1):43-75. [doi: 10.1177/0893318999131002]
45. Walker J, Leveille SG, Ngo L, Vodicka E, Darer JD, Dhanireddy S, et al. Inviting patients to read their doctors' notes: patients and doctors look ahead: patient and physician surveys. Ann Intern Med 2011 Dec 20;155(12):811-819 [FREE Full text] [doi: 10.7326/0003-4819-155-12-201112200-00003] [Medline: 22184688]
46. Swedish Research Council. Good Research Practice. URL: https://www.vr.se/download/18.5639980c162791bbfe697882/1555334908942/Good-Research-Practice_VR_2017.pdf [accessed 2020-08-25]
47. Daft RL, Lengel RH, Trevino LK. Message equivocality, media selection, and manager performance: implications for information systems. MIS Quart 1987 Sep;11(3):355. [doi: 10.2307/248682]
48. Nøhr C, Parv L, Kink P, Cummings E, Almond H, Nørgaard JR, et al. Nationwide citizen access to their health data: analysing and comparing experiences in Denmark, Estonia and Australia. BMC Health Serv Res 2017 Aug 07;17(1):534 [FREE Full text] [doi: 10.1186/s12913-017-2482-y] [Medline: 28784173]

Abbreviations

eHealth: electronic health
EHR: electronic health record

Edited by C Lovis; submitted 10.07.20; peer-reviewed by K Morse, D Pförringer; comments to author 31.07.20; revised version received 15.10.20; accepted 02.07.21; published 16.08.21.

Please cite as:
Jonnergård K, Petersson L, Erlingsdóttir G
Communicating the Implementation of Open Notes to Health Care Professionals: Mixed Methods Study
JMIR Med Inform 2021;9(8):e22391
URL: https://medinform.jmir.org/2021/8/e22391
doi:10.2196/22391
PMID:34398794

©Karin Jonnergård, Lena Petersson, Gudbjörg Erlingsdóttir. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 16.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Classification of Electronic Health Record–Related Patient Safety Incidents: Development and Validation Study

Sari Palojoki[1,2*], PhD; Kaija Saranto[3*], PhD; Elina Reponen[2*], MD, PhD; Noora Skants[2*], MD, PhD; Anne Vakkuri[2*], MD, PhD; Riikka Vuokko[1*], PhD

[1]Department of Steering of Healthcare and Social Welfare, Ministry of Social Affairs and Health, Helsinki, Finland

[2]Department of Anesthesiology, Intensive Care and Pain Medicine, Peijas Hospital, Helsinki University Hospital, Vantaa, Finland

[3]Faculty of Social Sciences and Business Studies, University of Eastern Finland, Kuopio, Finland

[*]all authors contributed equally

Corresponding Author:
Sari Palojoki, PhD
Department of Steering of Healthcare and Social Welfare
Ministry of Social Affairs and Health
P.O. Box 33
Helsinki, 00023
Finland
Phone: 358 29516001
Email: sari.palojoki@gmail.com

Abstract

Background: It is assumed that the implementation of health information technology introduces new vulnerabilities within a complex sociotechnical health care system, but no international consensus exists on a standardized format for enhancing the collection, analysis, and interpretation of technology-induced errors.

Objective: This study aims to develop a classification for patient safety incident reporting associated with the use of mature electronic health records (EHRs). It also aims to validate the classification by using a data set of incidents during a 6-month period immediately after the implementation of a new EHR system.

Methods: The starting point of the classification development was the Finnish Technology-Induced Error Risk Assessment Scale tool, based on research on commonly recognized error types. A multiprofessional research team used iterative tests on consensus building to develop a classification system. The final classification, with preliminary descriptions of classes, was validated by applying it to analyze EHR-related error incidents (n=428) during the implementation phase of a new EHR system and also to evaluate this classification's characteristics and applicability for reporting incidents. Interrater agreement was applied.

Results: The number of EHR-related patient safety incidents during the implementation period (n=501) was five-fold when compared with the preimplementation period (n=82). The literature identified new error types that were added to the emerging classification. Error types were adapted iteratively after several test rounds to develop a classification for reporting patient safety incidents in the clinical use of a high-maturity EHR system. Of the 427 classified patient safety incidents, interface problems accounted for 96 (22.5%) incident reports, usability problems for 73 (17.1%), documentation problems for 60 (14.1%), and clinical workflow problems for 33 (7.7%). Altogether, 20.8% (89/427) of reports were related to medication section problems, and downtime problems were rare (n=8). During the classification work, 14.8% (74/501) of reports of the original sample were rejected because of insufficient information, even though the reports were deemed to be related to EHRs. The interrater agreement during the blinded review was 97.7%.

Conclusions: This study presents a new classification for EHR-related patient safety incidents applicable to mature EHRs. The number of EHR-related patient safety incidents during the implementation period may reflect patient safety challenges during the implementation of a new type of high-maturity EHR system. The results indicate that the types of errors previously identified in the literature change with the EHR development cycle.

(JMIR Med Inform 2021;9(8):e30470) doi:10.2196/30470

KEYWORDS

classification; electronic health records; hospitals; medical informatics; patient safety; risk

XSL•FO
RenderX

Introduction

Background

The key components of health information technology (HIT) and electronic health records (EHRs) play a crucial role in patient management, care interventions, and effective health care services [1]. The literature indicates that HIT can improve patient safety and quality of care [2-4]. Despite evidence that improvements have helped with the adoption and implementation of EHR systems, EHR adaptation is not without obstacles or challenges [5,6]. EHR adoption may cause unintended consequences, safety risks, and other outcomes [7-9].

Data on error types specifically for high-maturity EHRs [10-12] remain scarce, and available studies have focused on EHRs from the earlier development stages; otherwise, the development stage is not described in detail [13]. Varied patient safety issues related to EHRs and documented in research include poor usability, inadequate communication of laboratory test results, EHR downtime, system-to-system interface incompatibilities, drug overdoses, inaccurate patient identification, care-related timing errors, and incorrect graphical display of test results [14-20].

Many researchers share the view that technology-induced errors arise from several sources in a complex health care environment [6-8,15,21]. Risks associated with EHRs have been identified as being related to technologies, apps, and their use [21-24]. Many EHR errors are latent and involve technological features, user behavior, and regulations, thereby making error anticipation challenging while underscoring the importance of identifying vulnerable areas [25]. The patient safety incident reporting system is fundamental to obtaining and processing patient safety–related information for improving work. Incident reporting aims to detect problems and investigate underlying causes; as a result, there is a possibility of using organizational learning to prevent such incidents from happening again [26-29].

In 2012, the Institute of Medicine recommended that information produced by HIT-related patient safety incidents should be used to improve patient safety [30]. The open sharing of HIT-related patient safety incident data using a uniform structure or other standards could help institutions learn the best practices for EHR implementation. Simultaneously, it is essential to recognize the limitations of patient safety incident reporting to avoid data misinterpretation. However, this information is not shared frequently, so organizations are constantly reinventing the wheel to address EHR issues and improve functionality [2,31]. There is a concern that benefits from HIT-related safety data are lost because of the absence of a mechanism to classify HIT-related events; yet, it is not well established how to define and classify incidents in these systems [19,28,29]. It has been suggested that research evidence, testing, and development of classifications applicable specifically for high-maturity EHRs are needed [10-12,28].

Implementing or upgrading an EHR system is a major endeavor for health care organizations. Decisions on the implementation process, such as user training and customization of the product, can have long-term implications on the usability of EHRs and thus safety related to EHR use [12,32-34]. Our capacity to reap the benefits of new technologies and manage new threats is contingent on understanding the potential threats to patient safety [19]. In the following sections, we describe our study design and results after developing and testing a new problem classification for reporting patient safety incidents while implementing and using a high-maturity EHR system [10-13]. Implementation of this system occurred in a Finnish university hospital with a first go-live phase that began in 2018. For clinical personnel, this meant a change from a previous EHR system to a new high-maturity EHR system. Our research data comprised incident reports from periods as early as 6 months before implementation and as late as 6 months immediately following the beginning of implementation.

Objectives

The aims of our study are specified as follows:

1. Our primary aim is to develop an error classification applicable to EHR-related patient safety incidents involving high-maturity EHRs.
2. Our secondary aim is to validate technology-induced error classification using real-world patient safety incidents, including the assessment of interrater agreement.

Methods

Study Design

A study design was proposed to develop and validate a classification for patient safety incidents. In this study, the concept of technology-induced errors was applied to define EHR-related patient safety incidents [35]. Classifications and taxonomies are used widely in clinical contexts; however, in the literature, they are based on practical needs to standardize medical data in documentation, with less emphasis on theorizing and characterizing classifications and other terminological systems [36,37]. In a clinical setting, classifications can be applied for various reasons, for example, to support clinical thinking to help establish guidelines for diagnosis and treatment [38]. The classification and other core concepts used in this study are listed in Textbox 1. Our primary focus—developing a classification for technology-induced errors—was based on previous research; however, we assumed that further development was required for the classification to be applicable with high-maturity EHRs. At a conceptual level, error classification captures both the instance and its conditions portrayed in patient safety incident reports. However, in the class descriptions, we also used the term *problem* to describe the reporting professional's experience of a situation that needs to be reported and remedied.

XSL•FO
RenderX

Textbox 1. Key concepts and abbreviations used in this study.

Classification (taxonomy)

- Taxonomies (classifications) are modes of information management that have been used successfully in areas such as medicine and information technology to describe, classify, and organize items based on common features. In this paper, we use the term

  classification

  [36,38,39].

Technology-induced errors

- These errors result from the design and development of technology, the implementation and customization of a technology, and the interplay between the operation of a technology and the new work processes that arise from the use of technology [35,40].

Electronic health records (electronic medical record and electronic patient record)

- Medical Subject Headings conceptualizes electronic health records as "media that facilitate transportability of pertinent information concerning (a) patient's illness across varied providers and geographic locations." Synonyms for electronic health records include electronic patient records, electronic medical records, computerized patient records, and digital medical records. In hospitals, electronic health records are often software apps that contain or interact with other apps. They cover apps for computerized provider order entry, clinical decision support, test results storage, and medication administration systems. These software apps need networked hardware and clinical data structures to operate [41,42]. In this paper, we use the abbreviation electronic health record.

Electronic health record (Electronic Medical Record) Maturity Model

- One of the electronic health record maturity models is the Electronic Medical Record Adoption Model, developed by Healthcare Information and Management Systems Society Analytics. It has become a universally recognized maturation model of a hospital's electronic medical record environment. The Electronic Medical Record Adoption Model is an eight-stage maturation model that reflects hospitals' electronic medical record capabilities, ranging from a completely paper-based environment (stage 0) to a highly advanced paperless and digital patient record environment (stage 7) [10-12]

Our starting point for the classification development in this study was based on previous research by Sittig and Singh [21,22]. The initial coding framework followed the structure of the Finnish Technology-Induced Error Risk Assessment tool comprising eight main categories: EHR downtime; system-to-system interface errors; open, incomplete, or missing orders; incorrect identification; time measurement errors; incorrect item selected; failure to heed a computer-generated alert; and failure to find or use the most recent patient data [14,15,43]. This tool-based coding framework was refined and extended through analysis and development by our research team based on the clinical experience of medical doctors using the studied EHR.

In addition to data-based analysis, to review and update the classification based on the latest research, articles on EHR error types were gathered from PubMed (MEDLINE complete). We searched for EHR error types with Medical Subject Headings using the keywords *electronic health records*, *patient safety*, and *medical informatics*, and *technology-induced error* was applied as a search term, although it is not yet a Medical Subject Heading term.

Study Materials and Research Context

We collected patient safety incident reports, which illustrate typical errors with an older EHR system and a new system to be implemented in a Finnish university hospital. The hospital district is among the largest in Finland, with 25,916 employees. In 2019, 680,000 patients were treated at the hospital, with 2.9 million outpatient visits and 92,000 surgeries performed. Since 2007, the hospital has been using a fully paperless EHR system [15,44]. The implementation of a new high-maturity EHR (Healthcare Information and Management Systems Society 6-7) started in 2018 at the first site to cover emergency services and several medical specialties. Data on all types of patient safety incident reports and 12 medical specialties were retrieved on July 17, 2020, from the university hospital's database.

The research data used comprised real-world patient safety incident data to develop and assess the emerging classification identified in the literature and in previous studies and expanded in our research. The Finnish patient safety incident reporting model and instrument, called HaiPro (Awanic), was developed in 2006. It is anonymous, nonpunitive, and not integrated into any EHR system. All personnel—including all nurses, physicians, and academic hospital workers (eg, pharmacists)—have been trained and are encouraged to report patient safety incidents through HaiPro. Although HaiPro contains structured data, the main content of incident reports is descriptive [44].

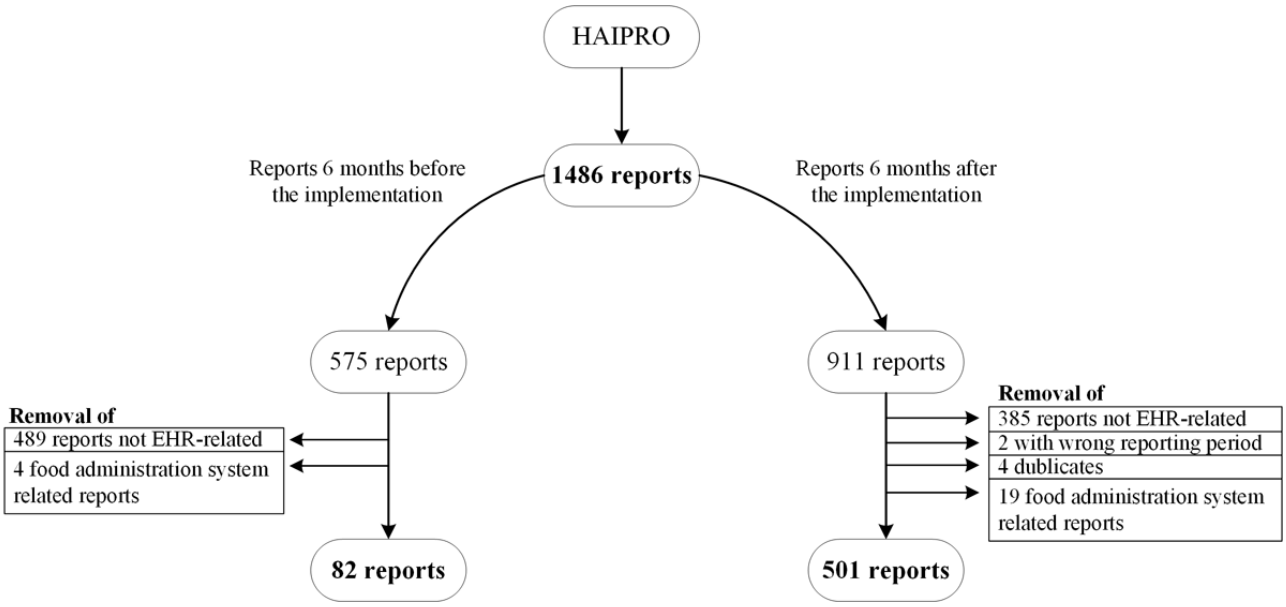
The research review process of the university hospital organization approved the study protocol (study permission update March 23, 2020, License org.id/200/2020). In the collected research data, no connection to patients or professionals exists because of the nature of the anonymized data, which do not contain any identification details. Psychiatric reports were excluded because of data sensitivity. To allow for comparisons in terms of the number of patient safety incident reports, we included all safety incidents reported through the HaiPro system during the 6-month period before the implementation of a new EHR system in 2018. A similar selection process with a full reading of reports was also applied during the implementation.

XSL•FO
RenderX

Data Cleaning, Data Analysis, and Validation

The incident report data were processed before starting the analysis, as shown in Figure 1. To clean up the research data, patient safety and informatics experts read all the reports in the database thoroughly to identify the EHR-related cases. Duplicates and reports concerning food administration information systems have been removed. All reports that met the inclusion criteria (EHR related) were selected for this study. Two clinical experts (medical doctors) with 2 years of experience in implementing and studying EHR systems and extensive experience with patient safety reporting made detailed and documented decisions on cases in which the definition of EHR-related error incidents was not clear. Our research team comprised 3 clinicians with 3 clinical informatics and classification experts.

Figure 1. The process of categorizing the reports for data analysis with 82 reports from before the implementation and 501 reports from after the implementation remaining for a blinded review with the proposed classification. EHR: electronic health record.



For our research purposes, 82 reports from before the implementation and 501 reports from after the implementation remained for a blinded review with the proposed classification. The process of classifying the data and reviewing the results are presented in Textbox 2. A more detailed process of data analysis and validation is provided in Multimedia Appendix 1. During the data analysis and review of the research team, we developed the original classification by adding several classes or subcategories. Finally, we validated the classes based on the distribution of incidents.

Textbox 2. Study design for data analysis and validation.

Patient safety incident report data quality analysis and validation

- Agreement upon preparatory classes and their descriptions; common rules for classifying data
- Blinded reviews of the data with the classification; research team agreements for classification revisions and refinement
- Blinded testing of revised classification
- Classification validation and results from data analysis finalized

First, to perform a classification-based analysis, the research team agreed on preparatory classes and their descriptions at the start of data analysis, as well as common classification rules. During the next research phase, 2 researchers with substantial experience in classification development and informatics independently reviewed a set of reports and applied the classification in a blinded fashion, along with 2 researchers with clinical experience. Disagreements were discussed among the research team, and the study design was adjusted accordingly. After each set of test rounds, the interpretation of the classes was discussed to update the wording of the classes and their descriptions. Altogether, seven classification rounds for multidisciplinary consensus and validation procedures were conducted to perform the iterative development of the emerging classification (Multimedia Appendix 2). Selecting the same main category created a match while choosing a different category or failing to find the category at all was viewed as a nonmatch. Disagreements were discussed by the research team. Percentage agreement was applied to perform the interrater reliability measurement.

During the third research phase, informatics and clinical experts tested the revised version of the classification to validate the results. Finally, the data analysis was completed after a 7-month research period that ended in March 2021. The research team reviewed the final results and revised the classification by refining the descriptions of the final classes.

XSL•FO
RenderX

Results

Overview

Here, we present the results from the patient safety incident report data analysis based on the results from the error classification that emerged during our iterative data analysis. In addition to presenting the results from validation, we also present observations regarding the development of the classification. Development needs for an original structure were realized during the analysis, and more subclasses were needed.

Data Analysis

The total number of all types of patient safety incident reports (excluding psychiatry) during a 1-year period was 1486. There were 38.69% (575/1486) reports during the 6-month period before the implementation of a new EHR system, of which 14.2% (82/575) of cases were related to EHRs. Altogether, 61.31% (911/1486) of reports were entered into the database 6 months after the implementation of a new EHR system, of which EHR-related incidents totaled 54.9% (501/911).

The total reporting volume during the implementation phase increased by 58.5%, with the number of cases related to the EHR system during the postimplementation period was five times higher (510%) than before implementation.

During classification, 14.8% (74/501) of EHR-related incident reports were rejected and thus remained unclassified. Decisions concerned situations wherein information was insufficient to classify the event reliably, or it was possible that the notification was not related to the EHR system.

The interrater agreement was 97.7%. During the blinded review, 10 discrepancies between reviewers were found in the final data (n=427), which were accepted for the classified data. Moreover, the previously mentioned rejected incident reports created discrepancies during the classification.

Validation of Classification

Our final analyses of EHR-related error types comprised 427 classified incidents. A detailed distribution (classification and frequencies of error types by main categories and subcategories) is provided in Multimedia Appendix 2.

The downtime (8/427, 1.9%) category was associated with the problem of logging into a single part of the EHR system or application (2), or the entire EHR system (3), whereas the presence of planned downtime existed only in one report. An unplanned downtime did not exist in the research data. During classification with our data, we noticed that not all incidents fit the existing subcategories. We added a new subcategory for data entry during and after a period of downtime, and we split the system-logging-problem subcategory to relate to all or part of the system in use to better capture issues with a high-maturity EHR system.

Among the 22.5% (96/427) of interface problems, 36% (35/96) of incidents were found in the category of data transfer between different EHRs within the same organization. This was caused partially by the implementation that occurred in the first hospital site at that time, and multiple EHR systems were still in use in the entire hospital district. Data transfer within the different components of the same patient information system accounted for 39 incidents. On the basis of our data and classification reviews, we added several subcategories to capture the complex interface issues in EHR adoption, in which transference occurred as a change from one EHR to another in a highly competent environment of clinical and HIT ecosystems.

Problems with timing functions accounted for 5.9% (25/427) of cases. Most of the reports (n=21) concerned changes in medication and treatment scheduling because of the programming logic in the EHR. This category's original classes worked well with the data, but we decided to update the class descriptions to better reflect high-maturity EHRs.

The largest number of cases, at 20.8% (89/427) of reports, was related to the medication section, whereas the smallest number (1/427, 0.2%) was found in the mixed patient record problems category. We noticed that the original classification did not cover these incidents adequately to capture the complex issues; thus, a new class was added after reviewing this incident type in our research team.

The usability problem category (73/427, 17.1%) covered notifications as follows: most reports concerned problems with missing, incomplete, or wrong alarms, or alarm fatigue (n=29) and problems finding data (n=30). Problems with decision support accounted for two reports and printing problems in 11 cases. One of the usability problems remains unspecified. For the subcategories related to alarms, we updated the class descriptions and clarified the characteristics of decision support–related issues as they relate to other alarms or system notifications. After discussing the data analysis findings within the research team, we separated workflow problems from the usability class. As a problem category, workflow problems are typically more complex than mere usability issues.

Clinical workflow problems using EHRs were the underlying causes of errors in 7.7% (33/427) of cases, and competence problems were identified in 5.4% (23/427) cases. These were divided into two subcategories, of which 16 reports cited a lack of education. Obstacles to competence development caused by EHRs were cited in seven incidents. Within the emerging classification, workflow problems were deemed complex situations in which EHRs played a clearly identified role. Typically, these cases occur when the system cannot support the clinical workflow, or when the workflow is interrupted.

The documentation category (60/427, 14.1%) comprises four subcategories, the largest of which turned out to be unspecified documentation issues in its 27 cases. The lack of data structure, errors in data structure, or interpretation problems with data structure appeared in 20 notifications, whereas clinical classification deficiencies were found in one report. The loss of recorded information during documentation was identified as the cause of incidents in 11 cases, a well-established category in previous research. On the basis of our data analysis, we decided to clarify the class descriptions to make it easier for reporting professionals to differentiate documentation incidents from usability problems. Simultaneously, subcategories were added to capture the manifold issues of documenting. Unrecognized problems with data loss form a separate main

XSL•FO
RenderX

class, comprising 2.6% (11/427) of reports. Class descriptions for data loss are also updated to indicate clear differences in usability problems.

An examination of the data revealed that 1.9% (8/427) of cases were related to the category of general situations, in which patient safety is threatened because of the introduction of a patient information system. This class can be used to capture incidents that seem to portray situations involving the poor organization of work during ongoing implementation phases in complex health care environments that, based on our data, typically may include demanding activities such as multitasking, problem solving, and clinical reasoning.

The classification and frequency of error types in the main and subcategories are provided in Multimedia Appendix 2. After the research team agreed to classification updates, the classification system comprised 13 main classes, with additional subcategories for several classes.

Discussion

Principal Findings

There is a need to integrate research into the design, development, and implementation of health technologies for improving their safety and reducing technology-induced errors [35]. The evolution of knowledge in this area has witnessed growth [35], but a classification suitable for EHR users' clinical practices is needed to derive maximum benefit from safety information reported through these means [19,28,44]. During this study, error types were adapted iteratively after several test rounds to develop a classification for reporting patient safety incidents in the clinical use of high-maturity EHRs. Some of the categories for error types have been identified in the scientific literature [13-20]; thus, their rationale exists. However, reliable classification work requires a solid knowledge of the features of an EHR system. In this study, an effective understanding of the content of problem reports was ensured by a multidisciplinary research team that included 3 physicians using the EHR system daily.

As the classification work progressed, one compromising agreement had to be made to continue classification development and validation with these particular data. According to the data, the medication section of the studied EHR system caused incidents for which it was not possible to detect a specific root cause. However, it was clear from the descriptions that the incidents were caused by features in the EHR system's medication section. As a result, a category was created for these incidents, but a deeper analysis in future research is needed to address the underlying problems with the medication section. Only some incidents related to the medication section were related to a lack of competence and classified accordingly. Finally, the manner in which the study was conducted was time consuming in terms of manual classification and review by the research team, but such a methodological approach was very profitable in practice. However, it is evident that the new emerging classification requires further validation in different health care contexts and with different high-maturity EHR products. Moreover, clinical users should test the classification so that its functionality and applicability can be assessed from the clinician's perspective in real patient care situations.

The number of EHR-related patient safety incidents during the implementation period was five-fold as compared with the preimplementation period, which can be viewed only as an indicative figure with respect to the actual situation. However, while analyzing possible reasons for increases in safety events, how members of a clinical team are organized and assigned, and how patient care is coordinated and delivered, is of paramount importance [32]. In this study, because of illustrative incident descriptions, a category, *general situation of endangering patient safety due to the introduction of an electronic health record*, was developed. On the basis of professionals' descriptions, the implementation of a new EHR system may disrupt the conventional ways of organizing and coordinating patient care; thus, it is justified to include the category to examine the wider implications of the implementation of the EHR system from the perspective of corrective actions [27,44].

Of the 427 classified patient safety incidents, usability problems accounted for 73 (17.1%) incidents, documentation problems for 60 (14.1%) incidents, medication section for 89 (20.8%) incidents, and clinical workflow problems for 33 (7.7%) incidents. Downtime problems were rare (8/427, 1.9%), and unlike in previous studies [15,43], unplanned downtime did not exist. Owing to decreases in unplanned situations, we assumed that the hospital competence for EHR implementation has developed with experience from previous implementations. However, despite the new EHR system being a high-maturity EHR system, further efforts are recommended to improve its usability, make the medication section more user friendly, and devote more attention to the needs and perspectives related to clinical workflow in the development of EHR systems [12,13,16-18,20,23,32]. In doing so, the EHR system provides even more benefits as a tool for clinicians to improve patient safety [22].

Limitations

The study had several limitations: causal attributions for HIT-related risks and safety incidents are difficult to identify, as they generally involve interactions between technical and nontechnical factors, which are notoriously difficult to separate [22]. The development of the classification was time consuming, and practical challenges were encountered in the application of the classification. The biggest obstacles arose from the readymade data, which included the professionals' own descriptions of the incident. Not all professionals described the incident's features in sufficient detail. Typically, this caused a situation in which the research team could not always definitively ascertain which category applies to an incident. To ensure the reliability of the results, 74 incidents were rejected when the research team members held detailed discussions after the blinded review. Therefore, it is important to ensure that the organization continues to pay attention to making sufficiently detailed descriptions to benefit from the reporting [27,40,44].

Moreover, it should be noted that the nature and well-known limitations of patient safety incident reporting should be

XSL•FO
RenderX

considered while interpreting the volume of incident data. Reports do not provide exact frequencies of incidents; consequently, data do not provide exact error rates, but rather a descriptive analysis of typical EHR-related safety problem types [26-28,44].

Conclusions

The broad spectrum of patient safety incidents is best understood by assessing data from multiple sources using a uniform classification, and this study proposes such a system for high-maturity EHR systems, which are known contributors to patient harm. However, this study's results indicate that the error types previously identified in the literature change and are specified with the development cycles of EHR maturity. Technology-induced errors in high-maturity EHRs include at least suboptimally developed workflows, usability design challenges, and interface and documentation problems. Unlike previous studies, there were no unplanned downtimes. Further research is recommended to evaluate the suitability of the classification for clinical use and its possible wider applicability in health care systems.

Acknowledgments

The authors acknowledge Finnish Governmental Research Funding TYH2019244 provided for their study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Classification and data analysis process for the iterative development of electronic health record patient safety error classification.
[PNG File , 446 KB - medinform_v9i8e30470_app1.png]

Multimedia Appendix 2
Classification of error types for electronic health record–related incidents by main and subcategories, which are identified with class numbering, for example, for the first class, the main category is "1" and the subcategories are "1.1" and "1.2." Table columns illustrate class identifiers, names, and respective class descriptions. The table also provides the number (N) of classified error reports per class category.
[DOCX File , 22 KB - medinform_v9i8e30470_app2.docx]

References

1. Kruse CS, Stein A, Thomas H, Kaur H. The use of electronic health records to support population health: a systematic review of the literature. J Med Syst 2018 Sep 29;42(11):214 [FREE Full text] [doi: 10.1007/s10916-018-1075-6] [Medline: 30269237]
2. Payne TH, Corley S, Cullen TA, Gandhi TK, Harrington L, Kuperman GJ, et al. Report of the AMIA EHR-2020 Task Force on the status and future direction of EHRs. J Am Med Inform Assoc 2015 Sep;22(5):1102-1110. [doi: 10.1093/jamia/ocv066] [Medline: 26024883]
3. Kruse CS, Beane A. Health information technology continues to show positive effect on medical outcomes: systematic review. J Med Internet Res 2018 Feb 05;20(2):e41 [FREE Full text] [doi: 10.2196/jmir.8793] [Medline: 29402759]
4. Carayon P, Wooldridge A, Hose B, Salwei M, Benneyan J. Challenges and opportunities for improving patient safety through human factors and systems engineering. Health Aff (Millwood) 2018 Nov;37(11):1862-1869 [FREE Full text] [doi: 10.1377/hlthaff.2018.0723] [Medline: 30395503]
5. Kruse CS, Kristof C, Jones B, Mitchell E, Martinez A. Barriers to electronic health record adoption: a systematic literature review. J Med Syst 2016 Dec;40(12):252 [FREE Full text] [doi: 10.1007/s10916-016-0628-9] [Medline: 27714560]
6. Colicchio TK, Cimino JJ, Del Fiol G. Unintended consequences of nationwide electronic health record adoption: challenges and opportunities in the post-meaningful use era. J Med Internet Res 2019 Jun 03;21(6):e13313 [FREE Full text] [doi: 10.2196/13313] [Medline: 31162125]
7. Kuziemsky CE, Randell R, Borycki EM. Understanding unintended consequences and health information technology: Contribution from the IMIA organizational and social issues working group. Yearb Med Inform 2016 Nov 10(1):53-60 [FREE Full text] [doi: 10.15265/IY-2016-027] [Medline: 27830231]
8. Vanderhook S, Abraham J. Unintended consequences of EHR systems: a narrative review. In: Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care. 2017 May 15 Presented at: Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care; 2017; New Orleans, Louisiana, USA p. 218-225. [doi: 10.1177/2327857917061048]
9. Coiera E, Ash J, Berg M. The unintended consequences of health information technology revisited. Yearb Med Inform 2018 Mar 06;25(01):163-169 [FREE Full text] [doi: 10.15265/iy-2016-014]

XSL•FO
RenderX

10. HIMSS Adoption Model for Analytics Maturity (AMAM). Healthcare Information and Management System Society Inc. URL: https://www.himssanalytics.org/amam [accessed 2021-04-17]
11. Hertzum M, Ellingsen G. The implementation of an electronic health record: Comparing preparations for Epic in Norway with experiences from the UK and Denmark. Int J Med Inform 2019 Sep;129:312-317. [doi: 10.1016/j.ijmedinf.2019.06.026] [Medline: 31445272]
12. Bornstein S. An integrated EHR at Northern California Kaiser Permanente: pitfalls, challenges, and benefits experienced in transitioning. Appl Clin Inform 2012;3(3):318-325 [FREE Full text] [doi: 10.4338/ACI-2012-03-RA-0006] [Medline: 23646079]
13. Ratwani RM, Savage E, Will A, Arnold R, Khairat S, Miller K, et al. A usability and safety analysis of electronic health records: a multi-center study. J Am Med Inform Assoc 2018 Sep 01;25(9):1197-1201. [doi: 10.1093/jamia/ocy088] [Medline: 29982549]
14. Palojoki S, Pajunen T, Lehtonen L, Saranto K. FIN-TIERA: A tool for assessing technology induced errors. Methods Inf Med 2017 Jan 09;56(1):1-12. [doi: 10.3414/ME16-01-0097] [Medline: 27922661]
15. Palojoki S, Pajunen T, Saranto K, Lehtonen L. Electronic health record-related safety concerns: a cross-sectional survey of electronic health record users. JMIR Med Inform 2016 May 06;4(2):e13 [FREE Full text] [doi: 10.2196/medinform.5238] [Medline: 27154599]
16. Senathirajah Y, Kaufman DR, Cato KD, Borycki EM, Fawcett JA, Kushniruk AW. Characterizing and visualizing display and task fragmentation in the electronic health record: mixed methods design. JMIR Hum Factors 2020 Oct 21;7(4):e18484 [FREE Full text] [doi: 10.2196/18484] [Medline: 33084580]
17. Fujita K, Onishi K, Takemura T, Kuroda T. The improvement of the electronic health record user experience by screen design principles. J Med Syst 2019 Dec 10;44(1):21. [doi: 10.1007/s10916-019-1505-0] [Medline: 31823092]
18. Rayner J, Khan T, Chan C, Wu C. Illustrating the patient journey through the care continuum: Leveraging structured primary care electronic medical record (EMR) data in Ontario, Canada using chronic obstructive pulmonary disease as a case study. Int J Med Inform 2020 Aug;140:104159. [doi: 10.1016/j.ijmedinf.2020.104159] [Medline: 32473567]
19. Kim MO, Coiera E, Magrabi F. Problems with health information technology and their effects on care delivery and patient outcomes: a systematic review. J Am Med Inform Assoc 2017 Dec 01;24(2):246-250. [doi: 10.1093/jamia/ocw154] [Medline: 28011595]
20. Howe JL, Adams KT, Hettinger AZ, Ratwani RM. Electronic health record usability issues and potential contribution to patient harm. J Am Med Assoc 2018 Mar 27;319(12):1276-1278 [FREE Full text] [doi: 10.1001/jama.2018.1171] [Medline: 29584833]
21. Sittig DF, Singh H. Electronic health records and national patient-safety goals. N Engl J Med 2012 Nov 8;367(19):1854-1860 [FREE Full text] [doi: 10.1056/NEJMsb1205420] [Medline: 23134389]
22. Singh H, Sittig DF. Measuring and improving patient safety through health information technology: The Health IT Safety Framework. BMJ Qual Saf 2015 Sep 14:226-232 [FREE Full text] [doi: 10.1136/bmjqs-2015-004486] [Medline: 26369894]
23. Nolan M, Siwani R, Helmi H, Pickering B, Moreno-Franco P, Herasevich V. Health IT Usability Focus Section: Data use and navigation patterns among medical ICU clinicians during electronic chart review. Appl Clin Inform 2017 Dec 14;08(04):1117-1126. [doi: 10.4338/aci-2017-06-ra-0110]
24. Tsou A, Lehmann C, Michel J, Solomon R, Possanza L, Gandhi T. Safe practices for copy and paste in the EHR. Appl Clin Inform 2017 Dec 20;26(01):12-34. [doi: 10.4338/aci-2016-09-r-0150]
25. Patel VL, Kannampallil TG, Shortliffe EH. Role of cognition in generating and mitigating clinical errors. BMJ Qual Saf 2015 Jul;24(7):468-474. [doi: 10.1136/bmjqs-2014-003482] [Medline: 25935928]
26. Mitchell I, Schuster A, Smith K, Pronovost P, Wu A. Patient safety incident reporting: a qualitative study of thoughts and perceptions of experts 15 years after 'To Err is Human'. BMJ Qual Saf 2016 Feb;25(2):92-99. [doi: 10.1136/bmjqs-2015-004405] [Medline: 26217037]
27. Howell A, Burns EM, Hull L, Mayer E, Sevdalis N, Darzi A. International recommendations for national patient safety incident reporting systems: an expert Delphi consensus-building process. BMJ Qual Saf 2017 Feb;26(2):150-163. [doi: 10.1136/bmjqs-2015-004456] [Medline: 26902254]
28. Palojoki S, Vuokko R, Vakkuri A, Saranto K. Electronic health record system-related patient safety incidents - how to classify them? Stud Health Technol Inform 2020 Nov 23;275:157-161. [doi: 10.3233/SHTI200714] [Medline: 33227760]
29. Wyatt KD, Benning TJ, Morgenthaler TI, Arteaga GM. Development of a taxonomy for medication-related patient safety events related to health information technology in pediatrics. Appl Clin Inform 2020 Oct;11(5):714-724. [doi: 10.1055/s-0040-1717084] [Medline: 33113568]
30. Health IT and patient safety: building safer systems for better care. In: Committee on Patient Safety and Health Information Technology; Institute of Medicine. Washington (DC), US: National Academies Press; Nov 10, 2011. URL: https://www.ncbi.nlm.nih.gov/books/NBK189661/
31. Liang C, Zhou S, Yao B, Hood D, Gong Y. Corrigendum to "Toward systems-centered analysis of patient safety events: Improving root cause analysis by optimized incident classification and information presentation" [Int. J. Med. Inform. 135 (2020) 104054]. Int J Med Inform 2020 May;137:104103. [doi: 10.1016/j.ijmedinf.2020.104103] [Medline: 32113970]

XSL•FO
RenderX

32. Tutty MA, Carlasare LE, Lloyd S, Sinsky CA. The complex case of EHRs: examining the factors impacting the EHR user experience. J Am Med Inform Assoc 2019 Jul 01;26(7):673-677 [FREE Full text] [doi: 10.1093/jamia/ocz021] [Medline: 30938754]
33. Blijleven V, Koelemeijer K, Jaspers M. Exploring workarounds related to electronic health record system usage: a study protocol. JMIR Res Protoc 2017 Apr 28;6(4):e72 [FREE Full text] [doi: 10.2196/resprot.6766] [Medline: 28455273]
34. Cifuentes M, Davis M, Fernald D, Gunn R, Dickinson P, Cohen DJ. Electronic health record challenges, workarounds, and solutions observed in practices integrating behavioral health and primary care. J Am Board Fam Med 2015 Oct;28 Suppl 1:63-72 [FREE Full text] [doi: 10.3122/jabfm.2015.S1.150133] [Medline: 26359473]
35. Borycki E. Quality and Safety in eHealth: The need to build the evidence base. J Med Internet Res 2019 Dec 19;21(12):e16689 [FREE Full text] [doi: 10.2196/16689] [Medline: 31855183]
36. Bowker GC, Star SL. Sorting Things Out: Classification and Its Consequences. Cambridge, MA, US: MIT Press; 1999.
37. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems. I: Terminology and typology. Methods Inf Med 2000 Mar;39(1):16-21. [Medline: 10786065]
38. Zeng F, Sun X, Yang B, Shen H, Liu L. The theoretical construction of a classification of clinical somatic symptoms in psychosomatic medicine theory. PLoS One 2016 Aug 15;11(8):1-10 [FREE Full text] [doi: 10.1371/journal.pone.0161222] [Medline: 27525701]
39. Oxford Dictionary of English. Oxford, UK: Oxford University Press; 2010.
40. Borycki EM, Kushniruk AW. Towards a framework for managing risk associated with technology-induced error. Stud Health Technol Inform 2017;234:42-48. [Medline: 28186013]
41. Kruse CS, DeShazo J, Kim F, Fulton L. Factors associated with adoption of health information technology: a conceptual model based on a systematic review. JMIR Med Inform 2014;2(1):e9 [FREE Full text] [doi: 10.2196/medinform.3106] [Medline: 25599673]
42. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A survey of recent advances in deep learning techniques for Electronic Health Record (EHR) analysis. IEEE J Biomed Health Inform 2018 Dec;22(5):1589-1604. [doi: 10.1109/JBHI.2017.2767063] [Medline: 29989977]
43. Palojoki S, Saranto K, Lehtonen L. Reporting medical device safety incidents to regulatory authorities: an analysis and classification of technology-induced errors. Health Informatics J 2019 Sep;25(3):731-740 [FREE Full text] [doi: 10.1177/1460458217720400] [Medline: 28747134]
44. Palojoki S, Mäkelä M, Lehtonen L, Saranto K. An analysis of electronic health record-related patient safety incidents. Health Informatics J 2017 Jun;23(2):134-145. [doi: 10.1177/1460458216631072] [Medline: 26951568]

Abbreviations

EHR: electronic health record
HIT: health information technology

Edited by C Lovis; submitted 16.05.21; peer-reviewed by J Walsh, F Magrabi; comments to author 05.06.21; revised version received 10.06.21; accepted 10.07.21; published 31.08.21.

Please cite as:
Palojoki S, Saranto K, Reponen E, Skants N, Vakkuri A, Vuokko R
Classification of Electronic Health Record–Related Patient Safety Incidents: Development and Validation Study
JMIR Med Inform 2021;9(8):e30470
URL: https://medinform.jmir.org/2021/8/e30470
doi:10.2196/30470
PMID:34245558

©Sari Palojoki, Kaija Saranto, Elina Reponen, Noora Skants, Anne Vakkuri, Riikka Vuokko. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 31.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Gender Prediction for a Multiethnic Population via Deep Learning Across Different Retinal Fundus Photograph Fields: Retrospective Cross-sectional Study

Bjorn Kaijun Betzler[1*], MBBS; Henrik Hee Seung Yang[2*], MD; Sahil Thakur[3], MS; Marco Yu[3], PhD; Ten Cheer Quek[3], BEng; Zhi Da Soh[3], MPH; Geunyoung Lee[4], MSc; Yih-Chung Tham[2,3], PhD; Tien Yin Wong[2,3], MD, PhD; Tyler Hyungtaek Rim[2,3*], MD, MBA; Ching-Yu Cheng[1,2,3], MD, PhD

[1]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[2]Ophthalmology and Visual Science Academic Clinical Program, Duke-NUS Medical School, Singapore, Singapore

[3]Singapore Eye Research Institute, Singapore, Singapore

[4]Medi Whale Inc, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Tyler Hyungtaek Rim, MD, MBA
Ophthalmology and Visual Science Academic Clinical Program
Duke-NUS Medical School
8 College Rd
Singapore, 169857
Singapore
Phone: 65 65767228
Fax: 65 62252568
Email: tyler.rim@snec.com.sg

Abstract

Background: Deep learning algorithms have been built for the detection of systemic and eye diseases based on fundus photographs. The retina possesses features that can be affected by gender differences, and the extent to which these features are captured via photography differs depending on the retinal image field.

Objective: We aimed to compare deep learning algorithms' performance in predicting gender based on different fields of fundus photographs (optic disc–centered, macula-centered, and peripheral fields).

Methods: This retrospective cross-sectional study included 172,170 fundus photographs of 9956 adults aged ≥40 years from the Singapore Epidemiology of Eye Diseases Study. Optic disc–centered, macula-centered, and peripheral field fundus images were included in this study as input data for a deep learning model for gender prediction. Performance was estimated at the individual level and image level. Receiver operating characteristic curves for binary classification were calculated.

Results: The deep learning algorithms predicted gender with an area under the receiver operating characteristic curve (AUC) of 0.94 at the individual level and an AUC of 0.87 at the image level. Across the three image field types, the best performance was seen when using optic disc–centered field images (younger subgroups: AUC=0.91; older subgroups: AUC=0.86), and algorithms that used peripheral field images had the lowest performance (younger subgroups: AUC=0.85; older subgroups: AUC=0.76). Across the three ethnic subgroups, algorithm performance was lowest in the Indian subgroup (AUC=0.88) compared to that in the Malay (AUC=0.91) and Chinese (AUC=0.91) subgroups when the algorithms were tested on optic disc–centered images. Algorithms' performance in gender prediction at the image level was better in younger subgroups (aged <65 years; AUC=0.89) than in older subgroups (aged ≥65 years; AUC=0.82).

Conclusions: We confirmed that gender among the Asian population can be predicted with fundus photographs by using deep learning, and our algorithms' performance in terms of gender prediction differed according to the field of fundus photographs, age subgroups, and ethnic groups. Our work provides a further understanding of using deep learning models for the prediction of gender-related diseases. Further validation of our findings is still needed.

(JMIR Med Inform 2021;9(8):e25165) doi:10.2196/25165

XSL•FO
RenderX

KEYWORDS

deep learning; artificial intelligence; retina; gender; ophthalmology

Introduction

An individual's gender is associated with a variety of systemic and ocular diseases. Females have longer life expectancies compared to those of males, regardless of their educational, economic, political, and health statuses [1,2]. Decreased estrogen production predisposes postmenopausal women to degenerative conditions, including cataracts and age-related macular degeneration [3-8]. In contrast, males are predisposed to open-angle glaucoma [9], diabetic retinopathy [10], and pigment dispersion glaucoma [11].

Deep learning algorithms have been developed for the detection of systemic and eye diseases based on fundus photographs [12-21]. By using deep neural networks, Poplin et al [12] found that cardiovascular risk factors, including gender, can be predicted with fundus images and obtained good classification results with a data set comprising White individuals. More recently, Gerrits et al [17] and Kim et al [22] also predicted gender by using neural networks to analyze Qatari and South Korean data sets, respectively.

This study builds on preexisting literature in three ways. First, we predicted gender by using retinal fundus images from a Southeast Asian data set. Second, we evaluated how differing fundus photography fields could have an effect, if any, on gender classification results. This is worth exploring because the retina possesses features that can be affected by gender differences (eg, vessel structure; optic nerve, fovea, and macular morphology; and retinal pigmentation). Different fundus photography fields (optic disc–centered, macula-centered, and peripheral fields) capture these features to varying extents and affect these features' availability in a neural network. Rim et al [22] reported the good generalizability of similar deep learning algorithms that have been used to predict gender based on fundus photographs; however, intracohort subgroup comparisons were not performed. Understanding how model performance differs based on different ethnic, age, and image field subgroups will be useful [22].

Third, the diversity of our data set allowed for the comparison of algorithm performance across age and ethnic subgroups (Malay, Chinese, and Indian subgroups). The introduction of artificial intelligence in clinical medicine has brought about ethical concerns, of which one is problematic decision-making by algorithms that reflect biases that are inherent in the data used to train these algorithms [23]. Ensuring that our model generalizes well across different ethnicities is essential for avoiding inadvertent, subtle discrimination in health care delivery [24]. Cross-cultural analysis is a unique feature of our study—one that is lacking in existing literature on deep learning in ophthalmology because few populations are inherently diverse.

Methods

Ethics Statement

This retrospective cross-sectional study was approved by the institutional ethical committee and adhered to the tenets of the Declaration of Helsinki. The need to obtain written informed consent was waived due to the use of anonymized and deidentified data.

Study Population

The Singapore Epidemiology of Eye Diseases (SEED) study is a population-based study that recruited subjects from the three major ethnic groups (the Chinese, Malay, and Indian ethnic groups) in Singapore. The SEED study's baseline examinations were conducted from 2004 through 2011, and subsequent follow-up studies were performed, as follows: the Singapore Malay Eye Study (baseline examination: 2004-2006; follow-up examination: 2010-2013), the Singapore Indian Eye Study (baseline examination: 2007-2009; follow-up examination: 2013-2016), and the Singapore Chinese Eye Study (baseline examination: 2009-2011; follow-up examination: 2016-2018). The detailed methodology of the SEED study was published previously [25-28]. Briefly, an age-stratified random sampling method was used to select subjects aged ≥40 years from each ethnic group living across southwestern Singapore. In total, 3280 out of 4168 Malay individuals (78.7%), 3400 out of 4497 Indian individuals (75.6%), and 3353 out of 4606 Chinese individuals (72.8%) agreed to participate in the study. As such, an overall response rate of 75.6% was achieved. The entire data set, which included both visits, was split and used for algorithm development and testing.

Fundus Photography and Image Database

A digital, nonmydriatic retinal camera (Canon CR-1 Mark-II nonmydriatic, digital retinal camera; Canon Inc) was used to obtain fundus photographs according to Early Treatment for Diabetic Retinopathy Study (ETDRS) standard fields 1 to 5. This was done after performing pharmacological dilation with 1% tropicamide and 2.5% phenylephrine hydrochloride. A total of 175,038 fundus photographs from 10,033 SEED study participants were included in this study. Original fundus photographs (3504×2336 pixels) were extracted in the JPEG format, and the black space around the contours of each photograph was removed. All images were reformatted to 300×300-pixel images.

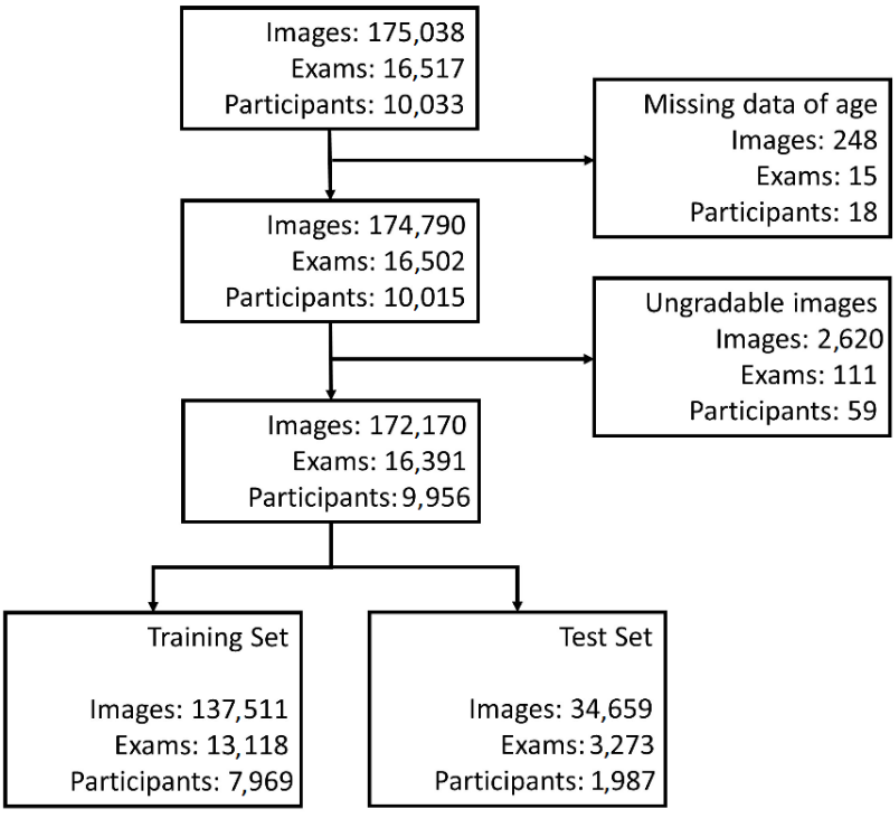
Model Development

Separate models for 3 different focus fields of fundus photographs were developed (optic disc–centered, macula-centered, peripheral fields) [29]. Images without age and gender information or those deemed ungradable were excluded from the analysis. The gradeability of fundus photographs was manually determined based on a modification of the Wisconsin Age-Related Maculopathy Grading System [30]. A total of 172,170 fundus photographs (from the 16,391 examinations of 9956 participants) were divided into a training

XSL•FO
RenderX

set (137,511/172,170, 79.9%) for developing our models and a test set (34,659/172,170, 20.1%), which was reserved to evaluate model performance. The photographs were stratified according to age groups, gender, and ethnic groups. Figure 1 and Table 1 describe this split in more detail. The test set was not used during model development. This division of photographs was based on the individual level rather than the image level to avoid class imbalances. Dividing photographs at the individual level ensured that there was an equal number of images for each individual, thereby avoiding the potential skew of data. Data augmentation (random rotation from −5 to 5 degrees and random brightness adjustment) was performed to introduce invariance in our neural network [31,32].

Figure 1. Flowchart depicting the inclusion and exclusion of study images and participants.



XSL•FO
RenderX

Table 1. Population characteristics.

Characteristics	Training set, n (%)	Test set, n (%)	*P* value
Fundus photographs (N=175,038)[a]			
Optic disc–centered photographs	56,814 (41.3)	14,231 (41.1)	.45
Macula-centered photographs	53,863 (39.2)	13,705 (39.5)	N/A[b]
Other peripheral photographs	26,834 (19.5)	6723 (19.4)	N/A
Examinations (N=16,517)[c]			
Age group (years)			.75
40-49	2145 (16.4)	551 (16.8)	
50-59	4447 (33.9)	1105 (33.8)	
60-69	3772 (28.8)	915 (28)	
≥70	2754 (21)	702 (21.5)	
Gender			>.99
Female	6725 (51.3)	1678 (51.3)	
Male	6393 (48.7)	1595 (48.7)	
Ethnic groups			.80
Malay	4067 (31)	1024 (31.3)	
Chinese	4609 (35.1)	1161 (35.5)	
Indian	4442 (33.9)	1088 (33.2)	

[a]The training set included a total of 137,511 fundus photographs, and the test set included a total of 34,659 fundus photographs.

[b]N/A: not applicable.

[c]The training set included data on a total of 13,118 examinations, and the test set included data on a total of 3273 examinations.

Our deep learning model, which was based on the Visual Geometry Group-16 neural network architecture [33], was developed, trained, and evaluated in TensorFlow [34,35]. The model had 13 convolutional layers after batch normalization and a fully connected layer after compressing the feature vector via global average pooling. The Adam optimizer with fixed weight decay was used to train our model; the learning rate was set to 0.0001 for 100 epochs. At the end of the neural network, a prediction score was generated for binary classification. A low prediction score was classified as “male,” while a high prediction score was classified as “female.” With regard to model explanation, saliency maps created via guided gradient-weighted class activation mapping (Grad-CAM) [36,37] were superimposed over input images to facilitate our understanding of how our model predicted gender.

Reference Standard

Gender information (male or female) was collected from the SEED study participants’ National Registration Identity Card, which is provided to all Singapore citizens.

Subgroups

Age was calculated based on the birth date indicated on participants’ National Registration Identity Card. The younger subgroup included participants aged 40 to 65 years, while the older subgroup included those aged ≥65 years. To classify the three ethnic subgroups, our study used criteria that were set by the Singapore census to define *Malay*, *Chinese*, and *Indian* [25,27].

Statistical Analysis

Python packages, including NumPy, SciPy, matplotlib, scikit-learn, were used to process the data [38]. Performance was evaluated by using the internal validation set, which included 34,659 fundus photographs (14,231 optic disc–centered field images, 13,705 macula-centered field images, and 6723 peripheral field images). Receiver operating characteristic curves for binary classification were plotted. The DeLong test for area under the receiver operating curve (AUC) comparisons was used [39]. Individual-based and image-based analyses were conducted.

Results

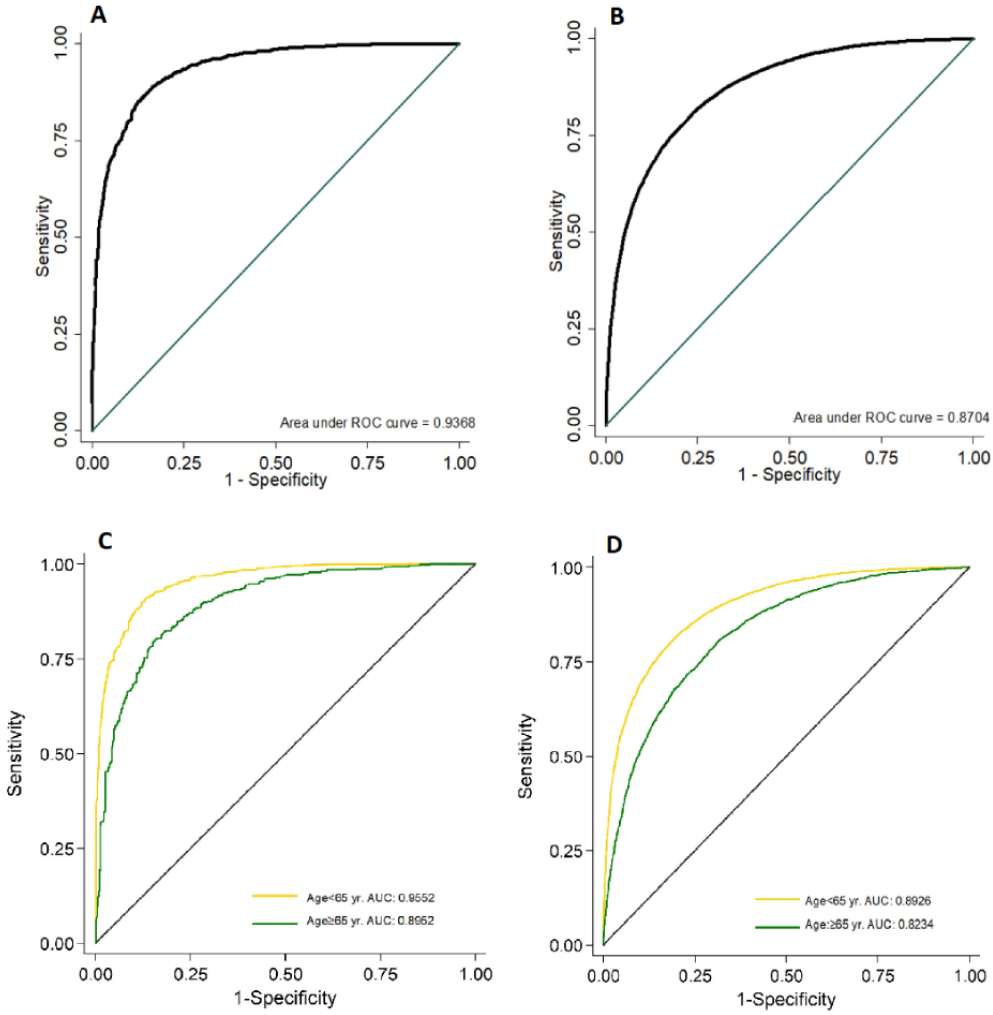
A total of 172,170 fundus photographs, including 71,045 optic disc–centered field images, 67,568 macula-centered field images, and 33,557 peripheral field images, were distributed among the training and test sets (Table 1). The mean age of participants was 60.8 years (SD 10.3 years; minimum: 40.0 years; maximum: 91.3 years), and 48.7% (7988/16,391) of the participants were male. The distribution of photographs between the training and test sets was stratified according to gender, age subgroups, and the three ethnic subgroups.

Upon validation, the model achieved an AUC of 0.94 (95% CI 0.93-0.95) at the individual level and an AUC of 0.87 (95% CI 0.87-0.87) at the image level (Figure 2). With regard to the age subgroup analysis at the individual level, model performance was better in the younger group (aged 40-65 years; AUC=0.96;

XSL•FO
RenderX

95% CI 0.95-0.96) than in the older group (aged >65 years; AUC=0.90; 95% CI 0.88-0.91; *P*<.001). At the image level, model performance in the younger group also surpassed model performance in the older group; AUCs of 0.89 (95% CI 0.89-0.90) and 0.82 (95% CI 0.82-0.83), respectively, were achieved (*P*<.001).

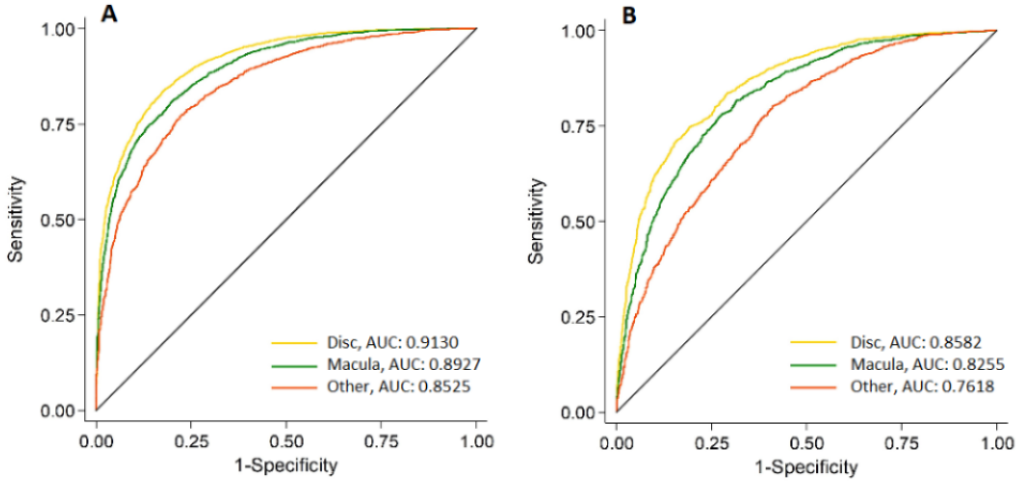
Figure 2. ROC curves at the individual and image levels based on the internal test set. A: Individual level; total population. B: Image level; total images. C: Individual level; age subgroups. D: Image level; age subgroups. Upon internal testing, the AUCs achieved were 0.937 and 0.870 at the individual and image levels (A and B), respectively. The AUCs achieved in the younger subgroups (aged <65 years) were 0.955 and 0.893 at the individual and image levels, respectively (*P*<.001). The AUCs achieved for the older subgroups were 0.895 and 0.823 at the individual and images levels, respectively (*P*<.001). AUC: area under the receiver operating curve; ROC: receiver operating curve.



We examined the differences in the model's predictions of gender across the three fundus photography fields at the image level. Figure 3 describes the corresponding AUC curves. The model's overall performance was better in the younger group (Figure 3) than in the older group (Figure 3). In both age groups, optic disc–centered images resulted in the best performance in terms of gender prediction. In the younger age group, the AUC was 0.91 (95% CI 0.91-0.92) for the optic disc–centered images, 0.89 (95% CI 0.89-0.90) for the macula-centered images, and 0.85 (95% CI 0.84-0.86) for the peripheral field images (*P*<.001). In the older age group, the AUC was 0.86 (95% CI 0.85-0.87) for the optic disc–centered images, 0.83 (95% CI 0.81-0.84) for the macula-centered images, and 0.76 (95% CI 0.84-0.86) for the peripheral field images (*P*<.001).

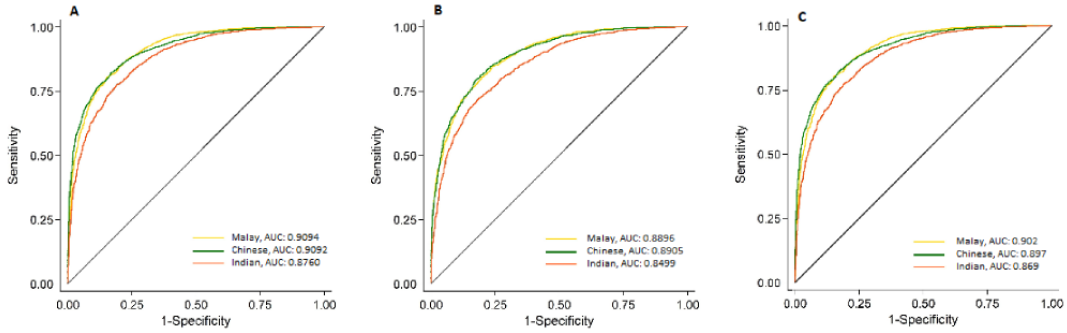
XSL•FO
RenderX

Figure 3. Comparison of the algorithms' performance in gender prediction between the different fundus photograph fields (optic disc–centered, macula-centered, and peripheral or other fields) A: Age<65 years. B: Age≥65 years. AUC: area under the receiver operating curve.



We also evaluated the model's gender prediction performance according to ethnic groups (the Malay, Chinese, Indian groups). Figure 4 depicts our algorithms' performance in analyzing photographs at the image level; the model fared relatively well for the Malay and Chinese ethnic groups but fared suboptimally for the Indian ethnic group. The model's overall performance was better when using optic disc–centered images (Figure 4) than when using macula-centered images (Figure 4). With regard to the optic disc–centered image group, the AUC was 0.91 (95% CI 0.90-0.92) for the Malay group, 0.91 (95% CI 0.90-0.92) for the Chinese group, and 0.88 (95% CI 0.87-0.89) for the Indian group (P<.001). With regard to the macular-centered image group, the AUC was 0.890 (95% CI 0.88-0.90) for the Malay group, 0.89 (95% CI 0.88-0.90) for the Chinese group, and 0.85 (95% CI 0.84-0.86) for the Indian group (P<.001). No significant performance differences were observed between the Malay and Chinese ethnic groups (optic disc–centered images: P=.98; macula-centered images: P=.90). Precision-recall curves were generated in addition to the receiver operating curves. These are provided in Multimedia Appendix 1.

Figure 4. Comparison of the algorithms' performance in gender prediction between ethnic groups. A: Optic disc–centered photographs. B: Macula-centered photographs. C: Overall. AUC: area under the receiver operating curve.



Saliency maps (heat maps) were generated via Grad-CAM for model explanation. Fundus photographs and overlaid heat maps that were strongly associated with males and females (extreme binary classification prediction scores) are shown in Figure 5 and Figure 6, respectively. The optic disc and the surrounding structures are activated in every heat map in Figure 5 and Figure 6. Selected heat maps of fundus images showing pathological lesions are presented in Figure 7. These heat maps suggested that the optic disc was an area of interest in gender prediction, despite the presence of random distractive elements (laser scars, diabetic retinopathy, hypertensive retinopathy, and age-related macular degeneration). A similar trend was noted in the heat maps of macula-centered images.

XSL•FO
RenderX

Figure 5. Original fundus photographs (A) and overlaid heat maps (B) with the features that were most associated with the male gender.

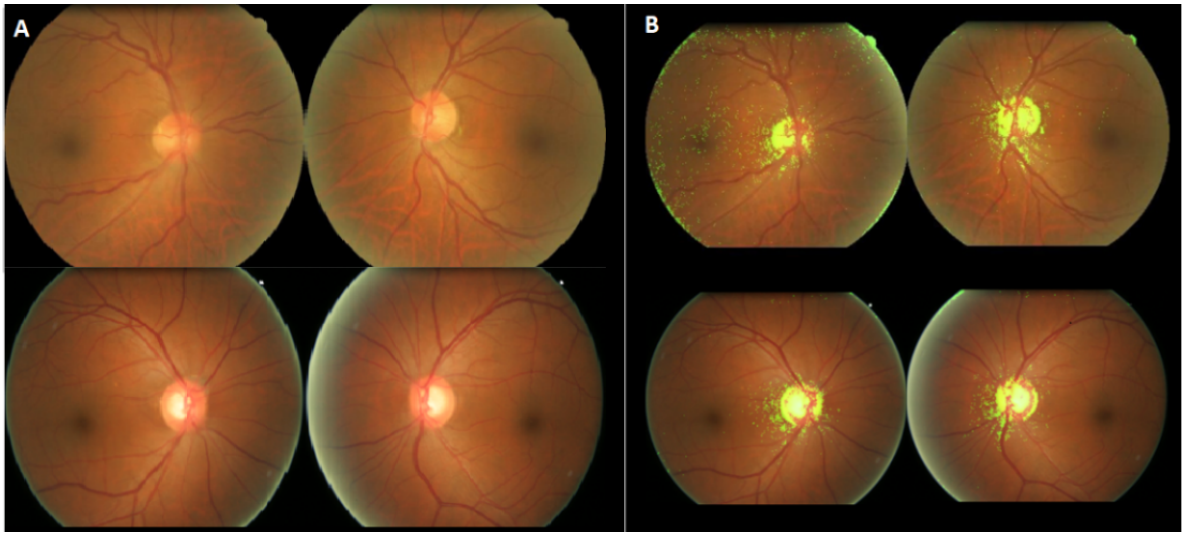


Figure 6. Original fundus photographs (A) and overlaid heat maps (B) with the features that were most associated with the female gender.

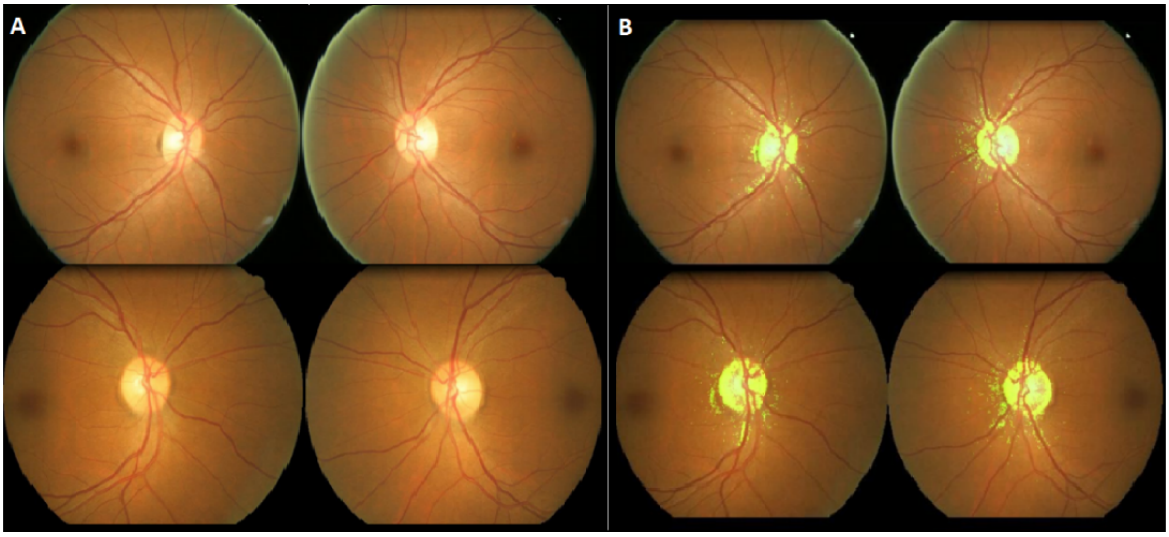
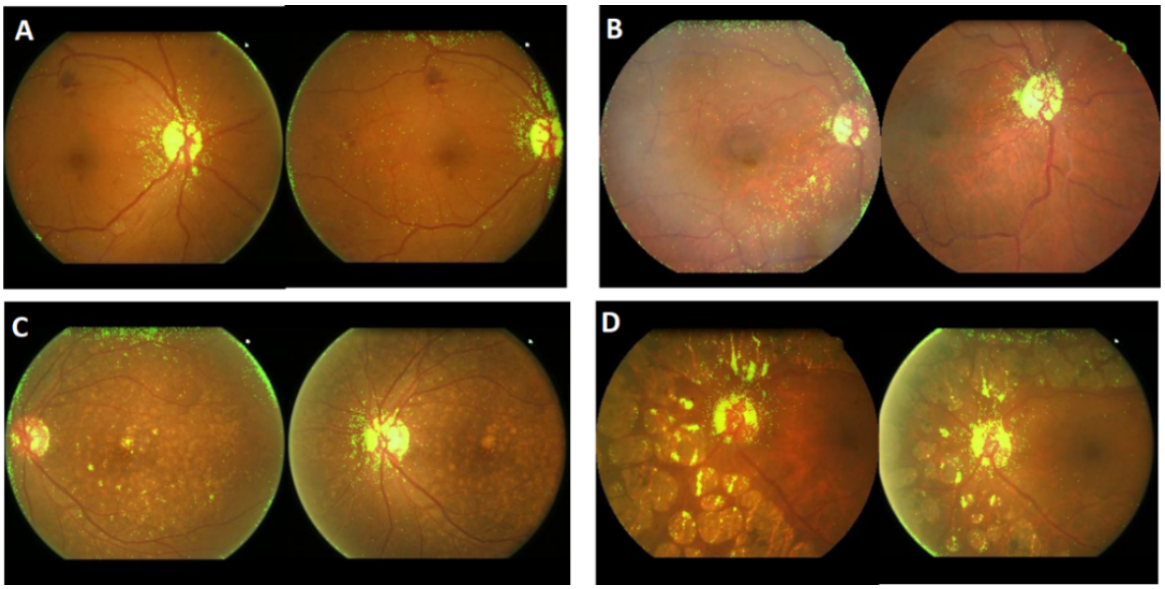


Figure 7. Selected heat maps of fundus images showing pathological lesions (all images are optic disc–centered images). A: Images of diabetic retinopathy. B: Images of hypertensive retinopathy. C: Images of age-related macular degeneration. D: Images of laser scars.



XSL•FO
RenderX

Discussion

Principal Findings

In this study, our results demonstrated the following points: (1) model performance was better in the younger subgroup (aged 40-65 years) than in the older subgroup (aged >65 years); (2) optic disc–centered images provided the most accurate predictions for gender, followed by macula-centered images; (3) the model's performance was better in the Malay and Chinese ethnic subgroups than in the Indian ethnic subgroup; and (4) the algorithms functioned well in the presence of possibly distractive attributes.

The deep learning algorithm from Poplin and colleagues [12] was developed based on 48,101 and 236,234 color fundus photographs from the UK Biobank and Eye Picture Archive Communication System (EyePACS) data sets, respectively. It successfully predicted gender and achieved an AUC of 0.97 (95% CI 0.97-0.97) and 0.97 (95% CI 0.96-0.98) with the UK Biobank and the EyePACS-2K validation sets, respectively. Compared to the model developed by Poplin and colleagues [12], our model, which achieved an AUC of 0.94 (95% CI 0.93-0.95), is slightly less precise. However, our model was trained on and validated with a wider range of age groups than those of Poplin et al [12], and this could explain the relatively weaker performance of our algorithm; we confirmed that the algorithms' performance was lower in older subgroups.

The ability of neural networks to use greater abstractions and tighter integrations comes at the cost of lower interpretability [40]. Saliency maps, which are also called *heat maps* or *attention maps*, are common model explanation tools that are used to visualize model thinking by indicating areas of local morphological changes within fundus photographs that carry more weight in modifying network predictions. After using saliency maps, which were created via Grad-CAM [36,37], we believe that our algorithms mainly used the features of the optic disc for gender prediction. This pattern is consistent with the observations made by Poplin et al [12] in 2018. Deep learning models that were trained by using images from the UK Biobank and EyePACS data sets primarily highlighted the optic disc, retinal vessels, and macula when soft attention heat maps were applied, although there appeared to be a weak signal distributed throughout the retina [12]. Given that the Poplin et al [12] study predominantly used data sets of White (UK Biobank) and Hispanic (EyePACS) individuals and our study used a Southeast Asian population (ie, Malay, Chinese, and Indian individuals), our results suggest that gender predictions based on fundus photographs will likely generalize well across different ethnic groups. Additional validations of our models based on other global population data sets would strengthen these findings.

Figure 4 shows representative fundus photographs with the most masculine and feminine features. The heat maps mainly highlighted the optic discs and the surrounding areas. Our algorithms work well even when there are obvious different clinical characteristics, such as retinal hemorrhages, ghost vessels, laser scars, and silicone oil tamponade eye. To further confirm that the optic disc is an area of interest in gender prediction, we performed an explorative analysis on a subset of fundus images that did not capture the optic disc. Of the 6723 peripheral field images from the test set, 649 images had fields that did not encompass the optic disc. The model validation analysis based on these 649 peripheral field images that did not capture the optic disc returned an AUC of 0.69 (95% CI 0.65-0.73). This explorative comparison found that the model's performance markedly decreased in the absence of features provided by the optic disc. We can therefore suggest with greater certainty that the optic disc is the main structure used by deep learning algorithms for gender prediction. Kim et al [22] explored this concept in a slightly different manner. They reported a decreased AUC when predicting gender by using subsets of artificially inpainted fundus images, in which either the fovea or retinal vessels were erased. Optic disc omission was not described, although their reported heat maps indicated activations in the fovea, optic disc, and retinal vessels [22]. In addition, Korot et al [41] reported poor performance when using images with foveal pathologies and used this finding to suggest that the fovea is an important input region for gender prediction. However, their saliency maps strongly attributed their model's predictive power to the optic disc. This is similar to the findings of our study. It is likely that both the fovea and optic disc provide critical feature inputs for gender prediction models, but we are unable to comment on their relative importance.

The consideration of clinical applicability is essential when developing a useful deep learning algorithm. In a real-world setting, clinicians often encounter a mixture of fundus photographs with different fields, and it is common to observe the incorrect sorting of fundus photographs within publicly available data sets [42]. Our results showed that the most precise predictions were obtained when using optic disc–centered images as the model input in both the primary and subgroup analyses. Researchers should be aware of the possible performance differences that arise due to using different image fields when predicting gender or gender-related systemic factors; using optic disc–centered images alone or a combination of macula-centered and optic disc–centered images may be the most prudent approach. Based on our model's suboptimal performance when using peripheral field images, such images are not ideal input data for gender prediction models.

A common ethical concern with regard to decision-making by algorithms is that biases that are inherent in the data used to train these algorithms will manifest during usage [23]. A study of facial recognition software evaluated the performance of three leading recognition systems (those of Microsoft Corporation, IBM Corporation, and Megvii) in a gender classification task based on human skin tones [43]. The results showed that darker-skinned females were the most misclassified group. The study reported error rates of up to 34.7% for this group. However, a maximum error rate of 0.8% was achieved for lighter-skinned males. The implications of this study raised broad questions about the fairness and accountability of artificial intelligence and contributed to the concept of algorithmic accountability [44]. Based on the ethnic subgroup analysis in our study, our model did not perform as well in predicting gender in the Indian ethnic group (AUC=0.88; 95% CI 0.87-0.89) as it did in predicting gender in the Chinese (AUC=0.91; 95% CI 0.90-0.92) and Malay (AUC=0.91; 95%

XSL•FO
RenderX

CI 0.90-0.92) ethnic groups (P<.001). Given that our results have shown an undesired disparity in performance among the three ethnic groups, efforts will be needed to refine the model so that gender prediction accuracies across different ethnic groups are reasonably on par. Ensuring that our model generalizes well across different ethnicities is essential for avoiding inadvertent, subtle discrimination in health care delivery [24].

A study limitation is that our model was developed and trained with data from a single center; therefore, the model was exposed to the inadvertent incorporation of systemic error. Ideally, an external validation data set that includes photographs that were taken by using the ETDRS standard fields should also be used to evaluate the algorithms. However, photographs that include only 1 field (eg, only macula-centered photographs) cannot be used alone for comparisons because of the systemic error involved. We were unable to find a well-organized data set that included images with different fundus photography fields for external validation. Training the model by using diverse, independent data sets that are captured by using different instruments and come from a variety of populations and clinical settings will also enhance the model's generalizability [45]. Another limitation is our algorithms' limited applicability to younger populations, as our study only included images from individuals aged ≥40 years.

Conclusions

In summary, our study is, to the best of our knowledge, the first to predict gender based on retinal fundus photographs of a Southeast Asian population. The ethnic diversity of our data set allowed us to make intercultural comparisons. The model's performance was better in the Malay and Chinese subgroups than in the Indian ethnic subgroup, and more work is required to refine the model and avoid an undesired disparity in performance among different ethnic groups. Our analysis of 3 different retinal fields provides evidence that the optic disc is a critical feature that is used by deep learning models for gender prediction. Algorithms that used peripheral field images had the lowest performance, followed by those that used macula-centered photographs. Algorithms that used optic disc–centered photographs had the best performance. Our work provides a further understanding of using deep learning models for the prediction of gender-related diseases, and we recommend using external validation sets to replicate our results.

Conflicts of Interest

THR was a scientific adviser to Medi Whale Inc. and received stocks as a part of the standard compensation package. THR also holds patents on a deep-learning system in ophthalmology, which are not directly related to this study. GL is an employee and owns stocks in Medi Whale Inc. TYW is an inventor of a patent on the various deep learning systems in ophthalmology. All other authors declare no competing interests.

Multimedia Appendix 1
Precision-recall curves.
[PNG File , 180 KB - medinform_v9i8e25165_app1.png]

References

1. Austad SN. Why women live longer than men: sex differences in longevity. Gend Med 2006 Jun;3(2):79-92. [doi: 10.1016/s1550-8579(06)80198-1] [Medline: 16860268]
2. Møller AP, Fincher C, Thornhill R. Why men have shorter lives than women: effects of resource availability, infectious disease, and senescence. Am J Hum Biol 2009;21(3):357-364. [doi: 10.1002/ajhb.20879] [Medline: 19189415]
3. Klein BE, Klein R, Linton KL. Prevalence of age-related lens opacities in a population. The Beaver Dam Eye Study. Ophthalmology 1992 Apr;99(4):546-552. [doi: 10.1016/s0161-6420(92)31934-7] [Medline: 1584573]
4. Lundström M, Stenevi U, Thorburn W. Gender and cataract surgery in Sweden 1992-1997. A retrospective observational study based on the Swedish National Cataract Register. Acta Ophthalmol Scand 1999 Apr;77(2):204-208 [FREE Full text] [doi: 10.1034/j.1600-0420.1999.770218.x] [Medline: 10321540]
5. Buch H, Nielsen NV, Vinding T, Jensen GB, Prause JU, la Cour M. 14-year incidence, progression, and visual morbidity of age-related maculopathy: the Copenhagen City Eye Study. Ophthalmology 2005 May;112(5):787-798. [doi: 10.1016/j.ophtha.2004.11.040] [Medline: 15878058]
6. Rudnicka AR, Jarrar Z, Wormald R, Cook DG, Fletcher A, Owen CG. Age and gender variations in age-related macular degeneration prevalence in populations of European ancestry: a meta-analysis. Ophthalmology 2012 Mar;119(3):571-580. [doi: 10.1016/j.ophtha.2011.09.027] [Medline: 22176800]
7. Rim THT, Kim M, Kim WC, Kim T, Kim EK. Cataract subtype risk factors identified from the Korea National Health and Nutrition Examination survey 2008-2010. BMC Ophthalmol 2014 Jan 10;14:4 [FREE Full text] [doi: 10.1186/1471-2415-14-4] [Medline: 24410920]
8. Yoo TK, Kim SH, Kwak J, Kim HK, Rim TH. Association between osteoporosis and age-related macular degeneration: The Korea National Health and Nutrition Examination Survey. Invest Ophthalmol Vis Sci 2018 Mar 20;59(4):AMD132-AMD142. [doi: 10.1167/iovs.18-24059] [Medline: 30372730]

XSL•FO
RenderX

9. Rudnicka AR, Mt-Isa S, Owen CG, Cook DG, Ashby D. Variations in primary open-angle glaucoma prevalence by age, gender, and race: a Bayesian meta-analysis. Invest Ophthalmol Vis Sci 2006 Oct;47(10):4254-4261. [doi: 10.1167/iovs.06-0299] [Medline: 17003413]
10. Zhang X, Saaddine JB, Chou CF, Cotch MF, Cheng YJ, Geiss LS, et al. Prevalence of diabetic retinopathy in the United States, 2005-2008. JAMA 2010 Aug 11;304(6):649-656 [FREE Full text] [doi: 10.1001/jama.2010.1111] [Medline: 20699456]
11. Scheie HG, Cameron JD. Pigment dispersion syndrome: a clinical study. Br J Ophthalmol 1981 Apr;65(4):264-269 [FREE Full text] [doi: 10.1136/bjo.65.4.264] [Medline: 7236571]
12. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018 Mar;2(3):158-164. [doi: 10.1038/s41551-018-0195-0] [Medline: 31015713]
13. Mitani A, Huang A, Venugopalan S, Corrado GS, Peng L, Webster DR, et al. Detection of anaemia from retinal fundus images via deep learning. Nat Biomed Eng 2020 Jan;4(1):18-27. [doi: 10.1038/s41551-019-0487-z] [Medline: 31873211]
14. Vaghefi E, Yang S, Hill S, Humphrey G, Walker N, Squirrell D. Detection of smoking status from retinal images; a Convolutional Neural Network study. Sci Rep 2019 May 09;9(1):7180 [FREE Full text] [doi: 10.1038/s41598-019-43670-0] [Medline: 31073220]
15. Jammal AA, Thompson AC, Mariottoni EB, Berchuck SI, Urata CN, Estrela T, et al. Human versus machine: Comparing a deep learning algorithm to human gradings for detecting glaucoma on fundus photographs. Am J Ophthalmol 2020 Mar;211:123-131 [FREE Full text] [doi: 10.1016/j.ajo.2019.11.006] [Medline: 31730838]
16. Kim YD, Noh KJ, Byun SJ, Lee S, Kim T, Sunwoo L, et al. Effects of hypertension, diabetes, and smoking on age and sex prediction from retinal fundus images. Sci Rep 2020 Mar 12;10(1):4623 [FREE Full text] [doi: 10.1038/s41598-020-61519-9] [Medline: 32165702]
17. Gerrits N, Elen B, Craenendonck TV, Triantafyllidou D, Petropoulos IN, Malik RA, et al. Age and sex affect deep learning prediction of cardiometabolic risk factors from retinal images. Sci Rep 2020 Jun 10;10(1):9432 [FREE Full text] [doi: 10.1038/s41598-020-65794-4] [Medline: 32523046]
18. Ting DSW, Cheung CY, Nguyen Q, Sabanayagam C, Lim G, Lim ZW, et al. Deep learning in estimating prevalence and systemic risk factors for diabetic retinopathy: a multi-ethnic study. NPJ Digit Med 2019 Apr 10;2:24 [FREE Full text] [doi: 10.1038/s41746-019-0097-x] [Medline: 31304371]
19. Liu H, Li L, Wormstone IM, Qiao C, Zhang C, Liu P, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. JAMA Ophthalmol 2019 Dec 01;137(12):1353-1360 [FREE Full text] [doi: 10.1001/jamaophthalmol.2019.3501] [Medline: 31513266]
20. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. JAMA Ophthalmol 2017 Nov 01;135(11):1170-1176 [FREE Full text] [doi: 10.1001/jamaophthalmol.2017.3782] [Medline: 28973096]
21. 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]
22. Rim TH, Lee G, Kim Y, Tham Y, Lee CJ, Baik SJ, et al. Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms. Lancet Digit Health 2020 Oct;2(10):e526-e536 [FREE Full text] [doi: 10.1016/S2589-7500(20)30216-8] [Medline: 33328047]
23. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983 [FREE Full text] [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
24. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
25. Foong AWP, Saw S, Loo J, Shen S, Loon S, Rosman M, et al. Rationale and methodology for a population-based study of eye diseases in Malay people: The Singapore Malay eye study (SiMES). Ophthalmic Epidemiol 2007;14(1):25-35. [doi: 10.1080/09286580600878844] [Medline: 17365815]
26. Rosman M, Zheng Y, Wong W, Lamoureux E, Saw S, Tay W, et al. Singapore Malay Eye Study: rationale and methodology of 6-year follow-up study (SiMES-2). Clin Exp Ophthalmol 2012 Aug;40(6):557-568. [doi: 10.1111/j.1442-9071.2012.02763.x] [Medline: 22300454]
27. Lavanya R, Jeganathan VSE, Zheng Y, Raju P, Cheung N, Tai ES, et al. Methodology of the Singapore Indian Chinese Cohort (SICC) eye study: quantifying ethnic variations in the epidemiology of eye diseases in Asians. Ophthalmic Epidemiol 2009;16(6):325-336. [doi: 10.3109/09286580903144738] [Medline: 19995197]
28. Majithia S, Tham YC, Chee ML, Nusinovici S, Teo CL, Chee ML, et al. Cohort Profile: The Singapore Epidemiology of Eye Diseases study (SEED). Int J Epidemiol 2021;50(1):41-52 Erratum in Int J Epidemiol. 2021 Jun 28. [doi: 10.1093/ije/dyaa238] [Medline: 33393587]
29. Rim TH, Soh ZD, Tham Y, Yang HHS, Lee G, Kim Y, et al. Deep learning for automated sorting of retinal photographs. Ophthalmol Retina 2020 Aug;4(8):793-800. [doi: 10.1016/j.oret.2020.03.007] [Medline: 32362553]

XSL•FO
RenderX

30. Cheung CMG, Li X, Cheng C, Zheng Y, Mitchell P, Wang JJ, et al. Prevalence, racial variations, and risk factors of age-related macular degeneration in Singaporean Chinese, Indians, and Malays. Ophthalmology 2014 Aug;121(8):1598-1603. [doi: 10.1016/j.ophtha.2014.02.004] [Medline: 24661862]
31. Illingworth J, Kittler J. The adaptive hough transform. IEEE Trans Pattern Anal Mach Intell 1987 May;9(5):690-698. [doi: 10.1109/tpami.1987.4767964] [Medline: 21869428]
32. Graham B. Kaggle diabetic retinopathy detection competition report. University of Warwick 2015:24-26 [FREE Full text]
33. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv. Preprint posted online on April 10, 2015. [FREE Full text]
34. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. Tensorflow: A system for large-scale machine learning. 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16); 2016 Presented at: OSDI'16: The 12th USENIX conference on Operating Systems Design and Implementation; November 2-4, 2016; Savannah, Georgia p. 265-283.
35. TensorFlow. TensorFlow. URL: https://www.tensorflow.org/ [accessed 2021-07-20]
36. Xu K, Ba JL, Kiros R, Cho K, Courville A, Salakhudinov R, et al. Show, attend and tell: neural image caption generation with visual attention. 2015 Presented at: ICML'15: The 32nd International Conference on International Conference on Machine Learning; July 6-11, 2015; Lille, France p. 2048-2057.
37. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2019 Oct 11;128:336-359 [FREE Full text] [doi: 10.1007/s11263-019-01228-7]
38. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Front Neuroinform 2014 Feb 21;8:14 [FREE Full text] [doi: 10.3389/fninf.2014.00014] [Medline: 24600388]
39. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837-845. [Medline: 3203132]
40. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. Artificial intelligence in retina. Prog Retin Eye Res 2018 Nov;67:1-29 [FREE Full text] [doi: 10.1016/j.preteyeres.2018.07.004] [Medline: 30076935]
41. Korot E, Pontikos N, Liu X, Wagner SK, Faes L, Huemer J, et al. Predicting sex from retinal fundus photographs using automated deep learning. Sci Rep 2021 May 13;11(1):10286 [FREE Full text] [doi: 10.1038/s41598-021-89743-x] [Medline: 33986429]
42. Liu P, Gu Z, Liu F, Jiang Y, Jiang S, Mao H. Large-scale left and right eye classification in retinal images. 2018 Presented at: Ophthalmic Medical Image Analysis 2018 and Computational Pathology and Ophthalmic Medical Image Analysis 2018; September 16-20, 2018; Granada, Spain p. 261-268. [doi: 10.1007/978-3-030-00949-6_31]
43. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. 2018 Presented at: Conference on fairness, accountability and transparency; February 23-24, 2018; New York City.
44. Goodman B. A step towards accountable algorithms? Algorithmic discrimination and the European union general data protection. 2016 Presented at: 29th Conference on Neural Information Processing Systems (NIPS), Barcelona NIPS Foundation; 2016; Barcelona.
45. Ting DSW, Peng L, Varadarajan AV, Keane PA, Burlina PM, Chiang MF, et al. Deep learning in ophthalmology: The technical and clinical considerations. Prog Retin Eye Res 2019 Sep;72:100759. [doi: 10.1016/j.preteyeres.2019.04.003] [Medline: 31048019]

Abbreviations

AUC: area under the receiver operating curve
ETDRS: Early Treatment for Diabetic Retinopathy Study
EyePACS: Eye Picture Archive Communication System
Grad-CAM: gradient-weighted class activation mapping
SEED: Singapore Epidemiology of Eye Diseases

Edited by C Lovis; submitted 21.10.20; peer-reviewed by JA Benítez-Andrades, V Franzoni; comments to author 05.12.20; revised version received 06.04.21; accepted 22.06.21; published 17.08.21.

Please cite as:
Betzler BK, Yang HHS, Thakur S, Yu M, Quek TC, Soh ZD, Lee G, Tham YC, Wong TY, Rim TH, Cheng CY
Gender Prediction for a Multiethnic Population via Deep Learning Across Different Retinal Fundus Photograph Fields: Retrospective Cross-sectional Study
JMIR Med Inform 2021;9(8):e25165
URL: https://medinform.jmir.org/2021/8/e25165
doi:10.2196/25165
PMID:34402800

XSL•FO
RenderX

©Bjorn Kaijun Betzler, Henrik Hee Seung Yang, Sahil Thakur, Marco Yu, Ten Cheer Quek, Zhi Da Soh, Geunyoung Lee, Yih-Chung Tham, Tien Yin Wong, Tyler Hyungtaek Rim, Ching-Yu Cheng. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 17.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Worker-Centered Personal Health Record App for Workplace Health Promotion Using National Health Care Data Sets: Design and Development Study

Hyun Sang Park[1,2*], MS; Kwang Il Kim[3], MS; Ho-Young Chung[2], MD, PhD; Sungmoon Jeong[2], PhD; Jae Young Soh[1], MS; Young Ho Hyun[1]; Hwa Sun Kim[4*], RN, PhD

[1]Digital Healthcare Department, BIT Computer Co. Ltd., Seoul, Republic of Korea
[2]Department of Medical Informatics, Kyungpook National University, Daegu, Republic of Korea
[3]Finance Programs Department, Korea Occupational Safety and Health Agency, Ulsan, Republic of Korea
[4]Elecmarvels Co. Ltd., Daegu, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Hyun Sang Park, MS
Digital Healthcare Department
BIT Computer Co. Ltd.
BIT Building 33, Seocho-daero 74-gil, Seocho-gu
Seoul, 06621
Republic of Korea
Phone: 82 2 3486 1234 ext 507
Fax: 82 2 3486 1983
Email: hspark@bit.kr

Abstract

Background: Personal health record (PHR) technology can be used to support workplace health promotion, and prevent social and economic losses related to workers' health management. PHR services can not only ensure interoperability, security, privacy, and data quality, but also consider the user's perspective in their design.

Objective: Using Fast Healthcare Interoperability Resources (FHIR) and national health care data sets, this study aimed to design and develop an app for providing worker-centered, interconnected PHR services.

Methods: This study considered the user's perspective, using the human-centered design (HCD) methodology, to develop a PHR app suitable for occupational health. We developed a prototype after analyzing quantitative and qualitative data collected from workers and a health care professional group, after which we performed a usability evaluation. We structured workers' PHR items based on the analyzed data, and ensured structural and semantic interoperability using FHIR, Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT), and Logical Observation Identifiers Names and Codes (LOINC). This study integrated workers' health information scattered across different Korean institutions through an interface method, and workers' PHRs were managed through a cloud server, using Azure API for FHIR.

Results: In total, 562 workers from industrial parks participated in the quantitative study. The preferred data items for PHR were medication, number of steps walked, diet, blood pressure, weight, and blood glucose. The preferred features were ability to access medical checkup results, health information content provision, consultation record inquiry, and teleconsultation. The worker-centered PHR app collected data on, among others, life logs, vital signs, and medical checkup results; offered health care services such as reservation and teleconsultation; and provided occupational safety and health information through material safety data sheet search and health questionnaires. The app reflected improvements in user convenience and app usability proposed by 19 participants (7 health care professionals and 12 end users) in the usability evaluation. The After-Scenario Questionnaire (ASQ) was evaluated with a mean score of 5.90 (SD 0.34) out of 7, and the System Usability Scale (SUS) was evaluated a mean score of 88.7 (SD 4.83) out of 100.

Conclusions: The worker-centered PHR app integrates workers' health information from different institutions and provides a variety of health care services from linked institutions through workers' shared PHR. This app is expected to increase workers' autonomy over their health information and support medical personnel's decision making regarding workers' health in the

XSL•FO
RenderX

workplace. Particularly, the app will provide solutions for current major PHR challenges, and its design, which considers the user's perspective, satisfies the prerequisites for its utilization in occupational health.

(JMIR Med Inform 2021;9(8):e29184) doi:10.2196/29184

KEYWORDS

personal health record app; workplace health promotion; Fast Healthcare Interoperability Resources; national health care data set; human-centered design

Introduction

Background

Changes in lifestyle habits and the spread of chronic diseases have increased health problems within companies [1]. Workforce health is increasingly important for market relevance; the World Health Organization (WHO) showed the physical and mental health of workers to be imperative to companies' success and competitive edge [2]. Compared with the general public, workers are at an increased risk of stress caused by a heavy workload and unhealthy lifestyle, including lack of exercise and frequent drinking [3]. Workers' health may be directly or indirectly linked to work efficiency, corporate productivity, and industrial accidents beyond the individual level. Managing workers' health at the corporate level can prevent social and economic losses, and employers are increasingly interested in improving workers' health and welfare as a corporate strategy [4-6].

The workplace, where workers spend most time [7], is the best place to apply the concept of health promotion. The concept of workplace health promotion denotes that employers, workers, and local communities work together to improve workers' mental and physical health and welfare [8]. Workplace health promotion initiatives can foster an appropriate work environment and promote personal health management [9,10]. Its primary challenge is increasing worker participation; studies have shown participation rates of less than 50% [11] and average annual reduction rates of 28% [12]. These obstacles can be overcome by applying health care technology to workplace health promotion [13].

Applications of health care technology, such as the personal health record (PHR), can increase workers' interest, motivation, and participation in workplace health promotion [14,15] through its technology-based attributes [16]. PHR allows users to systematically collect, process, store, and share their health information with others, such as family members or medical personnel [17]. PHR users can easily access their medical records, prescription drug information, hospital test results, and health promotion information [18]. Given that the use of PHR promotes cooperation between medical personnel and workers through communication, it can help reduce medical expenses and strengthen disease prevention, management, and treatment activities [19,20]. Because of the expected effects of PHR, it is increasingly provided by employers as part of self-managed health care programs [21,22].

PHR is intended to help workers manage their health information, but privacy concerns have evoked obstacles to its use [21]. Workers are often reluctant to allow employers to access their PHR, raising direct practical problems [23]. Concerns about the exposure of personal information and fear of discrimination are often discussed as privacy and security issues of PHR [24], and workers may question the motives of employers who provide such services [25]. Various factors influence PHR system acceptance and use [21], with workers' acceptance of PHR being influenced by individual and organizational factors (eg, trust in employer, management support for PHR, communication, and awareness), along with technical factors [16]. Workers' participation depends on incentive provision and how PHR is presented to them [26].

Privacy issues, lack of motivation, and operational difficulties have been identified as major obstacles for the use of PHR [27], with various studies promoting the use of PHR. Pushpangadan and Seckman [28] argued that consumer adoption was slow because PHR was designed based on a clinically oriented design, without considering the consumers' perspective. Weinert and Cudney [29] showed that PHR efficiency depends not only on system evolution and complexity, but also on user-friendliness, easy-to-use design, and structured documents. Thus, developing a successful PHR app may entail considering users' perspectives from the design stage, coupled with a systematic design methodology.

In terms of data access, workers currently must collect their health information from individual institutions and workplaces, which complicates individuals' active participation in their health management. Interconnected PHR services, where workers collect and manage their health information in one place, with users controlling others' access to their information, may provide a solution to this challenge. Data exchange based on workers' authorization is possible only when the structural and semantic interoperability of the PHR is guaranteed. Interoperability [30-32] is important in workers' adoption of PHR and is known to be a major challenge for PHR, in addition to security and privacy [33] and data quality [17]. A successful workplace PHR app service can be developed and operated by making the app user centric, and ensuring interoperability, security and privacy, and data quality.

This study aimed to design and develop a PHR app providing a worker-centered interconnected PHR service. To this end, we designed a PHR app following the analysis of quantitative and qualitative data collected from workers and a group of health care professionals, employing the human-centered design (HCD) methodology. We developed the app based on national health care data sets using web technologies.

Prior Work

Studies have been conducted to standardize PHR and address interoperability issues. Simon et al [34] developed a PHR that

XSL•FO
RenderX

acquires measured data from a device through IEEE 11073, converts them to ASTM continuity of care record (CCR), and transmits them to a server. Marceglia et al [35] proposed a design based on Health Level Seven (HL7) clinical document architecture (CDA) that can be adopted when exchanging information between PHR and electronic health records (EHRs). Plastiras and O'Sullivan [36] developed an ontology-based architecture model that can ensure interoperability between PHR and EHR using various standards, such as CCR and CDA. Li [37] proposed a mobile PHR using various standards such as CDA, Digital Imaging and Communications in Medicine (DICOM), Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT), and Logical Observation Identifiers Names and Codes (LOINC).

Since the introduction of Fast Healthcare Interoperability Resources (FHIR), studies have been conducted to apply it to PHR. Hong et al [38] developed a PHR system using FHIR and internet of things cloud to build an interconnected PHR, thereafter conducting a clinical trial to develop an obesity management model for 500 patients. Saripalle et al [39] developed a prototype of a tethered mobile PHR using FHIR and OpenEMR, and the developed mobile PHR synchronizes user data stored in OpenEMR using an HL7 Application Programming Interface (HAPI) library [40].

Studies have been conducted to develop PHR by applying various design approaches. Farinango et al [41] developed a PHR system for metabolic syndrome management by applying the HCD methodology. Farinango et al further developed 3 prototypes through 5 iterations by collecting user information through a survey of 1187 respondents, 8 interviews, and focus group interviews (FGIs) with 7 people. Zhou et al [42] developed and evaluated a mobile PHR app through a user-centered design (UCD) methodology, which involved using survey data from 609 respondents, and then conducting a usability evaluation on 15 participants. The UCD methodology has been used in other studies as well. For instance, Massoudi et al [43] developed a PHR that supports lifestyle intervention by applying the UCD methodology. They conducted structured interviews with 42 participants (28 users, 8 health care professionals, and 6 personal trainers) and user tests on 16 participants. Marchak et al [44] also applied the UCD methodology to develop and evaluate a web-based PHR for survivors of childhood cancer; they conducted FGIs and structured interviews with 28 patients (3 patients with pediatric cancer, 11 parents, and 14 health care providers), and a usability evaluation with 16 participants.

Various studies have also been conducted on workers and employers to operationalize the PHR. Dawson et al [22] conducted a questionnaire survey to understand workers' perceptions (in large companies) of PHR; results showed that the reason for the low confidence in the PHR was a lack of trust in employers and other employees who may have access to employees' health information. Fernando et al [45] analyzed the demographic characteristics of workers and health-related productivity (absence and overwork) related to PHR; results showed that high performers had a high absenteeism rate, indicating that PHR needs to focus on high performers. Fernando et al [46] also conducted quantitative and qualitative research on workers and employers to design the data model of PHR, thereafter developing and evaluating web-based PHR prototypes [47].

Fast Healthcare Interoperability Resources

Occupational factors, such as patients' workplace environment, need to be considered when managing chronic diseases; thus, occupational information has been integrated into the EHR [48] or an occupational data for health model [49]. HL7 has been used to design an FHIR profile [50] to represent patients' occupational elements in PHR. The FHIR [51] was developed by HL7 in 2014 and is a next-generation standard for EHR exchange. It utilizes a reference information model, lightweight web services, and the latest web and app development principles. It was developed based on lessons learned from the HL7 standard and expert experiences. V2, which focused on the message-based exchange, required customization owing to semantic inconsistencies in its implementation [52]. The V3 reference information model provided a framework for expressing semantically consistent clinical statements, but owing to the complexity of its implementation, compatibility between system and document was hindered [53,54]. The FHIR was designed to be concise and easy to understand by adopting the advantages of the existing HL7 standard.

The FHIR simplifies various types of information generated in the medical field and expresses all contents as exchangeable resources. Each resource has its original form, and they refer to the URL of a resource only when the content of another resource is needed. When FHIR expresses EHR, it is expressed as a combination of various resources, such as Lego blocks, so that information can be easily recycled and only the necessary resources can be updated. Currently, there are over 150 resources, including clinical concepts (eg, allergy, condition, family member history, medication, and observation) and administrative information (eg, patient, practitioner, organization, and location). These resources are provided to external systems and clients through RESTful application programming interfaces (APIs). Regarding data exchange, transport layer security should be used, with OpenID Connect and OAuth being recommended for user identification, authentication, and authorization.

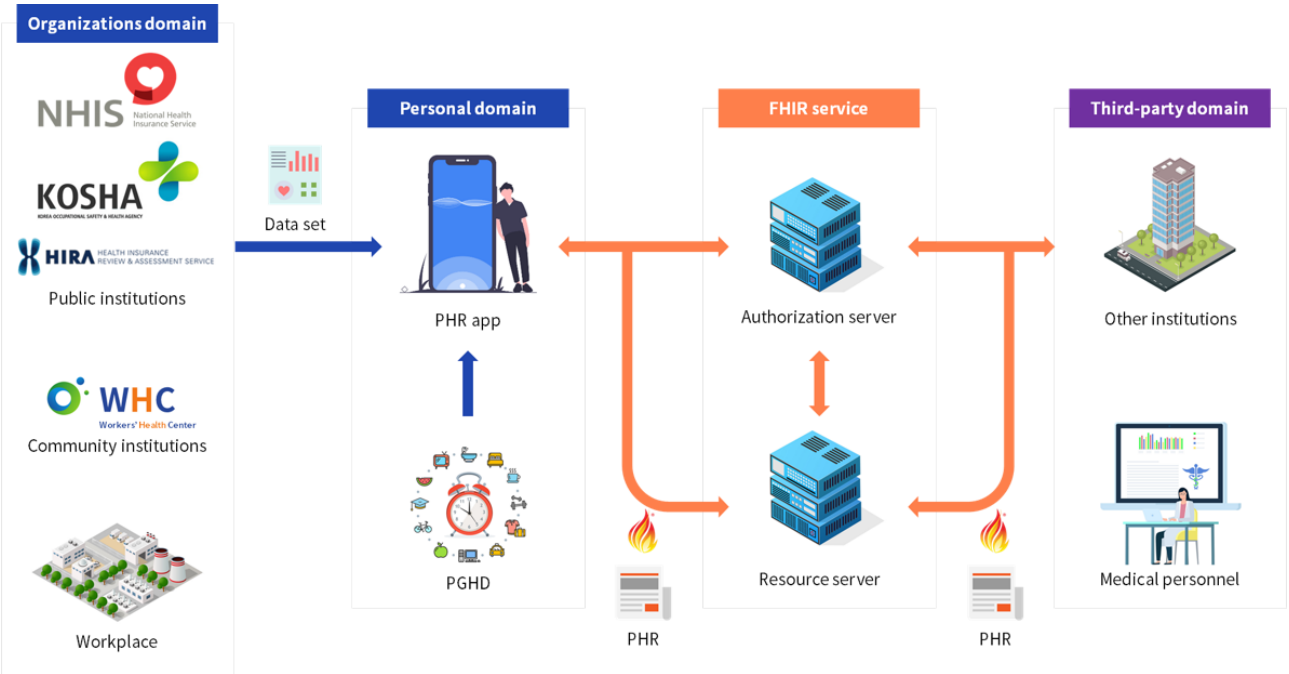
Methods

Service Design

Worker-centered PHR services ensure continuity of care outside the workplace by allowing workers to easily collect their health information from various sources and manage it as PHR (Figure 1). Although PHR has become technically safe, users still must manually input their data [38]. Generally, health information is generated from a variety of sources (eg, health care providers, insurance companies, social networks, mass media, and public institutions) [55], and the generation of interoperable PHR requires the integration of data from different sources [56].

XSL•FO
RenderX

Figure 1. Conceptual diagram of a worker-centered personal health record service. FHIR: Fast Healthcare Interoperability Resources; PGHD: patient-generated health data; PHR: personal health record.



This study collected users' health information through an API, and used an authentication method set by each institution. Institutions were classified into public and community institutions and workplaces, according to the data management entity. Public PHR sources were the National Health Insurance Service, Korea Occupational Safety and Health Agency, and the Health Insurance Review and Assessment Service. These institutions manage medical treatment history, prescription history, medical checkup results, and medical institution information according to the role of the institution, and users can view the data at their request. These data are national data generated when persons eligible for national health insurance access services provided by medical institutions.

Workers' health centers are community institutions in Korea that provide services to prevent occupational diseases among workers in industrial parks (incorporating various industries, including manufacturing plants and factories). Currently, there are 23 centers in operation. Each institution comprises professional personnel, such as occupational and environmental medicine specialists, occupational nurses, industrial hygiene safety engineers, physical therapists, and counseling psychologists, who provide comprehensive occupational health services, including occupational, cerebrovascular, and musculoskeletal disease prevention, and job stress prevention. All workers can visit their nearest center and use its services free of charge, similar to a workplace infirmary. Workers' health centers systematically manage the information of workers and workplaces in their area through an integrated system [57].

The workplace refers to the company employing the worker, where an occupational health manager manages workers' information generated through the workplace health promotion program. As such, workers' information is scattered across various sources, and needs to be managed in an integrated manner to ensure effective workplace health promotion.

Unlike an EHR, PHR can add patient-generated health data (PGHD). The PHR app from this study acquired data using various devices (eg, smartphone sensors, wearable bands, blood pressure monitors, blood glucose meters, and scales) and integrated these data with health information collected from each institution. These integrated data can be converted into an FHIR-compliant PHR according to the users' needs, and then managed through a cloud service. The FHIR service comprises authentication procedures and resource servers that allow safe data management in the cloud by restricting access to users' resources only to authorized institutions.

Design Methodology

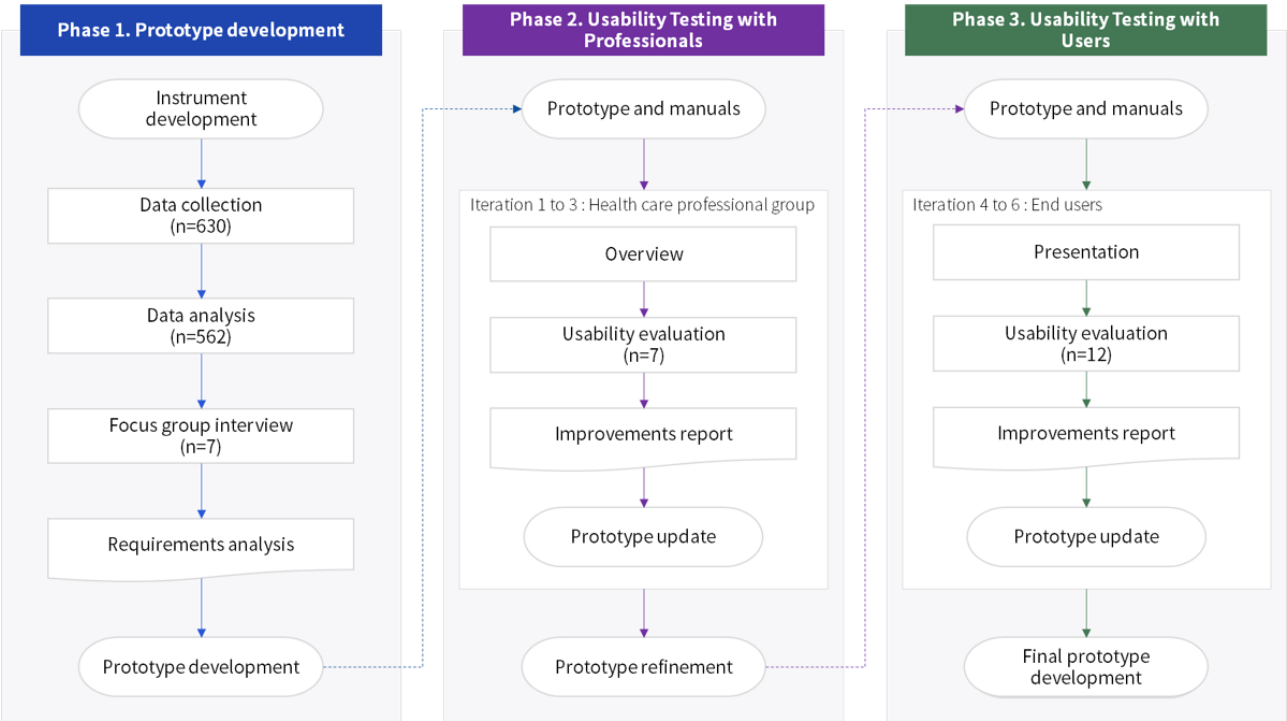
This study applied the HCD methodology to design and develop a worker-centered PHR app (Figure 2). The goal was to develop a prototype based on quantitative and qualitative data analysis, and improve it through repeated usability evaluations. After defining the features of the prototype PHR app through benchmarking and a literature review, a questionnaire was developed through consultation with a group of health care professionals with advanced practice nurse licenses, including occupational health managers with experience in computerization in the workplace, a public institution practitioner, and a professor. The questionnaire consisted of 17 items, of which 12 enquired about participants' general characteristics (sex, age, marital status, education, workplace, etc.), and 5 were configured to allow up to 3 responses on required data items and app features. Next, with the cooperation of the Korea Occupational Safety and Health Agency, we conducted a survey among workers in industrial parks in Korea, who had visited workers' health centers (21 in total). Considering regional distribution, we included 30 workers from each center. We explained the background and purpose of the study, as well as the envisioned PHR app, to the participants, and questionnaires were distributed to those who had provided their consent. The survey was conducted for approximately 3

XSL•FO
RenderX

weeks (from November 9, 2018, to November 30, 2018). In total, 630 questionnaires were distributed and 575 were collected. Of the collected questionnaires, 13 were excluded because they did not meet study aims or included insincere responses, thus being inappropriate for analysis.

We conducted a frequency analysis of participants' demographic data, and multiple response analysis of data items and app feature preferences. The results were relayed within FGI with the health care professional group, to inform the design of user profiles, requirements, interface concepts, and information architecture for the PHR app before developing the prototype.

Figure 2. A scheme of the phases for a human-centered design approach to developing a worker-centered personal health record app.



Usability Study

The prototype was evaluated by the health care professional group and end users (ie, workers who will be using the app). The health care professional group received applications from occupational health practitioners interested in participating in the study and usability evaluation. Participants in the evaluation study were selected based on their experience and occupation. The health care professional group meeting was conducted in a conference room with a large table to allow interaction among participants. First, we distributed the use cases and manuals to the health care professional group. Next, we performed a cognitive walkthrough of the prototype. For the health care professional group, we evaluated the usability of the scenario every time the task was completed with the After-Scenario Questionnaire (ASQ) [58], and the usability of the prototype was evaluated using the System Usability Scale (SUS) [59]. The SUS consisted of 10 items rated on a 5-point scale, ranging from 1 (Strongly disagree) to 5 (Strongly agree); it was converted into a total score between 0 and 100 points to evaluate the entire system. The ASQ consisted of 3 items, rated on a 7-point scale, ranging from 1 (Strongly disagree) to 7 (Strongly agree); each item of the ASQ evaluated the effectiveness, efficiency, and satisfaction of the task.

After reflecting on the prototype improvements derived through this process, the usability evaluation was performed for end users. The end users received online applications from individuals interested in participating in the study and usability evaluation, and the final participants were selected through random selection. End user evaluations were conducted individually to ensure privacy. The same methodology used for the health care professional group was applied to the 12 workers who participated in the usability evaluation; the research manager introduced the features of the app before performing the task, demonstrated the unique features of the app, and participants suggested improvements during interviews held after the task had been performed. The usability evaluation was conducted for approximately 6 months (from January to June 2019) and a total of 6 iterations were performed, 3 per group. At the end of each iteration, the prototype was improved based on the analyzed qualitative data, and tests were performed on existing participants (ie, 7 health care professionals and 12 end users). This study was conducted with the approval of the Korea Occupational Safety and Health Agency after a review of its research ethics (No. 211960314-00).

Structural and Semantic Interoperability

Before designing a PHR with guaranteed interoperability, we analyzed data from various sources and structured workers' PHR items by category. The basic information category comprised demographic information, personal history, family history, occupational history, and lifestyle. Data on these variables were collected by analyzing the database schema of the integrated system used in the workers' health centers, and the document received from 5 occupational health managers.

XSL•FO
RenderX

The treatment and prescription history category consisted of hospital information, visit date, treatment type, hospitalization days, pharmacy information, medication frequency, and drug information. Data on these variables were collected by linking public data provided by the National Health Insurance Service and the Health Insurance Review and Assessment Service.

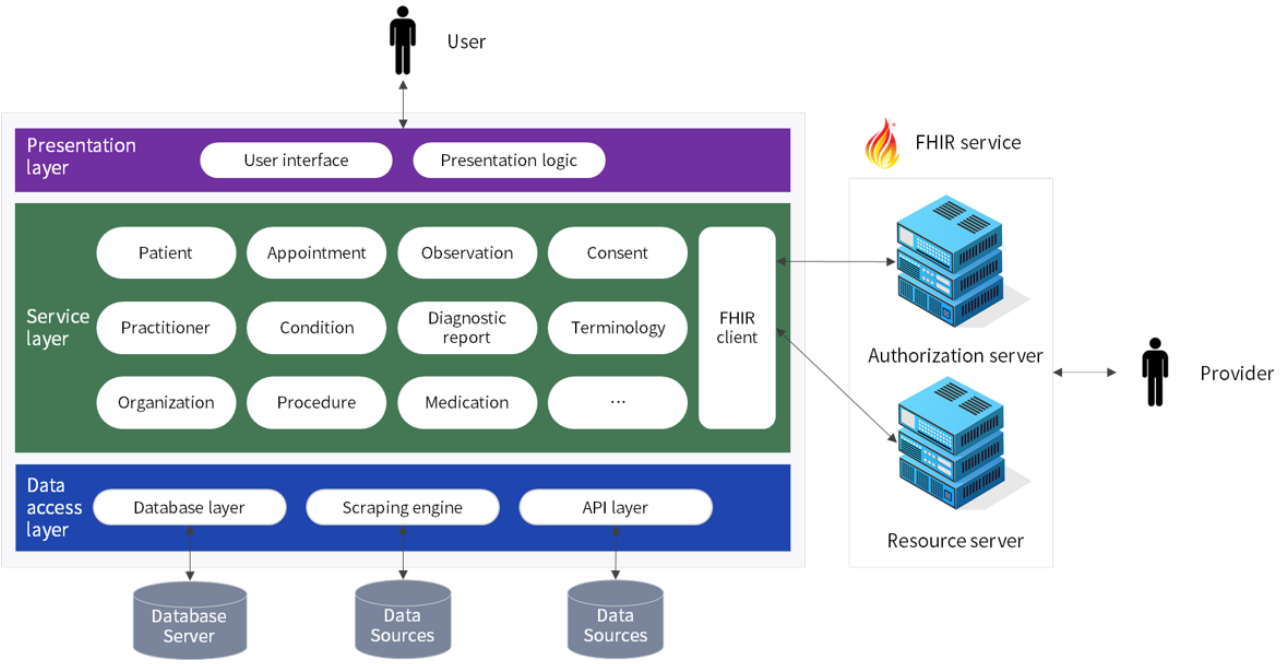
The medical checkup category referred to general medical checkup undertaken by the National Health Insurance Service, special medical checkup undertaken by the Korea Occupational Safety and Health Agency, target harmful factors, test methods, reference values, and units for each test item. Data on these variables were collected by analyzing the medical checkup results table and workers' medical checkup guidelines.

The standardization process was performed after establishing content validity (selection of items, review of classification, reference value, and units, etc.) of the structured PHR items of workers; validity was evaluated by 5 occupational health managers. For structural interoperability, workers' PHR was modeled through mapping between resource subitems and inspected PHR items after selecting FHIR resources corresponding to each category. For semantic interoperability, an appropriate code was defined through mapping between the concepts of SNOMED-CT and LOINC for the item representing the measured value of users. The mapped results were cross-validated by 2 experts: a laboratory medicine specialist and a medical informatics and nursing PhD graduate (HK).

Architecture

We developed a PHR app, named Workcare, which enables workers to systematically collect and store their health information from various sources and devices, and receive continuous health care services through data sharing. Workcare is an interconnected PHR app that secures ease of data entry, updates data using national health care data sets, guarantees the interoperability of PHR through a standardization process, and provides features for workers' health management through the linkage between independent modules (Figure 3).

Figure 3. The architecture of the interconnected personal health record app Workcare using FHIR. API: application programming interface; FHIR: Fast Healthcare Interoperability Resources.



The data access layer collects users' health information from various sources and stores it in a database. After user authentication, the API layer requests data through the API provided by each institution, and parses the response data; this allows access to test results and consultation records stored at the workers' health center, or information from institutions utilized by the user, gathered from the hospital and pharmacy information provided by the Health Insurance Review and Assessment Service. The Scraping Engine is a screen scraping module developed to collect information from institutions that do not currently provide an API. This module first processes the authentication agent of a web service using a certificate stored in the smartphone, and then delivers the session to process the content of a specific site. In this way users can access multiple institutional websites to collect scattered health information by providing authentication information only once.

The database layer stores users' collected health information in a database, and updates data generated by the events of users and third parties. The data exchange between the client and the database server complies with the JavaScript Object Notation (JSON) through hypertext transfer protocol over secure socket layer (HTTPS). Data are securely transmitted by applying secure socket layer (SSL), and personal information is encrypted and decrypted by applying ARIA256 and SHA256.

The service layer implements the function of the PHR app through linkage between other layers. In this study, the functions were configured according to the HL7 PHR-S FM [60], a framework that lists the functions required or desirable for PHR, and complied with the standardized model of the PHR system. This procedure combines the FHIR resources (patient, appointment, observation, etc.) and FHIR client to provide the essential functions defined in the PHR-S FM. The FHIR client

XSL•FO
RenderX

implements interconnected PHR services through the linkage between FHIR services. This converts health information that was collected based on the modeled workers' PHR into FHIR resources, and then transmits the PHR to the FHIR service, or even parses the PHR delivered to the FHIR service. Notwithstanding, before data exchange with the resource server, user authentication and authorization are checked from the authentication server through OAuth 2.0, and only those authorized by the users can access their resources. We employed the Azure API cloud services, provided by Microsoft, for FHIR services [61].

The presentation layer provides users with the user interface/user experience for using the PHR app. Workcare was developed as a hybrid app, thus providing the same screen for the user, regardless of the resolution of the operating system (Android, iOS) or device type, thus complying with the mobile design guidelines derived from the improvements report extracted through the usability evaluation.

Results

Quantitative Data Analysis

Participants were 562 workers who visited 21 workers' health centers in Korea. Most workers were women and older than 50 years, followed by those in their 40s and 30s. The most common duration of employment in the workplace was 1-4 years, and 63.9% (359/562) of the participants were employed in workplaces with less than 50 employees. Clerical and service-based businesses were more common than production and technical businesses (Table 1).

The results of the multiple response analysis for data items and feature preferences of the PHR apps are shown in Table 2. Regarding lifelogs to track, medication was the preferred feature, followed by the step count and diet. Regarding health data to track, blood pressure, weight, and blood glucose outranked body composition, body temperature, and oxygen saturation. Regarding information to manage, the highest preference was for examination result and the lowest for exercise. In terms of workplace health promotion, the preferences were, in order, for content provision, consultation record inquiry, and expert consultation. For other features, the preferences were, in order, data linkage, disease prediction, and material safety data sheet inquiry.

XSL•FO
RenderX

Table 1. Participants' characteristics (N=562).

Characteristic	Value, n (%)
Sex	
Male	195 (34.7)
Female	367 (65.3)
Age (years)	
<20	7 (1.2)
20-29	94 (16.7)
30-39	132 (23.5)
40-49	155 (27.6)
≥50	174 (31.0)
Marital status	
Single	207 (36.8)
Married	345 (61.4)
Widowed	7 (1.2)
Divorced or separated	3 (0.5)
Education	
Middle school	21 (3.7)
High school	142 (25.3)
College (2 years)	87 (15.5)
College (4 years)	270 (48.0)
Graduate school	42 (7.5)
Duration of employment in the workplace (years)	
<1	102 (18.1)
1-4	227 (40.4)
5-9	95 (16.9)
≥10	138 (24.6)
Number of employees in the workplace	
<5	62 (11.0)
5-9	75 (13.3)
10-29	98 (17.4)
30-49	124 (22.1)
50-99	37 (6.6)
≥100	166 (29.5)
Type of business	
Production	59 (10.5)
Clerical	227 (40.4)
Service based	185 (32.9)
Technical	33 (5.9)
Other	58 (10.3)
Previous experience with health care app	
Yes	189 (33.6)
No	373 (66.4)

XSL•FO
RenderX

Table 2. Summary of data items and feature preferences for the personal health record app.

Contents	Value, n (%)
Lifelogs to track (n=1040)	
Medication	272 (26.15)
Step count	257 (24.71)
Diet	159 (15.29)
Stress	89 (8.56)
Exercise	86 (8.27)
Smoking	54 (5.19)
Drinking	48 (4.62)
Caffeine	45 (4.33)
Water	30 (2.88)
Health data to track (n=1024)	
Blood pressure	352 (34.38)
Weight	272 (26.56)
Blood glucose	249 (24.32)
Body composition	87 (8.50)
Temperature	38 (3.71)
Oxygen saturation	26 (2.54)
Information to manage (n=1196)	
Examination result	239 (19.98)
Health data	221 (18.48)
Prescription history	187 (15.64)
Lifelogs	182 (15.22)
Diet	143 (11.96)
Treatment history	142 (11.87)
Exercise	82 (6.86)
Workplace health promotion (n=1178)	
Content provision	314 (26.66)
Consultation record inquiry	285 (24.19)
Expert consultation	244 (20.71)
Reservations	152 (12.90)
Campaigns	101 (8.57)
Community	82 (6.96)
Other features (n=1224)	
Data linkage	289 (23.61)
Disease prediction	235 (19.20)
Material safety data sheet inquiry	234 (19.12)
Body age analysis	198 (16.18)
Health questionnaire	181 (14.79)
Medical institution inquiry	87 (7.11)

XSL•FO
RenderX

Qualitative Data Analysis

Overview

In total, 19 participants were part of the usability evaluation, including 7 health care professionals (Table 3) and 12 end users (Table 4). Most health care professionals were women, and most were in their 40s. They were licensed as advanced practice nurses. Their most common occupation was occupational health manager; all had more than 5 years' experience in the related field, and 4 had previously used health care apps.

Similar to the health care professional group, most end users were women and in their 40s. Most had been employed in the same workplace for 5-9 years, and 5 had previously used health care apps.

The usability of the scenario (Table 5) and the prototype (Table 6) improved the results according to the iteration. The final ASQ was evaluated at a high level, with an average score of 5.90 (SD 0.43) out of 7. The final SUS was evaluated at an average score of 88.7 (SD 4.83) out of 100.

Table 3. Characteristics of health care professionals (N=7).

Characteristic	Value, n (%)
Sex	
Male	1 (14)
Female	6 (86)
Age (years)	
30-39	2 (29)
40-49	4 (57)
≥50	1 (14)
Marital status	
Single	1 (14)
Married	6 (86)
Education	
College (4 years)	1 (14)
Graduate school	6 (86)
Career(years)	
5-9	5 (71)
≥10	2 (29)
Type of occupation	
Occupational health manager	5 (71)
Professor	1 (14)
Official	1 (14)
Previous experience with health care app	
Yes	4 (57)
No	3 (43)

XSL•FO
RenderX

Table 4. Characteristics of end users (N=12).

Characteristic	Value, n (%)
Sex	
Male	3 (25)
Female	9 (75)
Age (years)	
30-39	2 (17)
40-49	7 (58)
≥50	3 (25)
Marital status	
Single	3 (25)
Married	9 (75)
Education	
High school	4 (33)
College (2 years)	2 (17)
College (4 years)	6 (50)
Duration of employment in the workplace (years)	
1-4	2 (17)
5-9	7 (58)
≥10	3 (25)
Previous experience with health care app	
Yes	5 (42)
No	7 (58)

Table 5. Usability evaluation results of scenario's task[a].

Section	Task 1[b]	Task 2[c]	Task 3[d]	Task 4[e]	Task 5[f]	Task 6[g]	Average	Target
Phase 2								Health care professional group
Iteration 1	5.29 (0.41)	4.67 (0.33)	5.38 (0.49)	4.95 (0.33)	4.62 (0.33)	5.33 (0.33)	5.04 (0.37)	
Iteration 2	5.67 (0.25)	5.24 (0.41)	5.86 (0.47)	5.00 (0.49)	5.10 (0.33)	5.67 (0.41)	5.42 (0.39)	
Iteration 3	6.19 (0.49)	5.62 (0.49)	6.29 (0.41)	5.76 (0.56)	5.76 (0.47)	6.10 (0.31)	5.95 (0.46)	
Phase 3								End users
Iteration 4	5.22 (0.34)	4.53 (0.29)	5.28 (0.38)	4.56 (0.42)	4.61 (0.48)	5.31 (0.48)	4.92 (0.40)	
Iteration 5	5.64 (0.32)	5.17 (0.34)	5.78 (0.61)	5.03 (0.46)	5.17 (0.47)	5.56 (0.56)	5.39 (0.46)	
Iteration 6	6.17 (0.43)	5.53 (0.24)	6.22 (0.37)	5.61 (0.42)	5.69 (0.32)	6.17 (0.59)	5.90 (0.43)	

[a]All values are presented as mean (SD).

[b]Task 1: After entering your account information, log in to the personal health record app.

[c]Task 2: After providing certification, import the national health care data sets.

[d]Task 3: After adding health data values, look at the stored values.

[e]Task 4: Select the data sharing range and upload the personal health record to the Fast Healthcare Interoperability Resources service.

[f]Task 5: After connecting the system, use the linked institutions' services.

[g]Task 6: Use the services after checking the provided occupational health content.

XSL•FO
RenderX

Table 6. Usability evaluation results of prototype.

Section	Mean (SD)	Target
Phase 2		Health care professional group
Iteration 1	83.2 (3.45)	
Iteration 2	85.4 (2.50)	
Iteration 3	86.9 (1.73)	
Phase 3		End users
Iteration 4	86.2 (6.83)	
Iteration 5	87.2 (5.05)	
Iteration 6	88.7 (4.83)	

Suggested Improvements: Health Care Professionals

The major improvements derived from the usability evaluation of the health care professional group are provided below. The health care professional group suggested improvements for the features and contents of the PHR app, and improvements with similar contents were integrated into a single category.

Lifelogs

> *Medication, smoking, and alcohol are essential items for managing workers' lifestyle habits and calculating risk factors for cerebrovascular disease. If workers can calculate risk factors for developing cerebrovascular diseases by data collected via the PHR app and self-tests, it will be a motivation for health management.* [Health care professional 6]

> *Since the type of food, calories, and nutritional contents differ by database, accurate information [about food intake] cannot be recorded and managed [in the app]. Also, according to past experiences of using existing health care apps, the process of searching and recording food intake was cumbersome.* [Health care professional 1]

Medical Checkup

> *Most construction workers are daily workers, so it is not easy to manage their medical checkup results. If daily workers can manage their individual medical checkup results through the PHR app, it will be of great help to occupational health managers who have recently moved to new workplaces.* [Health care professional 2]

> *Medical checkup results are sent to individual workers, and workers often lose them, so they do not bring them when consulting with an occupational health manager. The PHR app should enable the easy sharing of PHR to occupational health managers through user authentication and consent.* [Health care professional 1]

> *Most older adult workers do not have a certificate on their smartphones. In consideration of these classes, it is necessary to improve the feature of the app, so that the medical checkup results can be managed as images.* [Health care professional 3]

Harmful Factors

> *Even if workers are trained through material safety data sheets, they must be notified by the occupationalhealth manager. If it is possible to provide information on harmful factors for each user's work area through the PHR app, it may greatly help occupationalhealth managers' work convenience and workers' access to information.* [Health care professional 5]

> *It would be helpful if we could provide customized content according to the user's business and occupation. For example, it would be of great significance if workers could check information on precautions and harmful factors for their assigned processes through the PHR app.* [Health care professional 7]

Suggested Improvements: End Users

The major improvements derived from the usability evaluation of the end users are provided below. Generally, end users suggested improvements for data handling, and improvements with similar contents were integrated into 1 single category.

Data Input

> *I wish there were various ways to enter the result values. If I have to enter each value through the keypad, I think this will be a barrier for me to perform data entry.* [End user 10]

> *I would like to add a feature that can record the location in which I conducted the measurement. In my case, I tend to measure and record blood pressure and blood glucose in various places, such as my home, workplace, and the hospital.* [End user 2]

Data Output

> *It was difficult to concomitantly check the trends and values when there were separate lists and graphs, like in other existing health care apps. I wish I could see graphs and lists together on one screen.* [End user 3]

> *It should be possible to compare, at a glance, my current results with past medical checkup results. If you need to separately check the results of the medical checkup for every year, like now, it becomes*

XSL•FO
RenderX

inconvenient to check the trend of items that I want to carefully examine. Also, I wish I could see the categories of values of specific measurements according to a reference value. [End user 8]

It was good to be able to check the dosage guide and precautions [about a drug] in the prescription history. Can you not add the image of the drug? [End user 4]

Data Sharing

Can I not select the range (item and date, etc.) of information that I wish to share? I agree to share data for continuous health care services, but there are specific data that I do not want to share. [End user 1]

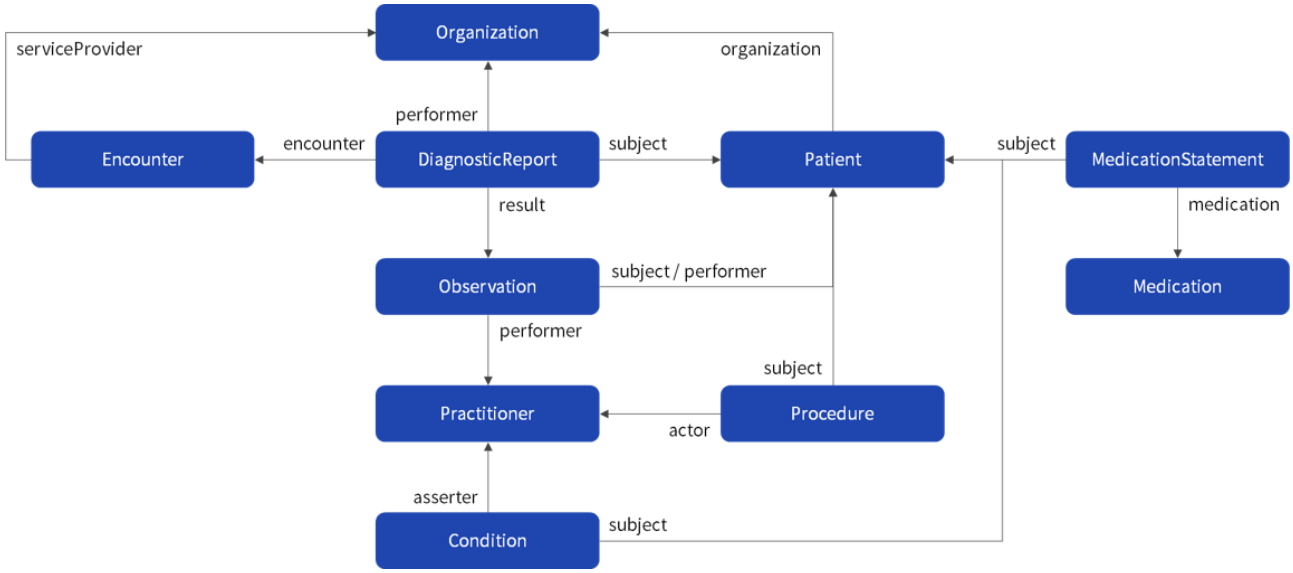
Data Security

Do you have any plans to add security features to the app? Even if the smartphone has a lock feature, it seems that a second authentication feature (fingerprint and password, etc.) is required to protect the sensitive personal information in the PHR app. [End user 2]

PHR Modeling

Among the FHIR resources, the structured PHRs of workers are shown in Figure 4. The Patient resource could be used to relay all information about patients and their surroundings, although this study focused only on representing workers' personal information. The Organization resource represented information from not only the workplace but also all other organizations used by workers, such as hospitals, pharmacies, and examination centers, collected through health care data sets. The DiagnosticReport resource could be used to describe a doctor's opinion based on information about a specific medical service and data measured in that medical service, although this study focused only on describing types of medical checkup and a doctor's opinion about the checkup. MedicationStatement and Medication resources were used for describing the prescription history, and the Procedure resources for relaying medical history and consultation records. Workers' PHR based on these resources were included in the Bundle resource and processed as a set when FHIR services interacted.

Figure 4. FHIR resource diagram of workers' personal health record.



Among the workers' PHR items, items requiring mapping comprised 40 general medical checkups, 289 special medical checkups, and 18 lifelogs (Table 7). General medical checkups are conducted for the early detection/prevention of diseases in workers, their dependents, and local subscribers. The types of examination for general medical checkups comprised general medical tests, oral tests, position tests (eg, for height, weight, obesity, and blood pressure), chest radiation, urinalysis, and blood examinations, etc., and the examination items differed by sex and age of the worker.

Special medical checkups are conducted to prevent occupational diseases and manage the health of workers engaged in jobs that expose them to harmful factors. Because there is a standardized test for each of the 179 harmful factors regulated by Korea's Occupational Safety and Health Act (eg, *N*, *N*-dimethylacetamide, benzene, acrylonitrile, vinyl chloride, dust), the test items for special medical checkups differed by work environment of the workers.

The lifelogs comprised items generated in daily life (eg, the number of steps and exercise) and about lifestyle (eg, the amount of drinking and smoking). As a result of the mapping, 347 items, except for 41, were mapped with the concepts of SNOMED-CT and LOINC.

XSL•FO
RenderX

Table 7. Mapping results of workers' personal health record items.

Section	Count	Mapping		Nonmapping
		SNOMED-CT[a]	LOINC[b]	
General medical checkups	40	40	35	—
Special medical checkups	289	208	234	41
Lifelogs	18	18	11	—
Total	347	266	280	41

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

[b]LOINC: Logical Observation Identifiers Names and Codes.

Most items that served as diagnostic tests (eg, general and special medical checkups) could be mapped with the concept of LOINC, although those that did not serve as diagnostic tests could not be mapped. Items that needed to be described in words rather than numbers (ie, doctor's opinion, medical history, occupational history) were mapped with the concept of SNOMED-CT and expressed as precoordinated, and items that needed to be partially specified were expressed as postcoordinated. Nonmapping items required specificity because they were ambiguous. For instance, the leukocyte percentage item, which was included in the hematopoietic classification, exists in various LOINC concepts (770-8, 35332-6 19023-1, 736-9, 42250-1, 5905-5, 713-8, 706-2) depending on the type of leukocyte. To summarize, when generating FHIR-based PHR, the concept of SNOMED-CT was used for lifelogs items, and the LOINC as a priority for general and special medical checkup items.

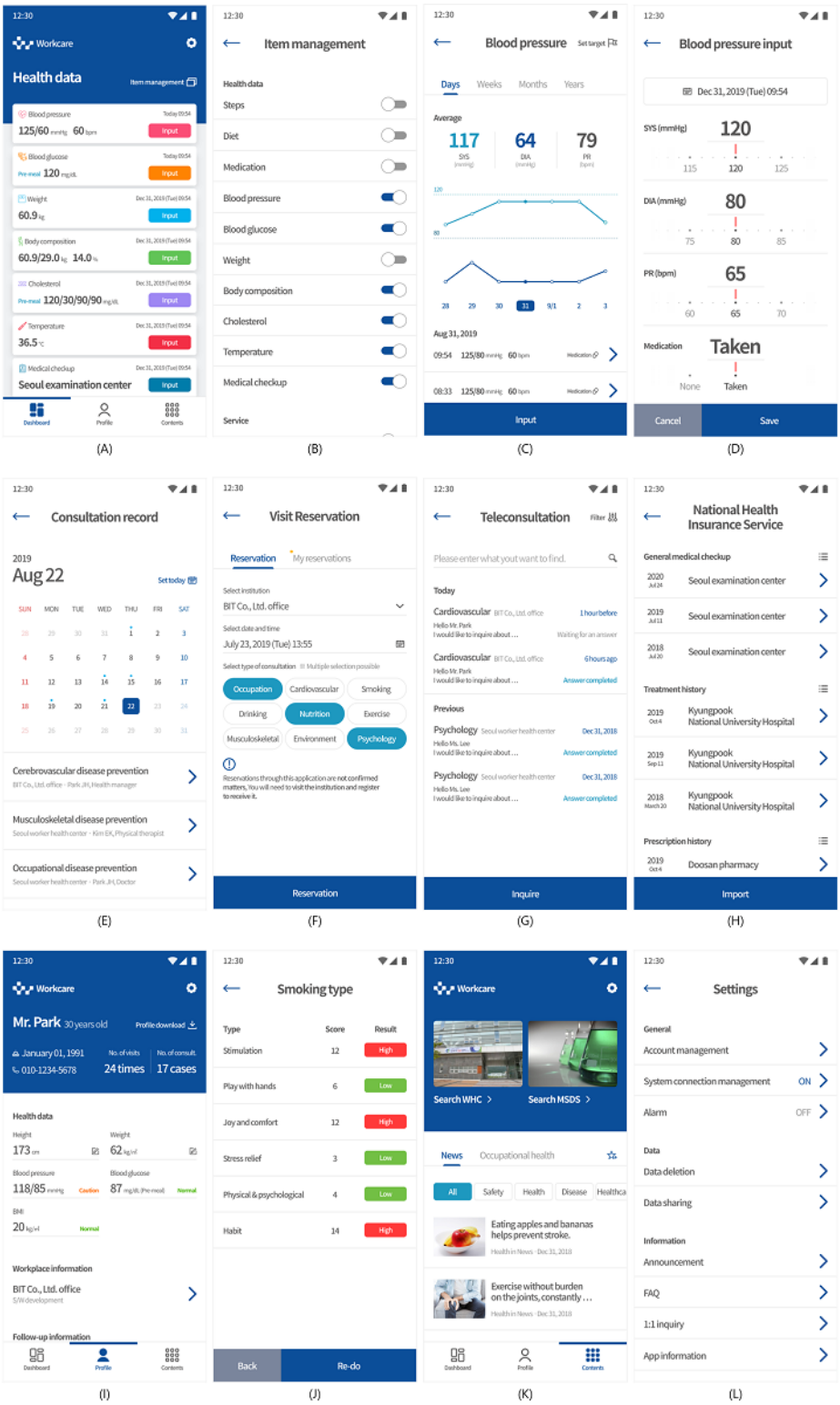
Final Developed Prototype App

Workcare provides users with PHR management according to the collection of data on workers' lifelogs, vital signs, medical checkup results, health care services (eg, reservation and teleconsultation), occupational safety and health information (eg, material safety data sheet search), and a health questionnaire. Users can access these features through more than 200 screens, and the app has an intuitive navigation system that minimizes the number of actions that users need to perform for accessing the desired content. The configuration of the screen was made in a way that frequently used features (eg, dashboards, profiles, and specific content) are placed on the bottom tab, with each screen being placed on its appropriate tab according to feature type.

The Dashboard tab (Figure 5A) provides the user with the main features for PHR management and health care services. The row in the Dashboard tab outputs the status of each feature, and frequently used features can be moved to the corresponding screen by clicking a button. For instance, the blood pressure row describes the latest measured value, and date and time at which it was measured; the user can also click the blood pressure row to move to the blood pressure screen (Figure 5C), or even click the input button to go to the blood pressure input screen (Figure 5D). Users can also enter the management screens for the number of steps, diet, medication, blood glucose, weight, body composition, cholesterol, body temperature, and general medical checkup results, or even view, through the API, values that were collected from and measured in various institutions. Users can also manage health-related tasks provided by API-linked institutions, such as consultation record (Figure 5E), visit reservation (Figure 5F), teleconsultation (Figure 5G), and data of the National Health Insurance Service (Figure 5H). The row within the Dashboard tab (Figure 5A) can select the order of items and the decision on whether to display them on the item management screen (Figure 5B) can be made by clicking the item management label at the top right.

XSL•FO
RenderX

Figure 5. Screenshots of different functions in the worker-centered personal health record app.



Regarding health data (eg, blood pressure and glucose), the screens were configured in a pattern similar to that of the dashboard. Typically, the blood pressure screen (Figure 5C) allows users to check blood pressure information (average, graph, and list) according to day, week, month, and year through the upper tab; in this screen, users can manipulate the graph by swiping left and right, with the measured value in the lower list and the average value at the top being updated according to the selected x-axis (ie, day, week, month, and year) in the graph. From the blood pressure input screen, users can directly input blood pressure data (Figure 5D) by clicking the input button, or even automatically enter measured values from a blood pressure device that has been paired with the app. The consultation record screen (Figure 5E) allows users to check health consultation records for visits to various health institutions (eg, the workers' health center). The visit reservation screen (Figure 5F) allows users to reserve a consultation in a specific institution; in the Reservation tab, users select the

XSL•FO
RenderX

institution they wish to visit, and the date, time, and type of the consultation. In the My Reservations tab, users can check information about the reservation, cancel it, or call the institution that made the reservation. The teleconsultation screen (Figure 5G) allows users to check responses from institutions after making an inquiry about health consultations; after reviewing the PHR shared through the FHIR service with medical personnel in the institution, users can also check the message sent to an institution. After completing certification in the login screen, users can collect and check the health care data present in public institutions; in other words, after completing certification of security and logging in the login screen, the National Health Insurance Service screen (Figure 5H) becomes available to users, who can then save their medical treatment history, prescription history, medical checkup results, and medical institution information by clicking the import button. Saved data can be viewed in detail on the screen by clicking a row.

The Profile tab allows users to check the main information of the user who is logged in (Figure 5I). In the upper area, basic information (eg, users' name, date of birth, and phone number) is described, with the health information of the user being output below the basic information in the upper area. By clicking on the Blood pressure, Blood glucose, and Body mass labels, users can check the share of the measured values according to a reference value through a graph that appears on the screen. By clicking on the workplace information row, users are moved to the screen that outputs information related to the workplace to which users belong; by clicking on the follow-up information row, users are moved to a screen that outputs the doctor's opinion about the results of the medical checkup. Based on the collected data and on the health questionnaire, users can self-evaluate their risk of cerebral heart disease, risk factors of cerebral cardiovascular disease, cerebral cardiovascular disease occurrence probability, and body age. Users can undergo health questionnaires on the smoking type (Figure 5J), nicotine dependence, job stress, psychological stress, and check the trend of the results.

The Contents tab (Figure 5K) provides users with information on occupational safety and health. In the upper area, images are arranged in a way to allow users to search for workers' health centers and material safety data sheets. The lower area outputs a list containing useful news and information on occupational safety and health. By clicking on the workers' health center search, users can check the locations, phone numbers, and home pages of 23 workers' health centers nationwide. The material safety data sheet provides detailed information on 16 categories, including chemical hazards, first aid measures, countermeasures in case of chemical exposure, and toxicity information; users can click a star icon to select a topic they wish to be displayed in the favorites screen.

The settings screen (Figure 5L) provides users with the main features for configuring the app environment. The account management row allows users to select whether they want to automatically log-in, change password, log out, or cancel their membership. The system connection management row displays a list of systems that have requested access to users' resources through the FHIR service, and users can add or delete these connections. The alarm row allows users to configure the app to produce push messages for major events, such as reservations, health counseling appointments, and goal achievements. The data deletion option allows users to delete all their data (after self-certification), while the data sharing option allows for uploading and synchronizing users' PHR according to the selected item and date. Finally, users can check important information necessary for service use through announcements, frequently asked questions, 1:1 inquiry, and app information.

Discussion

This study aimed to develop a PHR app that can provide worker-centered interconnected PHR services to support workplace health promotion by using health care standards, cloud services, and national health care data sets to solve known major challenges of PHR (ie, interoperability, security and privacy, and data quality), and by applying the HCD methodology to design an app based on users' perspectives.

We designed a service that integrates workers' health information that is scattered across various sources, and manages PHR through FHIR services; we used national health care data sets to ensure data entry, update, and quality. In 2017, the Republic of Korea revised the Act on Providing and Utilizing Public Data to guarantee the public's right to know about and access public data, as well as to ensure that most institutions provide data sets to the public. Accordingly, the National Health Insurance Service, while operating the national health insurance system, built a database comprising information on medical treatment history, prescription history, medical checkup results, and medical institution information; this database allows Korean citizens to check their data through self-certification. To prevent occupational diseases in workers, medical personnel need data on patients' treatment and prescription history, medical checkup results, and workers' PHR, as such thorough data can support medical personnel's decision making. Knowing the inherent problems of PHR (ie, regarding data input, update, and quality), we endeavored to acquire high-quality data that are managed by the Korean government through an interface method with institutions related to the management of workers' PHR. Nonetheless, the type of data that are measured by workers' visits to health institutions (eg, medical checkup results) has limitations regarding the identification of workers' health status at specific periods. Therefore, the PHR app we developed allows workers to measure and store PGHD through various devices, as well as to include these data in workers' PHR, so that medical personnel can identify workers' status even during periods when they will not or cannot visit a health institution.

Interconnected PHR is the ideal implementation of PHR, but the literature reports hindrances in standardizing the format and terminology for PHR information exchange. Previous studies on PHR have been conducted, but they differ from our study in several ways. First, previous studies [34,37] using document standards (eg, CCR and CDA) treated PHR as a single document; therefore, in previous studies the entire document must be updated when updating a single item. By contrast, PHR using FHIR, such as the one we used, does not incur such problems; items can be updated separately because they are

XSL•FO
RenderX

managed in a server by resource unit. Second, previous studies [35,37] that developed tethered PHR are dependent on specific electronic medical records (EMRs) and EHR; our app is not dependent on a specific system because the information is collected from various sources and is integrated and managed in the FHIR service according to the users' will. Besides, the users can have complete management authority over their health information. Third, some previous studies [38,39] used FHIR for managing PHR, but did not address privacy, security, and authentication issues. Our study, notwithstanding, developed an app that requested user authentication, confirmation, and authorization to access health resources through OAuth 2.0, also applying SSL, ARIA256, and SHA256 to solve privacy and security issues.

To ensure user convenience and usability, we designed the PHR app while considering the users' perspective through the HCD methodology. Previous studies [30,62-65] have shown that users expect PHR apps to assist in their health management by providing user-friendly and patient-centered features. Hence, this study considered data items, features, and interfaces that are suitable for user profiles through both quantitative and qualitative data analysis. Data items comprised lifelogs (eg, number of steps, diet, medication), health data (eg, blood pressure, blood glucose, and weight), medical checkup data (ie, general and special medical checkup results), and treatment and prescription history data. According to a systematic review of the literature by Roehrs et al [30], the common data items in PHR were allergy, vaccination, test results, and drugs, with little data on vital signs. Originally, we included allergy and vaccination items in the questionnaire of this study, but they were excluded through consultation with a health care professional group; this exclusion occurred because these items were considered less important than other items for occupational health.

Accordingly, we were able to derive various improvements to the app by conducting usability evaluations with both a health care professional group and an end user group. Regarding older adult workers, we added a feature to manage and show medical checkup results as an image; based on the opinion of the health care professional group, 1 out of 3 end users aged over 50 years are likely to not have a certificate on their smartphone, and thus, they would not be able to save the medical checkup results. Thus, amid the improvements to our prototype, we developed a feature that allowed users to directly input medical checkup results, capture a picture of the results through a smartphone camera, and save the picture. Moreover, we improved the diet management feature of the app using the integrated database; albeit the end users confirmed the need for information on dietary preferences through the survey results, the health care professional group did not confirm this addition because they were concerned about the lack of a unified system for the type of food, calories, and nutritional contents. We used the food nutrition ingredient database provided by the Ministry of Food and Drug Safety to solve the concerns of the health care professional group. The inconvenience of data recording about food, remarked by the health care professional group as another concern, was also dealt with by developing a feature to include frequently searched food (My Food) and image add-ons.

We developed a PHR app that can support workers' self-health management. Through the app, workers can collect and monitor their health information through the Dashboard tab, schedule a visit to a linked institution, or receive teleconsultations. The data items of this study were similar to those of a previous study [46], but some items (eg, water, alcohol, and smoking) were not included in the app. These items showed a low preference in the survey results of our study, and through the usability evaluation interview, we confirmed that users deemed the recording of frequent daily behaviors (eg, water and alcohol intake and smoking) as difficult. Therefore, we saved data on users' lifestyles through the inclusion of a health questionnaire for evaluating the risk of cerebrovascular disease, not through specific data collection items. The Profile tab allows users to check the status of their measured value according to a reference value, or even to check their cerebrovascular disease evaluation and body age based on the collected data. Zhou et al [42] did not support the analysis of data input by users; thus, this feature may have a limitation, in that the status of the measured values cannot be checked. In this study, we used the reference value of the structured workers' PHR item to determine whether the measured value of each item is normal. Nevertheless, based on the guideline [66] of the Korea Occupational Safety and Health Agency, it is possible to probabilistically predict the development of cerebrovascular disease by analyzing users' stored data on lifestyle and medical checkup results.

Recent changes in the social environment caused by COVID-19 have had a great impact on the distribution and production industries. The explosive increase in the volume and sorting of parcel deliveries owing to the COVID-19 pandemic led to the overwork of parcel workers, triggering an opportunity for the government and companies to review the state of workers' health improvement in the workplace. However, after the establishment of the employee assistance program provided by the Ministry of Employment and Labor in 2007, Korean workers have been provided with limited offline consultation opportunities; most employee assistance programs in Korea focus on mental health care, while research and investment in workers' health care services using technology have been insufficient. This study was, to the best of our knowledge, the first to develop a PHR app suitable for occupational health in Korea. Our PHR app can contribute to workers' personal health management by improving accessibility to their data and enabling the collection and management of their health information held by various institutions in one place. Registered users can continue to receive occupational health services by accessing and viewing their PHR at other institutions that comply with standards, even if they leave the workplace. This lays the foundation for ultimate workplace health promotion.

Most previous studies have focused on developing a PHR app for patients and older adults, while few researchers have endeavored to develop a PHR app for workers and support workplace health promotion. Thus, this study is meaningful in that it developed a worker-centered PHR app for workplace health promotion; however, it also had limitations. We attempted to integrate workers' health information that was scattered across various sources, but did not include data from hospitals. In order to activate worker-centered data exchange, hospital participation

XSL•FO
RenderX

is essential, and data standard issues must be resolved for each hospital. Even though Korea's EMR introduction rate is over 90%, it is difficult to utilize these data due to low standardization levels. The PHR app we developed enables information exchange between systems that comply with the standard through the FHIR service; however, a medical infrastructure that can guarantee continuity of treatment to patients is currently being developed in Korea. Since 2018, the national project for enabling the exchange of medical information between hospitals has been expanded with an EMR certification system (a system to verify national standards and conformity for EMR has been implemented in June 2020). Despite these advances, the possibility of integrating data from EMRs of hospitals visited by workers was still limited at the time of this study. However, given that EMRs include relevant health-related data (eg, vital signs, drugs, allergies, test results, and radiographic images), we believe that linkage between EMRs is necessary to ensure the provision of a wider number of services for users. Future studies are warranted to confirm the exchange of workers' medical information through the linkage between systems that have received the EMR certification system in Korea, and design a PHR app for workers that includes EMR data. Accordingly, future research may expand the service range of Workcare by linking it with the cloud EMR of BIT Computer Co. Ltd., to which the lead author (HP) is affiliated. Further, to confirm the clinical effectiveness of PHR services in the workplace, case–control and prospective studies will be conducted, and studies to analyze the satisfaction of workers and medical personnel with PHR services will also be conducted.

Acknowledgments

This work was supported by the Creative Industrial Technology Development Program (20002708, development and commercialization of the personalized health care service for employees based on the PHR platform), which is funded by the Ministry of Trade, Industry and Energy in Korea. We thank the Korea Occupational Safety & Health Agency, the Korean Association of Occupational Health Nurses, and the 21 workers' health centers nationwide for their help in data collection and the usability evaluation.

Conflicts of Interest

None declared.

References

1. Barron G. Going global with health and wellbeing analytics. Strategic HR Review 2012 Nov 23;12(1):5-9. [doi: 10.1108/14754391311282423]
2. Burton J. WHO healthy workplace framework and model: background and supporting literature and practices. Geneva, Switzerland: World Health Organization; 2010. URL: https://www.who.int/occupational_health/healthy_workplace_framework.pdf [accessed 2021-03-18]
3. Kim NJ. Relation between employees' life patterns and health conditions. Korean J Health Edu Promot 2007 Jun 30;24(2):75.
4. Parry T, Sherman B. Workforce Health--The Transition From Cost to Outcomes to Business Performance. Benefits Q 2015;31(1):32-38. [Medline: 26540941]
5. Winkler J. Reconsidering employer-sponsored health care: four paths to long-term strategic change. Benefits Q 2013;29(2):8-15. [Medline: 23943950]
6. Mudge-Riley M, McCarthy M, Persichetti T. Incorporating wellness into employee benefit strategies--why it makes sense. Benefits Q 2013;29(4):30-34. [Medline: 24730097]
7. Healthy People 2010: Understanding and improving health. Washington, DC, USA: US Government Printing Office Press; Oct 2011.
8. Federal Institute for Occupational Safety and Health. Luxembourg Declaration on Workplace health promotion in the European Union. 1997. URL: https://www.enwhp.org/resources/toolip/doc/2018/05/04/luxembourg_declaration.pdf [accessed 2021-03-18]
9. Dickson-Swift V, Fox C, Marshall K, Welch N, Willis J. What really improves employee health and wellbeing: findings from regional Australian workplaces. Int J Workplace Health Manag 2014 Sep 02;7(3):138-155. [doi: 10.1108/IJWHM-10-2012-0026]
10. Ljungblad C, Granström F, Dellve L, Åkerlind I. Workplace health promotion and working conditions as determinants of employee health. Intl J of Workplace Health Mgt 2014 Jun 03;7(2):89-104. [doi: 10.1108/IJWHM-02-2013-0003]
11. Robroek SJ, van Lenthe FJ, van Empelen P, Burdorf A. Determinants of participation in worksite health promotion programmes: a systematic review. Int J Behav Nutr Phys Act 2009 May 20;6:26 [FREE Full text] [doi: 10.1186/1479-5868-6-26] [Medline: 19457246]
12. Bull SS, Gillette C, Glasgow RE, Estabrooks P. Work site health promotion research: to what extent can we generalize the results and what is needed to translate research to practice? Health Educ Behav 2003 Oct 30;30(5):537-549. [doi: 10.1177/1090198103254340] [Medline: 14582596]

XSL•FO
RenderX

13. Griffiths F, Lindenmeyer A, Powell J, Lowe P, Thorogood M. Why are health care interventions delivered over the internet? A systematic review of the published literature. J Med Internet Res 2006 Jun 23;8(2):e10 [FREE Full text] [doi: 10.2196/jmir.8.2.e10] [Medline: 16867965]
14. Cook RF, Billings DW, Hersch RK, Back AS, Hendrickson A. A field test of a web-based workplace health promotion program to improve dietary practices, reduce stress, and increase physical activity: randomized controlled trial. J Med Internet Res 2007 Jun 19;9(2):e17 [FREE Full text] [doi: 10.2196/jmir.9.2.e17] [Medline: 17581811]
15. Hasson H, Brown C, Hasson D. Factors associated with high use of a workplace web-based stress management program in a randomized controlled intervention study. Health Educ Res 2010 Aug;25(4):596-607. [doi: 10.1093/her/cyq005] [Medline: 20150531]
16. Agarwal R, Anderson C, Zarate J, Ward C. If we offer it, will they accept? Factors affecting patient use intentions of personal health records and secure messaging. J Med Internet Res 2013 Feb 26;15(2):e43 [FREE Full text] [doi: 10.2196/jmir.2243] [Medline: 23470453]
17. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 2006 Mar;13(2):121-126 [FREE Full text] [doi: 10.1197/jamia.M2025] [Medline: 16357345]
18. Huba N, Zhang Y. Designing patient-centered personal health records (PHRs): health care professionals' perspective on patient-generated data. J Med Syst 2012 Dec 30;36(6):3893-3905. [doi: 10.1007/s10916-012-9861-z] [Medline: 22644130]
19. Reti SR, Feldman HJ, Safran C. Governance for personal health records. J Am Med Inform Assoc 2009;16(1):14-17 [FREE Full text] [doi: 10.1197/jamia.M2854] [Medline: 18952939]
20. Kaelber D, Pan E. The value of personal health record (PHR) systems. AMIA Annu Symp Proc 2008 Nov 06:343-347 [FREE Full text] [Medline: 18999276]
21. Burkhard RJ, Schooley B, Dawson J, Horan TA. Information Systems and Healthcare XXXVII: When Your Employer Provides Your Personal Health Record—Exploring Employee Perceptions of an Employer-Sponsored PHR System. CAIS 2010;27(19):323-338. [doi: 10.17705/1cais.02719]
22. Dawson J, Schooley B, Tulu B. A real world perspective: employee perspectives of employer sponsored personal health record (PHR) systems. : IEEE; 2009 Jan 20 Presented at: 2009 42nd Hawaii International Conference on System Sciences; January 5–8, 2009; Waikoloa, HI, USA. [doi: 10.1109/hicss.2009.34]
23. Smolij K, Dun K. Patient health information management: searching for the right model. Perspect Health Inf Manag 2006 Dec 12;3:10 [FREE Full text] [Medline: 18066368]
24. Vezyridis P, Timmons S. On the adoption of personal health records: some problematic issues for patient empowerment. Ethics Inf Technol 2015 Jun 26;17(2):113-124. [doi: 10.1007/s10676-015-9365-x]
25. Galvin RS, Delbanco S. Between a rock and a hard place: understanding the employer mind-set. Health Aff (Millwood) 2006;25(6):1548-1555. [doi: 10.1377/hlthaff.25.6.1548] [Medline: 17102179]
26. Moore J. Employers Taking Long-View Look to PHRs.: Chilmark Research LLC; 2008 Apr 07. URL: http://www.chilmarkresearch.com/2008/04/07/employers_adoption_phrs/ [accessed 2021-03-18]
27. Undem T. Consumers and Health Information Technology: A National Survey.: California HealthCare Foundation; 2010 Apr 13. URL: https://www.chcf.org/publication/consumers-and-health-information-technology-a-national-survey/ [accessed 2021-03-18]
28. Pushpangadan S, Seckman C. Consumer perspective on personal health records: A review of the literature. In: Online J Nurs Inform. Chicago, USA: Healthcare Information and Management Systems Society; Feb 2015.
29. Weinert C, Cudney S. My Health Companion©: A Low-Tech Personal Health Record Can Be an Essential Tool for Maintaining Health. OJRNHC 2012 May;12(1):3-15. [doi: 10.14574/ojrnhc.v12i1.36]
30. Roehrs A, da Costa CA, Righi RDR, de Oliveira KSF. Personal Health Records: A Systematic Literature Review. J Med Internet Res 2017 Jan 06;19(1):e13 [FREE Full text] [doi: 10.2196/jmir.5876] [Medline: 28062391]
31. Thompson MJ, Reilly JD, Valdez RS. Work system barriers to patient, provider, and caregiver use of personal health records: A systematic review. Appl Ergon 2016 May;54:218-242. [doi: 10.1016/j.apergo.2015.10.010] [Medline: 26851482]
32. Alyami M, Song Y. Removing barriers in using personal health record systems. In: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science.: IEEE; 2016 Aug 25 Presented at: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science; June 26–29, 2016; Okayama, Japan. [doi: 10.1109/icis.2016.7550810]
33. Carrión Señor I, Fernández-Alemán JL, Toval A. Are personal health records safe? A review of free web-accessible personal health record privacy policies. J Med Internet Res 2012 Aug 23;14(4):e114 [FREE Full text] [doi: 10.2196/jmir.1904] [Medline: 22917868]
34. Simon S, Anbananthen K, Lee S. A ubiquitous personal health record (uPHR) framework. : Atlantis Press; 2013 Aug Presented at: Proceedings of the 2013 International Conference on Advanced Computer Science and Electronics Information; July 25–26, 2013; Beijing, China. [doi: 10.2991/icacsei.2013.105]
35. Marceglia S, Fontelo P, Rossi E, Ackerman MJ. A Standards-Based Architecture Proposal for Integrating Patient mHealth Apps to Electronic Health Record Systems. Appl Clin Inform 2015;6(3):488-505 [FREE Full text] [doi: 10.4338/ACI-2014-12-RA-0115] [Medline: 26448794]

XSL•FO
RenderX

36. Plastiras P, O'Sullivan DM. Combining Ontologies and Open Standards to Derive a Middle Layer Information Model for Interoperability of Personal and Electronic Health Records. J Med Syst 2017 Oct 28;41(12):195. [doi: 10.1007/s10916-017-0838-9] [Medline: 29081012]
37. Li J. A service-oriented approach to interoperable and secure personal health record systems. : IEEE; 2017 Jun 08 Presented at: 11th IEEE Symposium on Service-Oriented System Engineering; April 6–9, 2017; San Francisco, CA, USA. [doi: 10.1109/sose.2017.20]
38. Hong J, Morris P, Seo J. Interconnected personal health record ecosystem using IoT cloud platform and HL7 FHIR. : IEEE; 2017 Sep 14 Presented at: 2017 IEEE International Conference on Healthcare Informatics; August 23–26, 2017; Park City, UT, USA. [doi: 10.1109/ichi.2017.82]
39. Saripalle R, Runyan C, Russell M. Using HL7 FHIR to achieve interoperability in patient health record. J Biomed Inform 2019 Jun;94:103188 [FREE Full text] [doi: 10.1016/j.jbi.2019.103188] [Medline: 31063828]
40. The Open Source FHIR API for Java.: HAPI URL: https://hapifhir.io/ [accessed 2021-03-18]
41. Farinango C, Benavides J, Cerón JD, López DM, Álvarez RE. Human-centered design of a personal health record system for metabolic syndrome management based on the ISO 9241-210:2010 standard. J Multidiscip Healthc 2018;11:21-37 [FREE Full text] [doi: 10.2147/JMDH.S150976] [Medline: 29386903]
42. Zhou L, DeAlmeida D, Parmanto B. Applying a User-Centered Approach to Building a Mobile Personal Health Record App: Development and Usability Study. JMIR Mhealth Uhealth 2019 Jul 05;7(7):e13194 [FREE Full text] [doi: 10.2196/13194] [Medline: 31278732]
43. Massoudi BL, Olmsted MG, Zhang Y, Carpenter RA, Barlow CE, Huber R. A web-based intervention to support increased physical activity among at-risk adults. J Biomed Inform 2010 Oct;43(5 Suppl):S41-S45 [FREE Full text] [doi: 10.1016/j.jbi.2010.07.012] [Medline: 20696275]
44. Marchak JG, Cherven B, Williamson Lewis R, Edwards P, Meacham LR, Palgon M, et al. User-centered design and enhancement of an electronic personal health record to support survivors of pediatric cancers. Support Care Cancer 2020 Aug;28(8):3905-3914 [FREE Full text] [doi: 10.1007/s00520-019-05199-w] [Medline: 31853699]
45. Fernando M, Sahama T, Fidge C, Hewagamage K. Personal health records as sources of productivity evidence. : IEEE; 2016 Jul 14 Presented at: 2016 IEEE International Conference on Communications; May 22–27, 2016; Kuala Lumpur, Malaysia. [doi: 10.1109/icc.2016.7510830]
46. Fernando M, Fidge C, Sahama T. An overall health and well-being data model for employersponsored personal health records. : ACM; 2019 Jan 29 Presented at: Proceedings of the Australasian Computer Science Week Multiconference; January 29-31, 2019; Sydney, Australia p. 1-10. [doi: 10.1145/3290688.3290727]
47. Fernando M, Fidge C, Sahama T. Design guidelines for effective occupation-based personal health records. : ACM; 2020 Feb 04 Presented at: Proceedings of the Australasian Computer Science Week Multiconference; February 4-6, 2020; Canberra, Australia p. 1-10. [doi: 10.1145/3373017.3373042]
48. Incorporating occupational information in electronic health records: Letter report. Washington, DC, USA: National Academies Press; Sep 26, 2011.
49. Rajamani S, Chen E, Lindemann E, Aldekhyyel R, Wang Y, Melton G. Representation of occupational information across resources and validation of the occupational data for health model. J Am Med Inform Assoc 2018 Feb 01;25(2):197-205 [FREE Full text] [doi: 10.1093/jamia/ocx035] [Medline: 28444213]
50. HL7 FHIR Profile: Occupational Data for Health (ODH), Release 1.: HL7 International URL: http://hl7.org/fhir/us/odh/2018Sep/ [accessed 2021-03-18]
51. Introducing HL7 FHIR.: HL7 International URL: https://www.hl7.org/fhir/summary.html [accessed 2021-03-18]
52. Bender D, Sartipi K. HL7 FHIR: an agile and RESTful approach to health care information exchange. : IEEE; 2013 Oct 10 Presented at: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems; June 20–22, 2013; Porto, Portugal. [doi: 10.1109/CBMS.2013.6627810]
53. Fyfe J, Bender D, Edwards HK. Everest. SIGHIT Rec 2012 Mar;2(1):24-24. [doi: 10.1145/2180796.2180816]
54. Smith B, Ceusters W. HL7 RIM: an incoherent standard. Stud Health Technol Inform 2006;124:133-138. [Medline: 17108516]
55. Agarwal R, Khuntia J. Personal health information and the design of consumer health information technology: Background report. Rockville, US: Agency for Health care Research and Quality; 2009. URL: https://digital.ahrq.gov/sites/default/files/docs/citation/09-0075-EF.pdf [accessed 2021-03-18]
56. Katehakis DG, Kondylakis H, Koumakis L, Kouroubali A, Marias K. Integrated Care Solutions for the Citizen: Personal Health Record Functional Models to Support Interoperability. ejbi 2017;13(1):51-58. [doi: 10.24105/ejbi.2017.13.1.8]
57. Park HS, Kim KI, Soh JY, Hyun YH, Lee BE, Lee JH, et al. Development and Operation of a Video Teleconsultation System Using Integrated Medical Equipment Gateway: a National Project for Workers in Underserved Areas. J Med Syst 2020 Oct 01;44(11):194 [FREE Full text] [doi: 10.1007/s10916-020-01664-w] [Medline: 33006060]
58. Lewis JR. Psychometric evaluation of an after-scenario questionnaire for computer usability studies. SIGCHI Bull 1991 Jan 01;23(1):78-81. [doi: 10.1145/122672.122692]
59. Brooke J. SUS: a quick and dirty usability scale. In: Usability Evaluation In Industry. London: CRC Press; 1996.

XSL•FO
RenderX

60. PHR-S FM Personal Health Record System Functional Model. HL7 International. URL: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=88 [accessed 2021-07-26]
61. Lighting up healthcare data with FHIR: Announcing the Azure API for FHIR.: Microsoft Azure URL: https://azure.microsoft.com/ko-kr/blog/lighting-up-healthcare-data-with-fhir-announcing-the-azure-api-for-fhir/ [accessed 2021-03-18]
62. Kruse CS, Argueta DA, Lopez L, Nair A. Patient and provider attitudes toward the use of patient portals for the management of chronic disease: a systematic review. J Med Internet Res 2015 Feb 20;17(2):e40 [FREE Full text] [doi: 10.2196/jmir.3703] [Medline: 25707035]
63. Abramson E, Patel V, Edwards A, Kaushal R. Consumer perspectives on personal health records: a 4-community study. Am J Manag Care 2014 Apr;20(4):287-296 [FREE Full text] [Medline: 24884860]
64. Wen K, Kreps G, Zhu F, Miller S. Consumers' perceptions about and use of the internet for personal health records and health information exchange: analysis of the 2007 Health Information National Trends Survey. J Med Internet Res 2010 Dec 18;12(4):e73 [FREE Full text] [doi: 10.2196/jmir.1668] [Medline: 21169163]
65. Ronda MCM, Dijkhorst-Oei L, Rutten GEHM. Reasons and barriers for using a patient portal: survey among patients with diabetes mellitus. J Med Internet Res 2014 Nov 25;16(11):e263 [FREE Full text] [doi: 10.2196/jmir.3457] [Medline: 25424228]
66. Risk assessment and follow-up guidelines for prevention of brain and cardiovascular diseases at work. In: KOSHA GUIDE. Ulsan, Korea: KOSHA Press; Dec 2018.

Abbreviations

API: application programming interface
ASQ: After-Scenario Questionnaire
CCR: continuity of care record
CDA: clinical document architecture
DICOM: Digital Imaging and Communications in Medicine
EHR: electronic health record
EMR: electronic medical records
FGI: focus group interview
FHIR: Fast Healthcare Interoperability Resources
HAPI: HL7 Application Programming Interface
HCD: human-centered design
HL7: Health Level Seven
HTTPS: hypertext transfer protocol over secure socket layer
JSON: JavaScript Object Notation
LOINC: logical observation identifiers names and codes
PGHD: patient-generated health data
PHR: personal health record
SNOMED-CT: Systematized Nomenclature of Medicine–Clinical Terms
SSL: secure socket layer
SUS: System Usability Scale
UCD: user-centered design
WHO: World Health Organization

Edited by G Eysenbach; submitted 30.03.21; peer-reviewed by L Zhou, H Pratomo; comments to author 21.04.21; revised version received 09.06.21; accepted 23.06.21; published 04.08.21.

Please cite as:
Park HS, Kim KI, Chung HY, Jeong S, Soh JY, Hyun YH, Kim HS
A Worker-Centered Personal Health Record App for Workplace Health Promotion Using National Health Care Data Sets: Design and Development Study
JMIR Med Inform 2021;9(8):e29184
URL: https://medinform.jmir.org/2021/8/e29184
doi:10.2196/29184
PMID:34346894

©Hyun Sang Park, Kwang Il Kim, Ho-Young Chung, Sungmoon Jeong, Jae Young Soh, Young Ho Hyun, Hwa Sun Kim. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 04.08.2021. This is an open-access article

XSL•FO
RenderX

distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Review

The Unified Medical Language System at 30 Years and How It Is Used and Published: Systematic Review and Content Analysis

Xia Jing[1], MD, PhD

Department of Public Health Sciences, College of Behavioral, Social and Health Sciences, Clemson University, Clemson, SC, United States

Corresponding Author:
Xia Jing, MD, PhD
Department of Public Health Sciences
College of Behavioral, Social and Health Sciences
Clemson University
511 Edwards Hall
Clemson, SC, 29634
United States
Phone: 1 8646563347
Fax: 1 8646566227
Email: xjing@clemson.edu

Abstract

Background: The Unified Medical Language System (UMLS) has been a critical tool in biomedical and health informatics, and the year 2021 marks its 30th anniversary. The UMLS brings together many broadly used vocabularies and standards in the biomedical field to facilitate interoperability among different computer systems and applications.

Objective: Despite its longevity, there is no comprehensive publication analysis of the use of the UMLS. Thus, this review and analysis is conducted to provide an overview of the UMLS and its use in English-language peer-reviewed publications, with the objective of providing a comprehensive understanding of how the UMLS has been used in English-language peer-reviewed publications over the last 30 years.

Methods: PubMed, ACM Digital Library, and the Nursing & Allied Health Database were used to search for studies. The primary search strategy was as follows: UMLS was used as a Medical Subject Headings term or a keyword or appeared in the title or abstract. Only English-language publications were considered. The publications were screened first, then coded and categorized iteratively, following the grounded theory. The review process followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.

Results: A total of 943 publications were included in the final analysis. Moreover, 32 publications were categorized into 2 categories; hence the total number of publications before duplicates are removed is 975. After analysis and categorization of the publications, UMLS was found to be used in the following emerging themes or areas (the number of publications and their respective percentages are given in parentheses): natural language processing (230/975, 23.6%), information retrieval (125/975, 12.8%), terminology study (90/975, 9.2%), ontology and modeling (80/975, 8.2%), medical subdomains (76/975, 7.8%), other language studies (53/975, 5.4%), artificial intelligence tools and applications (46/975, 4.7%), patient care (35/975, 3.6%), data mining and knowledge discovery (25/975, 2.6%), medical education (20/975, 2.1%), degree-related theses (13/975, 1.3%), digital library (5/975, 0.5%), and the UMLS itself (150/975, 15.4%), as well as the UMLS for other purposes (27/975, 2.8%).

Conclusions: The UMLS has been used successfully in patient care, medical education, digital libraries, and software development, as originally planned, as well as in degree-related theses, the building of artificial intelligence tools, data mining and knowledge discovery, foundational work in methodology, and middle layers that may lead to advanced products. Natural language processing, the UMLS itself, and information retrieval are the 3 most common themes that emerged among the included publications. The results, although largely related to academia, demonstrate that UMLS achieves its intended uses successfully, in addition to achieving uses broadly beyond its original intentions.

(JMIR Med Inform 2021;9(8):e20675) doi:10.2196/20675

KEYWORDS

Unified Medical Language System; systematic literature analysis; biomedical informatics; health informatics

XSL•FO
RenderX

Introduction

Background

The Unified Medical Language System (UMLS) [1] is a critical resource in biomedical and health informatics. It was created and released by the National Library of Medicine, an institute of the National Institutes of Health (NIH). The first edition of UMLS Knowledge Sources was distributed in 1991 [1], although its conceptualization can be traced to 1986 [2]. Currently, there are three UMLS Knowledge Sources: Metathesaurus, Semantic Network, and SPECIALIST Lexicon and Lexical Tools. The Metathesaurus contains approximately 4.4 million concepts and 16 million unique concept names, which are from 218 source vocabularies in 25 languages worldwide (2021AA release). The Semantic Network provides consistent categorization for all concepts included in UMLS [3]. The SPECIALIST Lexicon and Lexical Tools provide large syntactic lexicon tools that have been used broadly in the biomedical and health fields to normalize strings and lexical variants.

UMLS brings together many broadly used vocabularies and standards in the biomedical field to facilitate interoperability and semantic understanding among different computer systems and software applications [4,5]. UMLS has been maintained and further developed by the National Library of Medicine over the past 30 years. In the initial publication, UMLS was intended to be used in four main areas: patient care, medical education, library service, and product development [1]. A comprehensive evaluation of the UMLS would be a large project; however, a close examination of the literature in the form of peer-reviewed publications can provide a perspective on how the UMLS is used in academia, which is the *rationale* for this literature review.

Objective

The year 2021 is the 30th anniversary of UMLS. Despite its longevity, there is no comprehensive publication analysis of UMLS. To call attention to the importance of UMLS and highlight its critical role in advancing biomedical informatics, health informatics, medicine, and health care, this systematic analysis was conducted to demonstrate how UMLS has been used, based on peer-reviewed publications in English over the past 30 years, which is the objective of this literature review.

Methods

Literature Search Sources and Strategies

Overview

PubMed, ACM Digital Library, and the Nursing & Allied Health Database were used for the search. The primary strategy was to search literature that either used UMLS as a MeSH (Medical Subject Headings) term or a keyword or had UMLS or *unified medical language system* in the title or abstract.

Search Strategy in PubMed on April 28, 2020

unified medical language system [MeSH term]

Search Strategy in ACM Digital Library on April 28, 2020

Searches were conducted within the ACM Guide to Computing Literature:

[Publication title: umls] OR [Publication title: *unified medical language system**] OR [Abstract: umls] OR [Abstract: *unified medical language system**]

The following journals were excluded because they are indexed in PubMed: *Journal of Biomedical Informatics*, *Artificial Intelligence in Medicine*, and *Bioinformatics*.

Search Strategy in the Nursing & Allied Health Database on April 28, 2020

Searches were conducted within peer-reviewed publications:

mesh (*unified medical language system*) OR ti(umls) OR ti(*unified medical language system*) OR ab(umls) OR ab (*unified medical language system*)

Literature Examination

Literature Examination Process

The literature examination process followed the grounded theory. The steps for the content analysis were as follows: all duplicate publications were removed before the literature examination. The exclusion criteria included the following: UMLS not mentioned in the abstract, abstract unavailable, or non-English publications.

The first step of the content analysis was to go over and then code (or index) each title and to record the repeated themes or topics. The second step was to go over each abstract one by one to code (or index) each abstract again, record the repeated themes or topics, and exclude the irrelevant publications. The third step was to organize the themes and group them according to their similarities. Subsequently, each publication was classified into the corresponding theme, and additional themes were created during the process.

The classification step was conducted iteratively. The first round began with obvious and repeated themes. Each publication was examined and, as appropriate, categorized by theme. I began with the relatively obvious themes, each of which had relatively fewer publications. The initial group of themes included artificial intelligence (AI) tools and applications, other language UMLS studies, medical education, patient care, medical subdomains, digital library, and degree-related theses. The publications were then classified, one by one, for the following themes: UMLS itself, information retrieval, terminology study, natural language processing (NLP), ontology and modeling, data mining, and knowledge discovery. The publications that fell outside of these themes during the coding (or indexing) process were classified last. The classification process stopped when all publications were classified into themes without the need for additional consideration. The themes were adjusted whenever needed during the iterative classification processes. The publications were then analyzed, categorized, synthesized iteratively, counted, and recorded into each category.

XSL•FO
RenderX

A word cloud picture (Multimedia Appendix 1) based on the titles included in this comprehensive literature analysis was generated by removing all commonly used words. The Pro Word Cloud function within Microsoft Word (Microsoft Corporation) was used to generate the word cloud picture.

Literature Classification Principles

The following principles were followed during classification: the primary principle is that when a publication is analyzed, the objectives of the publication, not the methods implemented, are the prioritized reasons for the categorization. The secondary principle is to maximize the possibility that a publication will stand out among the publications in each category; that is, if a publication has an approximately equal possibility to be classified into 2 categories, the one with fewer publications wins. The third principle is to give publications on applications and patient care a higher priority than methodology development or foundational studies, in general. The fourth principle is to classify a publication into the most specific category whenever possible. The rationale for following these principles is based on the literature review. Instead of providing a comprehensive evaluation of all aspects of the UMLS, I attempted to determine how the UMLS is used in the real world. I focused on its application as a critical factor. As the UMLS is found in medicine, patient care is a higher priority.

In addition, I used this opportunity to recognize my peers' contributions by maximizing the possibility of their publications standing out because only a small fraction of the work can be awarded a prize. These principles helped me to classify all the publications in a more consistent, clear, reproducible, and objective manner.

Literature Review Guideline

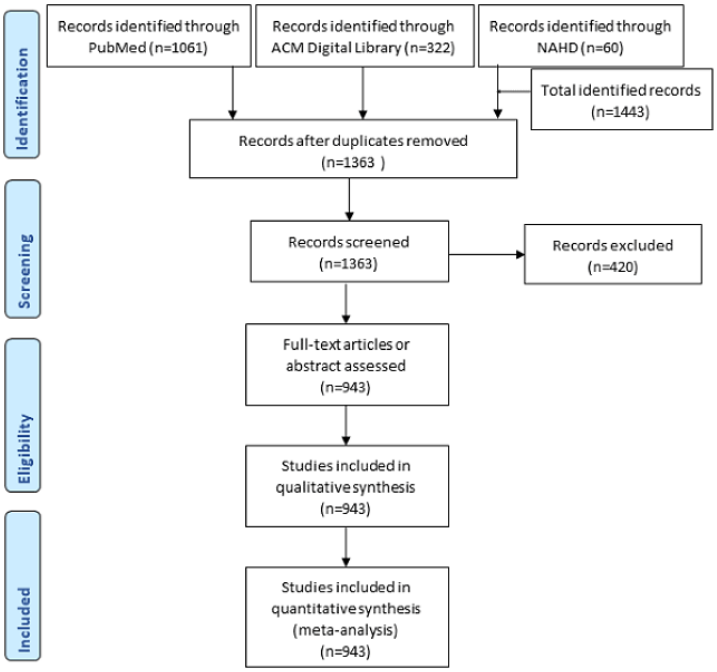
The systematic literature analysis protocol has not been registered. The data items used in this literature review including title, author, publication year, journal or conference proceeding, abstract, MeSH terms or keywords, PubMed ID (if available), full-text for some publications, and what was UMLS used for. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [6] were followed and most of the checklist items were included. The PRISMA checklist is provided in Multimedia Appendix 2.

Results

Overview

The search strategies yielded 1061 records in PubMed, 322 in the ACM Digital Library, and 60 in the Nursing & Allied Health Database. After removing the duplicates, records without abstracts, non-English records, and abstracts that did not mention UMLS, 943 records were retained for the final analysis. Figure 1 [6] shows detailed records of the literature search, screening, and analysis.

Figure 1. Flowchart of the literature search, screening, analysis, and its records. NAHD: Nursing & Allied Health Database.



Multimedia Appendix 3 shows the yearly number of the included UMLS publications over the last 30 years. Table 1 presents the themes that emerged and the corresponding number of publications for each category. This table provides an overview of the results of the systematic analysis. Multimedia Appendix 4 presents the major themes, topics, and corresponding publication counts.

XSL•FO
RenderX

Table 1. Results of the Unified Medical Language System systematic literature analysis: emerging themes, subtopics, and the number of publications in each category before and after removing duplicates.

Themes and subtopics	Publication counts (n=975), n (%)	After removing the duplicates (n=943), n (%)
Artificial intelligence tools and applications	46 (4.7)	46 (4.9)
Automatic annotation or interpretation	7 (15.2)	7 (15.2)
Automatic coding	7 (15.2)	7 (15.2)
Automatic summarization	15 (32.6)	15 (32.6)
Question-answering systems	10 (21.7)	10 (21.7)
Other intelligent tools	7 (15.2)	7 (15.2)
Data mining and knowledge discovery	25 (2.6)	25 (2.7)
Degree-related theses	13 (1.3)	8 (0.8)
Digital library	5 (0.5)	4 (0.4)
Information retrieval	125 (12.8)	119 (12.6)
Image retrieval	20 (16)	20 (16.8)
Indexing	33(26.4)	30 (25.2)
Information retrieval	34 (27.2)	32 (26.9)
Information retrieval system and search engine	8 (6.4)	8 (6.7)
Performance	13 (10.4)	12 (10.1)
Query	17 (13.6)	17 (14.3)
Medical education	20 (2.1)	19 (2)
Medical subdomains (34 subdomains)	76 (7.8)	76 (8.1)
NLP[a]	230 (23.6)	230 (24.4)
Abbreviation	11 (4.8)	11 (4.8)
Feature identification or extraction or phenotyping	4 (1.7)	4 (1.7)
Lexicon and/or inventory	7 (3)	7 (3)
Semantic	165 (71.7)	165 (71.7)
Concept recognition or extraction	42 (25.5)	42 (25.5)
Name entity recognition or extraction	18 (10.9)	18 (10.9)
Natural language, vocabulary, question generation	3 (1.8)	3 (1.8)
Natural language understanding	3 (1.8)	3 (1.8)
Relationship recognition or extraction	45 (27.3)	45 (27.3)
Semantic similarity, relatedness, or distance	20 (12.1)	20 (12.1)
Word sense disambiguation	34 (20.6)	34 (20.6)
Syntax	18 (7.8)	18 (7.8)
Parsing	5 (27.8)	5 (27.8)
Tagging	5 (27.8)	5 (27.8)
Terminology extraction	8 (44.4)	8 (44.4)
Text classification	10 (4.4)	10 (4.4)
Other NLP-related publications	15 (6.5)	15 (6.5)
Ontology and modeling	80 (8.2)	79 (8.4)
Classification or taxonomy	21 (26.3)	21 (26.6)
Modeling	18 (22.5)	17 (21.5)
Ontology	29 (36.3)	29 (36.7)
Representation	12 (15)	12 (15.2)

XSL•FO
RenderX

Themes and subtopics	Publication counts (n=975), n (%)	After removing the duplicates (n=943), n (%)
Other languages (10 languages)	53 (5.4)	47 (5)
Patient care	35 (3.6)	27 (2.9)
Terminology study	90 (9.2)	90 (9.5)
Comparison of terminologies	6 (6.7)	60 (6.7)
Construction of terminology or taxonomy	19 (21.1)	19 (21.1)
Harmonization	46 (51.1)	46 (51.1)
Interoperability	7 (7.8)	7 (7.8)
Quality assurance	7 (7.8)	7 (7.8)
Other publications of terminology	5 (5.6)	5 (5.6)
UMLS[b] itself	150 (15.4)	146 (15.5)
Applications or tools for UMLS	25 (16.7)	25 (17.1)
Auditing of UMLS	24 (16)	24 (16.4)
Components of UMLS or UMLS	78 (52)	76 (52.1)
Coverage of UMLS	23 (15.3)	21 (14.4)
UMLS for other purposes	27 (2.8)	27 (2.9)
Auditing	3 (11.1)	3 (11.1)
Consumer health	4 (14.8)	4 (14.8)
Integrated system or data	17 (63)	17 (63)
Other research use	3 (11.1)	3 (11.1)

[a]NLP: natural language processing.

[b]UMLS: Unified Medical Language System.

Themes, Subtopics, and Publications Under Each Category

After the included publications were examined carefully, the following themes emerged during analysis and synthesis.

UMLS Is Used in AI Tools and Applications

The UMLS has been used in developing AI tools and applications since 1994 [7] (publication; the actual work started many years ago). The AI tools include question-answering systems, automatic summarization, automatic coding, automatic annotation, and plagiarism detection. Question-answering systems focus on the medical domain. Some question-answering systems focus specifically on answering consumers' questions. Automatic summarization focuses mainly on summarizing medical literature, textbooks, and patient records. This category also includes methodology exploration. Multimedia Appendix 5 includes the 46 UMLS publications in this category.

I recognize that there is an overlap between AI tools and NLP. The criterion used concerned whether a publication focused on the final products. If so, it was classified into the AI tools and applications category; if a publication focused on the middle-layer methodology to enhance performance, it was classified into the NLP category.

Automatic translation can also be categorized into this theme; however, the publications were categorized on automatic translation into the other language UMLS studies category, using a more detailed description. Similarly, intelligent tutoring systems were classified into medical education instead of AI tools and applications. These categories should be cross-referenced accordingly.

UMLS-Based Data Mining and Knowledge Discovery

UMLS is used broadly as a critical tool in data mining and knowledge discovery in the biomedical field. However, there are large overlaps between this category and the subcategory under NLP, namely, relationship extraction. The following categorization criteria were implemented: if a publication could be dissected into a relationship (eg, drug-drug interaction, condition-treatment, and association rule mining) extraction, identification, or discovery, the publication was classified under the relationship extraction subcategory of NLP; otherwise, the publication was included in the data mining and knowledge discovery category. Multimedia Appendix 6 lists all 25 included UMLS publications related to data mining, knowledge discovery, data analysis, and text analysis.

UMLS in Degree-Related Theses

Notably, there are 13 doctoral theses [8-20] included from the ACM Digital Library that used the UMLS as a key component in conducting the research. I believe that it is very likely that there is greater use of the UMLS in doctoral or master theses that might not be captured through the title, abstract, or keywords. My own doctoral thesis used UMLS as a critical foundational tool to build a knowledge base; however, UMLS was not listed as a keyword.

XSL•FO
RenderX

UMLS for Digital Libraries

A digital library is another initial goal of UMLS. In this systematic literature analysis, 5 publications related to the UMLS and a digital library were identified. Of these, one publication used machine learning to process information extracted from a digital library, in which UMLS served as an information source [21]. In terms of a digital library, UMLS is also used for navigation purposes [22], for the semantic query [23], to improve the functions of the digital library [24], and to extract knowledge from a digital library [25]. There could be additional publications on the topic that do not necessarily use *digital library* as the key term.

UMLS in Information Retrieval

Since its inception, UMLS has been used to achieve and improve information retrieval. A total of 125 publications were identified in this theme, which is the third most active theme in this review. The subtopics of this emerging theme include image retrieval (eg, radiological images, pathological images, microscopic images, computed tomography scans, and electrocardiograms), indexing, information retrieval (including information needs), information retrieval systems, and search engines (eg, PubMed, MEDLINE, electronic health record systems, books, databases of texts, images, and sounds), performance or correct measures (including ranking), and query (from generic queries, query formulation, query expansion, and more accurate queries to evaluations). The information sources for retrieval purposes included documents, information within documents, metadata, scientific literature, and patient records. Multimedia Appendix 7 lists all 125 UMLS publications related to information retrieval.

UMLS in Medical Education

UMLS was planned for use in medical education [1,26-29]. Most of the publications in this category included curriculum mapping [30], continuing education [31,32], problem-based learning [33], tutoring systems [33-41], and educational resource development [31,32,42-44].

UMLS in Different Medical Subdomains

As the most comprehensive collection of medical terminologies, UMLS has been used in 34 medical subdomains in a variety of ways. The subdomains in which UMLS has been used include Alzheimer disease [45,46], anatomical structure [47-64], appendectomy [65], asthma [66,67], blood transfusion [68,69], breast biopsy [65], breast cancer [70,71], cardiovascular diseases [72-74], colorectal cancer [75,76], depression [77,78], dilated cardiomyopathies [79], epidemiology [80,81], falling injury risk assessment [82], HIV [83], hypertension [84-86], Kawasaki disease [87], liver cancer [88], liver diseases [89,90], lupus [91], neuropsychiatric disorders [92-94], occupational medicine [95,96], oncology [97,98], Parkinson disease [99], pneumonia [100], physical therapy [101], primary care [102-104], prostate cancer [105,106], rare diseases [107-111], respiratory tract infection [112], stroke thrombolysis [113], surveillance [114-116], traditional Chinese medicine [117,118], urology [119,120], and Zika virus [121]. There are significantly more publications about anatomy than about any other medical subdomain.

UMLS in NLP

UMLS is used as a critical component in NLP, the most active theme in the review, with 230 publications identified. The specific use of UMLS in this category includes abbreviation-related studies, feature identification, lexicon and inventory, semantic-related studies, syntax-related studies, text classification, and other NLP-related UMLS publications.

Semantic-related publications (165/230, 71.7%) included concept recognition and extraction, named entity recognition, natural language, vocabulary, question generation, natural language understanding, relationship recognition and extraction, semantic similarity or relevance or distance, and word sense disambiguation. Named entity recognition also included negation recognition. For concept recognition or extraction, the following groups were included: adverse drug event identification, contextual property identification, disorder recognition, and identification of treatment information. Relationship recognition and extraction included association recognition, medication-indication relationships, drug-drug interaction, and disease-manifestation relationships.

Syntax-related publications (18/230, 7.8%) included part-of-speech tagging, parsing, and terminology extraction.

Other NLP-related publications (47/230, 20.4%) included rule-based NLP, statistical NLP, corpus development, morphological similarity, word embedding, and stemming.

The source document types used in NLP are very rich and include discharge summaries, problem lists, clinical trial eligibility criteria, clinical trial protocols, clinical narrative notes, patient records, radiology reports, neuroradiology reports, pathology reports, histology reports, emergency department reports, surgical operative reports, medical progress notes, literature, social media, emails, and forum posts. Multimedia Appendix 8 presents a list of all 230 publications classified into the NLP category.

UMLS-Based Ontology and Modeling-Related Publications

UMLS is also a common tool used in ontology, classification, taxonomy, modeling, knowledge representation, and their associated studies. Although UMLS and terminology study are 2 existing categories, there are still some publications that cannot be categorized into either of these categories. If a publication can be included in a more specific subcategory, for example, a model of an information retrieval system, then it will be classified into the information retrieval system and search engine subcategory instead of the modeling subcategory. In this category, the publications were classified into corresponding subcategories only if the publication could not be included in any other category. Multimedia Appendix 9 presents a list of all 80 publications in this category.

UMLS English-Language Publications About Non-English Languages

There are efforts related to using UMLS in languages other than English, as well as multilingual studies. In this category, 10 additional languages and 53 publications were identified. Some publications are related to automatic translation, whereas others

XSL•FO
RenderX

are related to the coverage of an additional language of medical terms in addition to English. Languages other than English that relate to multilingual or cross-language uses of UMLS include Bulgarian [122], Dutch [123], French [63,124-145], German [76,146-149], Italian [150], Japanese [151-153], Korean [154-158], Portuguese [159], Spanish [160-162], and Swedish [146,163]. A total of 12 publications included more than two languages [146,159,163-172]. Clearly, there are more French-related UMLS publications than any other non-English language.

UMLS in Patient Care

One of the original goals of UMLS is to facilitate patient care directly or indirectly. As planned, UMLS has been used in patient care in many different ways, including the prediction of bariatric surgery outcomes, ensuring patient safety, development of a fall injury risk assessment instrument, patient outcome measurement, functional status measurement, clinical care quality assurance, computerized physician order entry, and clinical decision support systems. Multimedia Appendix 10 presents a list of all 35 publications in this category.

UMLS for Terminology Studies

As a critical tool, UMLS is used to conduct terminology studies. A total of 90 publications were classified into this category. The scope of the work includes a comparison of terminologies, construction of terminology, harmonization, interoperability, quality assurance, and other UMLS publications of terminology. UMLS is critical for achieving and advancing interoperability. The publications about the UMLS itself were classified into the UMLS category instead of under terminology studies. The roles of the UMLS in terminology studies include data sharing, aggregating data, harmonizing (including mapping among different terminologies), and vocabulary foundation. The publications on lexical mapping were classified into NLP. Multimedia Appendix 11 presents a list of all 90 UMLS publications on terminology studies.

Studies About the UMLS Itself

A total of 150 publications about the UMLS itself, which is the second most active theme after NLP, were identified. The scope of the publications ranged from auditing and enhancement of UMLS to the development of its own components, including Metathesaurus, SPECIALIST Lexicon and Lexical Tools, and Semantic Network, as well as its application tools MetaMap, MMTx, and SemRep. Furthermore, many efforts were related to increasing the coverage of UMLS in different subdomains, for example, in nursing, radiology, genetic disease, anatomy, and herbal supplements. In this category, the subtopics included applications or tools for UMLS, auditing of UMLS, components of UMLS, and coverage of UMLS. All studies in this category used UMLS as the study object. For example, auditing of UMLS includes publications on auditing-related studies that focus on the auditing of UMLS. If UMLS was used for other auditing purposes in a publication, then the publication was classified into UMLS in the other purposes category.

This category of publications also included modeling in UMLS. Other modeling-related publications that used UMLS were classified into the ontology and modeling categories. The publications that used UMLS to achieve different objectives (eg, identification of associations in texts) were classified into other categories based on their corresponding objectives. Multimedia Appendix 12 presents a list of all 150 UMLS publications on studies of the UMLS itself.

UMLS for Other Purposes

This category is used mainly for publications that use UMLS to achieve other purposes that cannot be covered by the themes noted above. In this category, auditing (not for UMLS auditing), consumer health, integrated system or data, and other research uses (including profile construction, management use, and deidentification) were included. Multimedia Appendix 13 presents a list of all 27 publications in this category.

Discussion

Summary and Interpretation of the Results

The results of the literature analysis showed the broad scope of the impact of UMLS in the academic world in the form of peer-reviewed journal publications, peer-reviewed conference publications, book chapters, and degree theses. What has been captured here, however, is only a *small fraction* of the real impact of UMLS. This literature analysis does not capture the following possible uses or impacts if no paper was published or if UMLS was not included in the title, abstract, or keywords: use of UMLS in the health information technology industry, health care delivery, software development, and any patent-related output.

The results show that UMLS has been broadly used, from basic science to applied projects in biomedical and health informatics. From the perspective of the number of publications, NLP, UMLS itself, and information retrieval are the 3 themes with the most publications. Anatomy is the medical subdomain with the most publications. French is the most active language, with a higher number of UMLS English-language publications of non-English languages. The large number of publications shows that certain themes are very active, although this literature analysis does not examine the overlap in different themes among different research projects. In addition, the number of publications should be used in a relative sense and with caution because a special issue of a journal or focused workshops or contests can skew the number of publications significantly.

In the *Results* section, the themes that repeatedly emerged during the literature analysis and synthesis have been listed. However, this is only an observation and a recording. From a purely ontological perspective, the same publications can be classified into different categories, depending on the axis. For example, a publication that focuses on automatic translation can be included in AI tools or applications; it can also be included in the multilingual category. Ideally, it will be useful to cross-reference each publication, which can then be classified into different categories. However, because of the large number of publications included in this literature analysis, such publications have been listed in only one category mostly (only 32/943, 3.4% of the publications was categorized into 2 categories; Table 1) instead of all possible categories. It is recognized that what was provided in this review is a *snapshot*

XSL•FO
RenderX

of the publications at the gross anatomy level, not a *panoramic view* of the publications with every single detail at the molecular level. This literature analysis serves as an archive of English-language UMLS peer-reviewed publications. The themes and subtopics and the publications under each theme or subtopic show only one perspective, not the only perspective, on the publications and their organizations. It is recognized that the search strategies can find only those publications for which UMLS plays a critical role. Some additional publications may use UMLS in their work; however, if UMLS was not listed in the title, MeSH terms, or abstract, then these publications will not be found through the search strategies. Therefore, the real impact of UMLS, even as academic output, is far larger than this review can represent.

Comparison With Existing UMLS-Use Publications

No systematic review or comprehensive literature analysis of UMLS was found during the literature search; however, there are publications on the use of UMLS through an analysis of UMLS annual reports [2] and the collection of surveys of UMLS users [173]. Nevertheless, the content of this literature analysis is complementary to these 2 studies [2,173]. The study by Fung et al [2] reported the geographical distribution of the users, the organizations of the UMLS license holders, types of information processed by UMLS, and areas of use of UMLS as well as users' support, communications, and feedback. The study [2] drew conclusions from 1427 UMLS annual reports for the year 2004.

Chen et al [173] reported the results of a 26-item survey sent to those on a UMLS mailing list (>600 subscribers). The research team analyzed the responses from 70 respondents, provided detailed categories of the users' employment and areas of use, and concluded that the top uses of UMLS were to access the source terminologies through UMLS and to achieve mapping among these terminologies. In addition, *terminology research*, *information retrieval*, *terminology translation*, *UMLS research*, and *NLP*, as well as *UMLS auditing*, were identified as the categories for the use of UMLS and as future priorities [173]. By comparison, this literature analysis paints a more comprehensive picture of publications in the last 30 years with regard to UMLS, by UMLS, and with UMLS. In analog language, this literature analysis is still at the level of *gross anatomy*; however, this review does provide more comprehensive categories, more detailed classifications, and clusters of publications on the topic. This literature analysis also lists degree-related doctoral theses in which the UMLS plays a critical role.

About UMLS

The original intended uses of UMLS involved four main areas: patient care, medical education, library service, and product development [1]. Comparing the results of this literature analysis with the originally intended uses, it is concluded that, although the literature analysis reflects an output largely within academic settings, the original intended uses have been achieved successfully. There are multiple themes and subtopics that can be matched to each of the 4 areas. For example, the patient care and medical subdomains can be placed in the patient care category. It was, however, recognized that such a literature analysis is not the best way to capture all the uses of UMLS in the real world, especially with regard to product development. Nevertheless, it is acknowledged that many electronic health records, AI, and NLP applications in the health field commonly use UMLS [5].

UMLS has been a cornerstone of academic activities in biomedical informatics, health informatics, and health information technology as a way to facilitate interoperability in broad medical and health fields. This literature analysis demonstrates only a small fraction of the true impact of UMLS. UMLS can be used as a terminology hub that hosts the most commonly used biomedical and health terminologies worldwide by using a universal concept unique identifier. A terminology hub is different from terminology in the same way that SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms) and UMLS are different but, at the same time, have some similarities. The 2 resources overlap but have mainly complementary purposes in the biomedical and health fields. SNOMED-CT is the most comprehensive medical terminology in the world, and UMLS includes SNOMED-CT and many additional terminologies. A common use of the UMLS is to provide machine-processable codes and meanings, which is similar to the use of SNOMED-CT; UMLS also provides mapping among different source terminologies. UMLS is critical for processing historical data and heterogeneous data sources, which will be a reality in health care in the near future. Therefore, to achieve seamless and effortless interoperability with a finer level of granularity in health care delivery sufficient to completely solve the puzzle described in e-patient Dave case study [174], at least at the front end, we need both SNOMED-CT and UMLS as well as many other resources.

However, UMLS is beyond a terminology hub. The intended uses of UMLS are mainly through software programs or systems. Many listed applications of UMLS include linking terms and codes in practice, pharmacy, and laboratory; facilitating mapping among different terminologies by providing terminology services; and serving as a lexical tool for NLP and AI, among others. Many additional UMLS applications have never been captured in the form of peer-reviewed publications. For example, my colleagues and I use UMLS as a teaching tool to introduce the concept of using controlled vocabularies to code medical records for health science major undergraduates.

Future Work

This literature analysis provides a descriptive observation of English-language peer-reviewed publications on UMLS over the last 30 years. It is an overview of the publications in terms of scope, as well as major themes and subtopics. More detailed content and literature analysis can be conducted for each theme. In this study, most of the publications were examined through an analysis of titles and abstracts, with some full-text publications when necessary. A more detailed full-text publication analysis may provide a more in-depth understanding of this topic.

Another possible direction is to examine the overlap among different themes and subtopics. For example, future research could analyze the overlaps by classifying a publication into as many categories as possible. If a publication has only 1 position within 1 theme or one subtopic, a theme graph can be generated

XSL•FO
RenderX

with all themes and subtopics (a graphical representation of Table 1) and all publications within each theme and each subtopic. Each publication would then have multiple positions in the theme graph. A visualization to consider the aggregated overlap (the same publication with multiple positions among multiple subtopics) among themes and subtopics can show or even inspire possible research collaboration opportunities among themes and subtopics.

Conclusions

This comprehensive literature analysis provides an overview with systematic evidence of the UMLS English-language peer-reviewed publications in the last 30 years. The analysis provides a descriptive observation of the themes and their subtopics of the publications and provides a detailed list of the publications in each category. UMLS has been used and published successfully in patient care, medical education, digital libraries, and software development in biomedicine, as well as in degree-related theses, building AI tools, data mining and knowledge discovery, and many more foundational works in methodology and middle layers that may lead to advanced products. The results, although largely in academia, demonstrate that UMLS achieves its intended uses successfully and has been used successfully and broadly beyond its original intentions. NLP, UMLS itself, and information retrieval are the three themes with the most publications. Anatomy is the most active medical subdomain. French is the most active language among the UMLS English-language publications of non-English languages. Nevertheless, this systematic literature analysis only captures publications in the English language; therefore, it should not be treated as a comprehensive impact description of UMLS, which should include English-language peer-reviewed publications and much more (eg, other language publications, patents, software, apps, care quality, and patient safety).

Acknowledgments

This study was partially supported by the National Library of Medicine of the NIH under award number R15LM012941 and partially supported by the National Institute of General Medical Sciences of the NIH under award numbers P20 GM121342 and R01GM138589. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIH.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Word cloud for all titles included in this systematic literature analysis (publication counts).
[PNG File , 335 KB - medinform_v9i8e20675_app1.png]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) extension for Scoping Reviews checklist.
[PDF File (Adobe PDF File), 543 KB - medinform_v9i8e20675_app2.pdf]

Multimedia Appendix 3
The yearly number of publications included in the literature analysis from 1990 to 2020.
[PNG File , 32 KB - medinform_v9i8e20675_app3.png]

Multimedia Appendix 4
Themes and major topics of applications of the Unified Medical Language System.
[PNG File , 58 KB - medinform_v9i8e20675_app4.png]

Multimedia Appendix 5
Unified Medical Language System is used for building artificial intelligence applications and tools.
[PDF File (Adobe PDF File), 246 KB - medinform_v9i8e20675_app5.pdf]

Multimedia Appendix 6
Publications on data mining, knowledge discovery, and text analysis using the Unified Medical Language System.
[PDF File (Adobe PDF File), 208 KB - medinform_v9i8e20675_app6.pdf]

Multimedia Appendix 7
Unified Medical Language System publications related to information retrieval.
[PDF File (Adobe PDF File), 337 KB - medinform_v9i8e20675_app7.pdf]

XSL•FO
RenderX

Multimedia Appendix 8
Unified Medical Language System in publications related to natural language processing.
[PDF File (Adobe PDF File), 469 KB - medinform_v9i8e20675_app8.pdf]

Multimedia Appendix 9
Unified Medical Language System publications in ontology and modeling.
[PDF File (Adobe PDF File), 286 KB - medinform_v9i8e20675_app9.pdf]

Multimedia Appendix 10
Publications that used the Unified Medical Language System in patient care.
[PDF File (Adobe PDF File), 225 KB - medinform_v9i8e20675_app10.pdf]

Multimedia Appendix 11
Unified Medical Language System publications about terminology studies.
[PDF File (Adobe PDF File), 293 KB - medinform_v9i8e20675_app11.pdf]

Multimedia Appendix 12
Publications about studies of the Unified Medical Language System itself.
[PDF File (Adobe PDF File), 344 KB - medinform_v9i8e20675_app12.pdf]

Multimedia Appendix 13
Unified Medical Language System publications for purposes other than the themes noted above.
[PDF File (Adobe PDF File), 211 KB - medinform_v9i8e20675_app13.pdf]

References

1. Humphreys BL, Lindberg DA, Hole WT. Assessing and enhancing the value of the UMLS Knowledge Sources. Proc Annu Symp Comput Appl Med Care 1991:78-82 [FREE Full text] [Medline: 1807711]
2. Fung KW, Hole WT, Srinivasan S. Who is using the UMLS and how - insights from the UMLS user annual reports. AMIA Annu Symp Proc 2006:274-278 [FREE Full text] [Medline: 17238346]
3. Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. J Biomed Inform 2003 Dec;36(6):414-432 [FREE Full text] [doi: 10.1016/j.jbi.2003.11.002] [Medline: 14759816]
4. Unified Medical Language System (UMLS). National Library of Medicine. 2004. URL: https://www.nlm.nih.gov/research/umls/about_umls.html [accessed 2021-07-31]
5. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
6. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Br Med J 2009;339:b2535 [FREE Full text] [Medline: 19622551]
7. Murphy SN, Barnett GO. Achieving automated narrative text interpretation using phrases in the electronic medical record. Proc AMIA Annu Fall Symp 1996:532-536 [FREE Full text] [Medline: 8947723]
8. Fu LS. A public domain unified medical language system (UMLS) patient database. Theses and dissertations: The University of Utah, Salt Lake City, UT. 1992. URL: https://dl.acm.org/doi/book/10.5555/166869 [accessed 2021-06-21]
9. Chen Y. Abstraction, Extension and Structural Auditing With the UMLS Semantic Network. Newark, NJ: New Jersey Institute of Technology; 2008.
10. Assefa S. Human Conceptual Representation and Knowledge Structure in The UMLS: A Coherence Analysis. Saarbrücken, Germany: VDM Verlag; 2009:1-128.
11. Mcinnes BT. Supervised and Knowledge-Based Methods for Disambiguating Terms in Biomedical Text Using the UMLS and Metamap. Minneapolis, MN: University of Minnesota; 2009.
12. Zhang L. Enriching and Designing Metaschemas for the UMLS Network. Newark, NJ: New Jersey Institute of Technology; 2004.
13. Min H. Structural Auditing Methodologies for Controlled Terminologies. Newark, NJ: New Jersey Institute of Technology; Oct 01, 2006.
14. Fowler RG, Gorry GA. The virtual object model for distributed hypertext. Theses and dissertations: Rice University. 1995. URL: https://scholarship.rice.edu/handle/1911/16822 [accessed 2021-06-21]
15. Leroy GA, Chen H. Facilitating knowledge discovery by integrating bottom-up and top-down knowledge sources: a text mining approach. Dissertations & Theses: The University of Arizona. 2003. URL: https://www.proquest.com/openview/8f6892b7db04631c2c1df8bcc62e3ffc/1?pq-origsite=gscholar&cbl=18750&diss=y [accessed 2021-06-21]

XSL•FO
RenderX

16. Gu H. Developing Techniques for Enhancing Comprehensibility of Controlled Medical Terminologies. Newark, NJ: New Jersey Institute of Technology; 1999.
17. Ruiz ME, Srinivasan P. Combining Machine Learning and Hierarchical Structures for Text Categorization. Iowa City, IA: The University of Iowa; 2001.
18. An YJ. Ontology Learning for the Semantic Deep Web. Newark, NJ: New Jersey Institute of Technology; 2008.
19. Zhou W. Knowledge-Intensive Conceptual Retrieval of Biomedical Literature. Chicago, IL: University of Illinois at Chicago; 2008.
20. Liu H. Corpus-Based Ambiguity Resolution of Biomedical Terms Using Knowledge Bases and Machine Learning. New York, NY: City University of New York; 2002.
21. Hu X, Lin TY, Song IY. A semi-supervised efficient learning approach to extract biological relationships from web-based biomedical digital library. Web Intelli Agent Sys 2006;4(3):327-339 [FREE Full text]
22. McCray AT. Digital library research and application. Stud Health Technol Inform 2000;76:51-62. [Medline: 10947501]
23. Kim EH, Oh JS, Song M. Exploring context-sensitive query reformulation in a biomedical digital library. In: Allen R, Hunter J, Zeng M, editors. Digital Libraries: Providing Quality Information. Cham: Springer; 2015:94-106.
24. Robinson J, de Lusignan S, Kostkova P, Madge B. Using UMLS to map from a library to a clinical classification: improving the functionality of a digital library. Stud Health Technol Inform 2006;121:86-95. [Medline: 17095807]
25. Mendonça EA, Cimino JJ. Automated knowledge extraction from MEDLINE citations. Proc AMIA Symp 2000:575-579 [FREE Full text] [Medline: 11079949]
26. Denny JC, Bastarache L, Sastre EA, Spickard A. Tracking medical students' clinical experiences using natural language processing. J Biomed Inform 2009 Oct;42(5):781-789 [FREE Full text] [doi: 10.1016/j.jbi.2009.02.004] [Medline: 19236956]
27. Kanter SL, Miller RA, Tan M, Schwartz J. Using POSTDOC to recognize biomedical concepts in medical school curricular documents. Bull Med Libr Assoc 1994 Jul;82(3):283-287 [FREE Full text] [Medline: 7920338]
28. Kanter SL. Using the UMLS to represent medical curriculum content. Proc Annu Symp Comput Appl Med Care 1993:762-765 [FREE Full text] [Medline: 8130579]
29. Denny JC, Smithers JD, Miller RA, Spickard A. "Understanding" medical school curriculum content using KnowledgeMap. J Am Med Inform Assoc 2003;10(4):351-362 [FREE Full text] [doi: 10.1197/jamia.M1176] [Medline: 12668688]
30. Komenda M, Schwarz D, Švancara J, Vaitsis C, Zary N, Dušek L. Practical use of medical terminology in curriculum mapping. Comput Biol Med 2015 Aug;63:74-82. [doi: 10.1016/j.compbiomed.2015.05.006] [Medline: 26037030]
31. Eysenbach G, Bauer J, Sager A, Bittorf A, Simon M, Diepgen T. An international dermatological image atlas on the WWW: practical use for undergraduate and continuing medical education, patient education and epidemiological research. Stud Health Technol Inform 1998;52 Pt 2:788-792. [Medline: 10384570]
32. Kumar A, Quaglini S, Stefanelli M, Ciccarese P, Caffi E. Modular representation of the guideline text: an approach for maintaining and updating the content of medical education. Med Inform Internet Med 2003 Jun;28(2):99-115. [doi: 10.1080/14639230310001600498] [Medline: 14692587]
33. Kazi H. A diverse and robust tutoring system for medical problem-based learning. In: Proceeding of the 15th International Conference on Computers in Education, ICCE 2007. 2007 Presented at: 15th International Conference on Computers in Education, ICCE 2007; November 5-9, 2007; Hiroshima, Japan p. 659-660 URL: https://www.researchgate.net/publication/221319250_A_Diverse_and_Robust_Tutoring_System_for_Medical_Problem-Based_Learning
34. Kazi H, Haddawy P, Suebnukarn S. Clinical reasoning gains in medical PBL: an UMLS based tutoring system. J Intell Inf Syst 2013 Apr 2;41(2):269-284 [FREE Full text] [doi: 10.1007/s10844-013-0244-9]
35. Kazi H, Haddawy P, Suebnukarn S. Employing UMLS for generating hints in a tutoring system for medical problem-based learning. J Biomed Inform 2012 Jun;45(3):557-565 [FREE Full text] [doi: 10.1016/j.jbi.2012.02.010] [Medline: 22429987]
36. Kazi H, Haddawy P, Suebnukarn S. Enriching Solution Space for Robustness in an Intelligent Tutoring System. In: Proceedings of the 2007 Conference on Supporting Learning Flow Through Integrative Technologies. Amsterdam: IOS Press; 2007:547-550.
37. Kazi H, Haddawy P, Suebnukarn S. Expanding the space of plausible solutions in a medical tutoring system for problem-based learning. Int J Artif Intell Edu 2009;19(3):309-334 [FREE Full text] [doi: 10.5555/1891970.1891974]
38. Kazi H, Haddawy P, Suebnukarn S. Leveraging a domain ontology to increase the quality of feedback in an intelligent tutoring system. In: Aleven V, Kay J, Mostow J, editors. Intelligent Tutoring Systems. Berlin: Springer; 2010:75-84.
39. Kazi H, Haddawy P, Suebnukarn S. Expanding the plausible solution space for robustness in an intelligent tutoring system. In: Intelligent Tutoring Systems. Berlin: Springer; 2008:583-592.
40. Kazi H, Haddawy P, Suebnukarn S. METEOR: medical tutor employing ontology for robustness. In: Proceedings of the 16th International Conference on Intelligent User Interfaces. 2011 Presented at: IUI '11: 16th International Conference on Intelligent User Interfaces; Feb 13-16, 2011; Palo Alto, CA p. 247-256. [doi: 10.1145/1943403.1943441]
41. Suebnukarn S, Haddawy P, Rhienmora P. A collaborative medical case authoring environment based on the UMLS. J Biomed Inform 2008 Apr;41(2):318-326 [FREE Full text] [doi: 10.1016/j.jbi.2007.08.007] [Medline: 17920337]
42. Zeng Y, Liu X, Wang Y, Shen F, Liu S, Rastegar-Mojarad M, et al. Recommending education materials for diabetic questions using information retrieval approaches. J Med Internet Res 2017 Oct 16;19(10):e342 [FREE Full text] [doi: 10.2196/jmir.7754] [Medline: 29038097]

XSL•FO
RenderX

43. Zou H, Lu QC, Durack JC, Chao C, Strasberg HR, Zhang Y, et al. Structured data management--the design and implementation of a web-based video archive prototype. Proc AMIA Symp 2001:786-790 [FREE Full text] [Medline: 11825293]
44. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A. The KnowledgeMap project: development of a concept-based medical school curriculum database. AMIA Annu Symp Proc 2003:195-199 [FREE Full text] [Medline: 14728161]
45. Song M, Heo GE, Lee D. Identifying the landscape of Alzheimer's disease research with network and content analysis. Scientometrics 2014 Jul 17;102(1):905-927 [FREE Full text] [doi: 10.1007/s11192-014-1372-x]
46. Dramé K, Diallo G, Delva F, Dartigues JF, Mouillet E, Salamon R, et al. Reuse of termino-ontological resources and text corpora for building a multilingual domain ontology: an application to Alzheimer's disease. J Biomed Inform 2014 Apr;48:171-182 [FREE Full text] [doi: 10.1016/j.jbi.2013.12.013] [Medline: 24382429]
47. Talos I, Rubin DL, Halle M, Musen M, Kikinis R. A prototype symbolic model of canonical functional neuroanatomy of the motor system. J Biomed Inform 2008 Apr;41(2):251-263 [FREE Full text] [doi: 10.1016/j.jbi.2007.11.003] [Medline: 18164666]
48. Rosse C, Mejino JL. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform 2003 Dec;36(6):478-500 [FREE Full text] [doi: 10.1016/j.jbi.2003.11.007] [Medline: 14759820]
49. Pyysalo S, Ananiadou S. Anatomical entity mention recognition at literature scale. Bioinformatics 2014 Mar 15;30(6):868-875 [FREE Full text] [doi: 10.1093/bioinformatics/btt580] [Medline: 24162468]
50. Rosse C, Ben Said M, Eno KR, Brinkley JF. Enhancements of anatomical information in UMLS knowledge sources. Proc Annu Symp Comput Appl Med Care 1995:873-877. [Medline: 8563417]
51. Sato L, McClure RC, Rouse RL, Schatz CA, Greenes RA. Enhancing the Metathesaurus with clinically relevant concepts: anatomic representations. Proc Annu Symp Comput Appl Med Care 1992:388-391. [Medline: 1482903]
52. Tran LT, Divita G, Carter ME, Judd J, Samore MH, Gundlapalli AV. Exploiting the UMLS Metathesaurus for extracting and categorizing concepts representing signs and symptoms to anatomically related organ systems. J Biomed Inform 2015 Dec;58:19-27. [doi: 10.1016/j.jbi.2015.08.024] [Medline: 26362345]
53. Bean CA. Formative evaluation of a frame-based model of locative relationships in human anatomy. Proc AMIA Annu Fall Symp 1997:625-629 [FREE Full text] [Medline: 9357701]
54. Sneiderman CA, Rindflesch TC, Bean CA. Identification of anatomical terminology in medical text. Proc AMIA Symp 1998:428-432 [FREE Full text] [Medline: 9929255]
55. Hishiki T, Ogasawara O, Tsuruoka Y, Okubo K. Indexing anatomical concepts to OMIM Clinical Synopsis using the UMLS Metathesaurus. In Silico Biol 2004;4(1):31-54. [Medline: 15089752]
56. Bashyam V, Taira RK. Indexing anatomical phrases in neuro-radiology reports to the UMLS 2005AA. AMIA Annu Symp Proc 2005 Dec:26-30. [Medline: 16778995]
57. Melgar HA, Beppler FD, Pacheco RC. Knowledge retrieval in the anatomical domain. In: Proceedings of the 1st ACM International Health Informatics Symposium. 2010 Presented at: IHI '10: ACM International Health Informatics Symposium; Nov 11-12, 2010; Arlington, VA p. 684-693. [doi: 10.1145/1882992.1883098]
58. Rosse C, Mejino JL, Modayur BR, Jakobovits R, Hinshaw KP, Brinkley JF. Motivation and organizational principles for anatomical knowledge representation: the digital anatomist symbolic knowledge base. J Am Med Inform Assoc 1998;5(1):17-40. [doi: 10.1136/jamia.1998.0050017] [Medline: 9452983]
59. Bowden DM, Song E, Kosheleva J, Dubach MF. NeuroNames: an ontology for the BrainInfo portal to neuroscience on the web. Neuroinformatics 2012 Jan;10(1):97-114. [doi: 10.1007/s12021-011-9128-8] [Medline: 21789500]
60. Mork P, Brinkley JF, Rosse C. OQAFMA Querying agent for the Foundational Model of Anatomy: a prototype for providing flexible and efficient access to large semantic networks. J Biomed Inform 2003 Dec;36(6):501-517 [FREE Full text] [doi: 10.1016/j.jbi.2003.11.004] [Medline: 14759821]
61. Cerveri P, Masseroli M, Pinciroli F. Remote access to anatomical information: an integration between semantic knowledge and visual data. Proc AMIA Symp 2000:126-130. [Medline: 11079858]
62. Mejino JL, Rosse C. The potential of the digital anatomist foundational model for assuring consistency in UMLS sources. Proc AMIA Symp 1998:825-829 [FREE Full text] [Medline: 9929334]
63. Merabti T, Soualmia LF, Grosjean J, Palombi O, Müller J, Darmoni SJ. Translating the Foundational Model of Anatomy into French using knowledge-based and lexical methods. BMC Med Inform Decis Mak 2011 Oct 26;11:65 [FREE Full text] [doi: 10.1186/1472-6947-11-65] [Medline: 22029629]
64. Lowe HJ, Huang Y, Regula DP. Using a statistical natural language Parser augmented with the UMLS specialist lexicon to assign SNOMED CT codes to anatomic sites and pathologic diagnoses in full text pathology reports. AMIA Annu Symp Proc 2009 Nov 14;2009:386-390. [Medline: 20351885]
65. Lamiell JM, Wojcik ZM, Isaacks J. Computer auditing of surgical operative reports written in English. Proc Annu Symp Comput Appl Med Care 1993:269-273. [Medline: 8130475]
66. Gabb HA, Blake C. An informatics approach to evaluating combined chemical exposures from consumer products: a case study of asthma-associated chemicals and potential endocrine disruptors. Environ Health Perspect 2016 Aug;124(8):1155-1165 [FREE Full text] [doi: 10.1289/ehp.1510529] [Medline: 26955064]

XSL•FO
RenderX

67. Choong MK, Tsafnat G, Hibbert P, Runciman WB, Coiera E. Linking clinical quality indicators to research evidence - a case study in asthma management for children. BMC Health Serv Res 2017 Jul 21;17(1):502 [FREE Full text] [doi: 10.1186/s12913-017-2324-y] [Medline: 28732500]
68. Achour SL, Dojat M, Rieux C, Bierling P, Lepage E. A UMLS-based knowledge acquisition tool for rule-based clinical decision support system development. J Am Med Inform Assoc 2001;8(4):351-360 [FREE Full text] [doi: 10.1136/jamia.2001.0080351] [Medline: 11418542]
69. Achour S, Dojat M, Brethon JM, Blain G, Lepage E. The use of the UMLS knowledge sources for the design of a domain specific ontology: a practical experience in blood transfusion. In: Horn W, Shahar Y, Lindberg G, Andreassen S, Wyatt J, editors. Artificial Intelligence in Medicine. Berlin: Springer; 1999:249-253.
70. Herskovic JR, Subramanian D, Cohen T, Bozzo-Silva PA, Bearden CF, Bernstam EV. Graph-based signal integration for high-throughput phenotyping. BMC Bioinformatics 2012;13 Suppl 13(Suppl 13):S2 [FREE Full text] [doi: 10.1186/1471-2105-13-S13-S2] [Medline: 23320851]
71. Zeng Z, Espino S, Roy A, Li X, Khan SA, Clare SE, et al. Using natural language processing and machine learning to identify breast cancer local recurrence. BMC Bioinformatics 2018 Dec 28;19(Suppl 17):498 [FREE Full text] [doi: 10.1186/s12859-018-2466-x] [Medline: 30591037]
72. Jadhav A, Sheth A, Pathak J. Analysis of online information searching for cardiovascular diseases on a consumer health information portal. AMIA Annu Symp Proc 2014;2014:739-748 [FREE Full text] [Medline: 25954380]
73. Varghese J, Sünninghausen S, Dugas M. Standardized cardiovascular quality assurance forms with multilingual support, UMLS coding and medical concept analyses. Stud Health Technol Inform 2015;216:837-841. [Medline: 26262169]
74. Shivade C, Malewadkar P, Fosler-Lussier E, Lai AM. Comparison of UMLS terminologies to identify risk of heart disease using clinical notes. J Biomed Inform 2015 Dec;58 Suppl(Suppl):S103-S110 [FREE Full text] [doi: 10.1016/j.jbi.2015.08.025] [Medline: 26375493]
75. Martínez M, Vázquez JM, Pereira J, Pazos A. Annotation of colorectal cancer data using the UMLS Metathesaurus. In: Knowledge-Based Intelligent Information and Engineering Systems. Berlin: Springer; 2008:58-65.
76. Becker M, Kasper S, Böckmann B, Jöckel K, Virchow I. Natural language processing of German clinical colorectal cancer notes for guideline-based treatment evaluation. Int J Med Inform 2019 Jul;127:141-146 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.04.022] [Medline: 31128826]
77. Du Y, Lin S, Huang Z. Making semantic annotation on patient data of depression. In: Proceedings of the 2nd International Conference on Medical and Health Informatics. 2018 Presented at: ICMHI '18: 2018 2nd International Conference on Medical and Health Informatics; June 8 - 10, 2018; Tsukuba Japan p. 134-137. [doi: 10.1145/3239438.3239453]
78. Kossman S, Jones J, Brennan PF. Tailoring online information retrieval to user's needs based on a logical semantic approach to natural language processing and UMLS mapping. AMIA Annu Symp Proc 2007 Oct 11:1015. [Medline: 18694113]
79. Gabetta M, Larizza C, Bellazzi R. A Unified Medical Language System (UMLS) based system for Literature-Based Discovery in medicine. Stud Health Technol Inform 2013;192:412-416. [Medline: 23920587]
80. Kim H, Song S, Kim Y, Song M. A display of conceptual structures in the epidemiologic literature. In: Proceedings of the ACM 8th International Workshop on Data and Text Mining in Bioinformatics. 2014 Presented at: CIKM '14: 2014 ACM Conference on Information and Knowledge Management; Nov 7, 2014; Shanghai, China p. 35. [doi: 10.1145/2665970.2665983]
81. Xu H, Lu Y, Jiang M, Liu M, Denny JC, Dai Q, et al. Mining biomedical literature for terms related to epidemiologic exposures. AMIA Annu Symp Proc 2010 Nov 13;2010:897-901 [FREE Full text] [Medline: 21347108]
82. Currie LM, Mellino LV, Cimino JJ, Bakken S. Development and representation of a fall-injury risk assessment instrument in a clinical information system. Stud Health Technol Inform 2004;107(Pt 1):721-725. [Medline: 15360907]
83. Bates J, Fodeh SJ, Brandt CA, Womack JA. Classification of radiology reports for falls in an HIV study cohort. J Am Med Inform Assoc 2016 Apr;23(e1):113-117 [FREE Full text] [doi: 10.1093/jamia/ocv155] [Medline: 26567329]
84. Kumar A, Ciccarese P, Smith B, Piazza M. Context-based task ontologies for clinical guidelines. Stud Health Technol Inform 2004;102:81-94. [Medline: 15853265]
85. Kumar A, Ciccarese P, Quaglini S, Stefanelli M, Caffi E, Boiocchi L. Relating UMLS semantic types and task-based ontology to computer-interpretable clinical practice guidelines. Stud Health Technol Inform 2003;95:469-474. [Medline: 14664031]
86. Campbell JR, Kallenberg GA, Sherrick RC. The clinical utility of META: an analysis for hypertension. Proc Annu Symp Comput Appl Med Care 1992:397-401 [FREE Full text] [Medline: 1482905]
87. Doan S, Maehara CK, Chaparro JD, Lu S, Liu R, Graham A, Pediatric Emergency Medicine Kawasaki Disease Research Group. Building a natural language processing tool to identify patients with high clinical suspicion for Kawasaki disease from Emergency Department notes. Acad Emerg Med 2016 May;23(5):628-636. [doi: 10.1111/acem.12925] [Medline: 26826020]
88. Ganzinger M, Knaup P. Semantic prerequisites for data sharing in a biomedical research network. Stud Health Technol Inform 2013;192:938. [Medline: 23920712]
89. Marquet G, Burgun A, Moussouni F, Guérin E, Le Duff F, Loréal O. BioMeKe: an ontology-based biomedical knowledge extraction system devoted to transcriptome analysis. Stud Health Technol Inform 2003;95:80-85. [Medline: 14663967]

XSL•FO
RenderX

90. Guérin E, Marquet G, Burgun A, Loréal O, Berti-Equille L, Leser U, et al. Integrating and warehousing liver gene expression data and related biomedical resources in GEDAW. In: Ludäscher B, Raschid L, editors. Data Integration in the Life Sciences. Berlin: Springer; 2005:158-174.
91. Turner CA, Jacobs AD, Marques CK, Oates JC, Kamen DL, Anderson PE, et al. Word2Vec inversion and traditional text classifiers for phenotyping lupus. BMC Med Inform Decis Mak 2017 Aug 22;17(1):126 [FREE Full text] [doi: 10.1186/s12911-017-0518-1] [Medline: 28830409]
92. Lyalina S, Percha B, LePendu P, Iyer SV, Altman RB, Shah NH. Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records. J Am Med Inform Assoc 2013 Dec;20(e2):297-305 [FREE Full text] [doi: 10.1136/amiajnl-2013-001933] [Medline: 23956017]
93. Zolnoori M, Fung KW, Patrick TB, Fontelo P, Kharrazi H, Faiola A, et al. A systematic approach for developing a corpus of patient reported adverse drug events: a case study for SSRI and SNRI medications. J Biomed Inform 2019 Feb;90:103091 [FREE Full text] [doi: 10.1016/j.jbi.2018.12.005] [Medline: 30611893]
94. Van Le D, Montgomery J, Kirkby KC, Scanlan J. Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting. J Biomed Inform 2018 Oct;86:49-58 [FREE Full text] [doi: 10.1016/j.jbi.2018.08.007] [Medline: 30118855]
95. Silverstein SM, Miller PL, Cullen MR. An information sources map for Occupational and Environmental Medicine: guidance to network-based information through domain-specific indexing. Proc Annu Symp Comput Appl Med Care 1993:616-620 [FREE Full text] [Medline: 8130548]
96. Harber P, Leroy G. Feasibility and utility of lexical analysis for occupational health text. J Occup Environ Med 2017 Jun;59(6):578-587. [doi: 10.1097/JOM.0000000000001035] [Medline: 28598934]
97. Sherertz DD, Tuttle MS, Olson NE, Hsu GT, Carlson RW, Fagan LM, et al. Accessing oncology information at the point of care: experience using speech, pen, and 3-D interfaces with a knowledge server. Medinfo 1995;8 Pt 1:792-795. [Medline: 8591330]
98. Berman JJ, Henson DE. Classifying the precancers: a metadata approach. BMC Med Inform Decis Mak 2003 Jun 20;3:8 [FREE Full text] [doi: 10.1186/1472-6947-3-8] [Medline: 12818004]
99. Sneiderman CA, Rindflesch TC, Aronson AR. Finding the findings: identification of findings in medical literature using restricted natural language processing. Proc AMIA Annu Fall Symp 1996:239-243 [FREE Full text] [Medline: 8947664]
100. Bejan CA, Xia F, Vanderwende L, Wurfel MM, Yetisgen-Yildiz M. Pneumonia identification using statistical feature selection. J Am Med Inform Assoc 2012;19(5):817-823 [FREE Full text] [doi: 10.1136/amiajnl-2011-000752] [Medline: 22539080]
101. Hardardottir A, Heimisdottir M, Aronson AR, Gunnarsdottir V. Standardized documentation in physical therapy: testing of validity and reliability of the PT-ITC and mapping it to the Metathesaurus. AMIA Annu Symp Proc 2008 Nov 06:964. [Medline: 18999048]
102. Westberg EE, Miller RA. The basis for using the internet to support the information needs of primary care. J Am Med Inform Assoc 1999;6(1):6-25 [FREE Full text] [Medline: 9925225]
103. He Z, Halper M, Perl Y, Elhanan G. Clinical clarity versus terminological order - the readiness of SNOMED CT concept descriptors for primary care. MIXHS 12 (2012) 2012;2012:1-6 [FREE Full text] [doi: 10.1145/2389672.2389674] [Medline: 26870837]
104. Mullins HC, Scanland PM, Collins D, Treece L, Petruzzi P, Goodson A, et al. The efficacy of SNOMED, Read Codes, and UMLS in coding ambulatory family practice clinical records. Proc AMIA Annu Fall Symp 1996:135-139 [FREE Full text] [Medline: 8947643]
105. Heintzelman NH, Taylor RJ, Simonsen L, Lustig R, Anderko D, Haythornthwaite JA, et al. Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. J Am Med Inform Assoc 2013;20(5):898-905 [FREE Full text] [doi: 10.1136/amiajnl-2012-001076] [Medline: 23144336]
106. Overton JA, Romagnoli C, Chhem R. Open Biomedical Ontologies applied to prostate cancer. Appl Ontol 2011;6(1):35-51 [FREE Full text] [doi: 10.3233/ao-2010-0081]
107. Fung KW, Richesson R, Bodenreider O. Coverage of rare disease names in standard terminologies and implications for patients, providers, and research. AMIA Annu Symp Proc 2014;2014:564-572 [FREE Full text] [Medline: 25954361]
108. Darmoni SJ, Soualmia LF, Letord C, Jaulent M, Griffon N, Thirion B, et al. Improving information retrieval using Medical Subject Headings Concepts: a test case on rare and chronic diseases. J Med Libr Assoc 2012 Jul;100(3):176-183 [FREE Full text] [doi: 10.3163/1536-5050.100.3.007] [Medline: 22879806]
109. Rance B, Snyder M, Lewis J, Bodenreider O. Leveraging terminological resources for mapping between rare disease information sources. Stud Health Technol Inform 2013;192:529-533 [FREE Full text] [Medline: 23920611]
110. Brandt M, Rath A, Devereau A, Aymé S. Mapping orphanet terminology to UMLS. In: Peleg M, Lavrač N, Combi C, editors. Artificial Intelligence in Medicine. Berlin: Springer; 2011:194-203.
111. Andrews JE, Shereff D, Patrick T, Richesson R. The question about questions: is DC a good choice to address the challenges of representation of clinical research questions and value sets? In: Proceedings of the DCMI International Conference on Dublin Core and Metadata Applications. 2010 Presented at: DCMI International Conference on Dublin Core and Metadata Applications; Oct 20-22, 2010; Pittusburg, PA p. 88-93 URL: https://dcpapers.dublincore.org/pubs/article/view/1032

XSL•FO
RenderX

112. Arif K, Qamar U, Wahab K, Riaz M. Building a biomedical ontology for respiratory tract infection. In: Proceedings of the 2019 7th International Conference on Computer and Communications Management. 2019 Presented at: 7th International Conference on Computer and Communications Management; July 27-29, 2019; Bangkok, Thailand p. 8-12 URL: https://doi-org.libproxy.clemson.edu/10.1145/3348445.3348461 [doi: 10.1145/3348445.3348461]
113. Sung S, Chen K, Wu DP, Hung L, Su Y, Hu Y. Applying natural language processing techniques to develop a task-specific EMR interface for timely stroke thrombolysis: a feasibility study. Int J Med Inform 2018 Dec;112:149-157. [doi: 10.1016/j.ijmedinf.2018.02.005] [Medline: 29500013]
114. Lu H, King C, Wu T, Shih F, Hsiao J, Zeng D, et al. Chinese chief complaint classification for syndromic surveillance. In: Intelligence and Security Informatics: Biosurveillance. Berlin: Springer; 2007:11-22.
115. Tolentino H, Matters M, Walop W, Law B, Tong W, Liu F, et al. Concept negation in free text components of vaccine safety reports. AMIA Annu Symp Proc 2006:1122 [FREE Full text] [Medline: 17238741]
116. Chapman WW, Fiszman M, Dowling JN, Chapman BE, Rindflesch TC. Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. Stud Health Technol Inform 2004;107(Pt 1):487-491. [Medline: 15360860]
117. Lau AS, Tse SH. Development of the ontology using a problem-driven approach: in the context of traditional Chinese medicine diagnosis. Int J Knowl Eng.Data Min 2010;1(1):37-49 [FREE Full text] [doi: 10.1504/ijkedm.2010.032579]
118. Zhu X, Lee KP, Cimino JJ. Knowledge representation of traditional Chinese acupuncture points using the UMLS and a terminology model. In: Proceedings of the IDEAS Workshop on Medical Information Systems: The Digital Hospital (IDEAS-DH'04). 2004 Presented at: IDEAS Workshop on Medical Information Systems: The Digital Hospital (IDEAS-DH'04); Sept 1-3, 2004; Beijing, China p. 40-48. [doi: 10.1109/ideadh.2004.15]
119. Burgun A, Botti G, Lukacs B, Mayeux D, Seka LP, Delamarre D, et al. A system that facilitates the orientation within procedure nomenclatures through a semantic approach. Med Inform (Lond) 1994;19(4):297-310. [doi: 10.3109/14639239409025335] [Medline: 7603121]
120. Burgun A, Delamarre D, Botti G, Lukacs B, Mayeux D, Bremond M, et al. Designing a sub-set of the UMLS knowledge base applied to a clinical domain: methods and evaluation. Proc Annu Symp Comput Appl Med Care 1994:968 [FREE Full text] [Medline: 7950072]
121. Moreira A, Alonso-Calvo R, Muñoz A, Crespo J. Enhancing collaborative case diagnoses through unified medical language system-based disambiguation: a case study of the zika virus. Telemed J E Health 2017 Dec;23(7):608-614. [doi: 10.1089/tmj.2016.0203] [Medline: 28092493]
122. Nikolova I. Angelova, identifying relations between medical concepts by parsing UMLS® definitions. In: Conceptual Structures for Discovering Knowledge. Berlin: Springer; 2011:173-186.
123. Afzal Z, Pons E, Kang N, Sturkenboom MC, Schuemie MJ, Kors JA. ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. BMC Bioinformatics 2014 Nov 29;15:373 [FREE Full text] [doi: 10.1186/s12859-014-0373-3] [Medline: 25432799]
124. Deléger L, Merabti T, Lecrocq T, Joubert M, Zweigenbaum P, Darmoni S. A twofold strategy for translating a medical terminology into French. AMIA Annu Symp Proc 2010 Nov 13;2010:152-156 [FREE Full text] [Medline: 21346959]
125. Fabry P, Baud R, Burgun A, Lovis C. Amplification of Terminologia anatomica by French language terms using Latin terms matching algorithm: a prototype for other language. Int J Med Inform 2006 Jul;75(7):542-552. [doi: 10.1016/j.ijmedinf.2005.08.008] [Medline: 16203172]
126. Merabti T, Massari P, Joubert M, Sadou E, Lecroq T, Abdoune H, et al. An automated approach to map a French terminology to UMLS. Stud Health Technol Inform 2010;160(Pt 2):1040-1044. [Medline: 20841842]
127. Maisonnasse L, Harrathi F, Roussey C, Calabretto S. Analysis combination and Pseudo relevance feedback in conceptual language model: LIRIS Participation at ImageCLEFMed. In: Multilingual Information Access Evaluation II. Multimedia Experiments. Berlin: Springer; 2009:203-210.
128. Joubert M, Abdoune H, Merabti T, Darmoni S, Fieschi M. Assisting the translation of SNOMED CT into French using UMLS and four representative French-language terminologies. AMIA Annu Symp Proc 2009 Nov 14;2009:291-295 [FREE Full text] [Medline: 20351867]
129. Abdoune H, Merabti T, Darmoni SJ, Joubert M. Assisting the translation of the CORE subset of SNOMED CT into French. Stud Health Technol Inform 2011;169:819-823. [Medline: 21893861]
130. Grabar N, Varoutas P, Rizand P, Livartowski A, Hamon T. Automatic acquisition of synonyms from French UMLS for enhanced search of EHRs. Stud Health Technol Inform 2008;136:809-814. [Medline: 18487831]
131. Le Duff F, Burgun A, Pouliquen B, Delamarre D, Le Beux P. Automatic enrichment of the unified medical language system starting from the ADM knowledge base. Stud Health Technol Inform 1999;68:881-886. [Medline: 10725024]
132. Joubert M, Peretti A, Darmoni S, Dahamna B, Fieschi M. Contribution to an automated indexing of French-language health web sites. AMIA Annu Symp Proc 2006:409-413 [FREE Full text] [Medline: 17238373]
133. Deléger L, Merkel M, Zweigenbaum P. Contribution to terminology internationalization by word alignment in parallel corpora. AMIA Annu Symp Proc 2006:185-189 [FREE Full text] [Medline: 17238328]
134. Ruiz M, Névéol A. Evaluation of Automatically Assigned MeSH Terms for Retrieval of Medical Images. In: Proceedings of the 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007. 2007 Presented at: 8th Workshop of the

XSL•FO
RenderX

Cross-Language Evaluation Forum, CLEF 2007; Sep 19-21, 2007; Budapest, Hungary p. 641-648. [doi: 10.1007/978-3-540-85760-0_7]
135. Tran TD, Garcelon N, Burgun A, Le Beux P. Experiments in cross-language medical information retrieval using a mixing translation module. Stud Health Technol Inform 2004;107(Pt 2):946-949. [Medline: 15360952]
136. Besana P. From french EHR to NCI ontology via UMLS. In: Proceedings of the 5th International Conference on Ontology Matching. 2010 Presented at: 5th International Conference on Ontology Matching; Nov 7, 2010; Shanghai, China p. 222-223 URL: http://www.dit.unitn.it/~p2p/OM-2010//om2010_proceedings.pdf
137. Bodenreider O, McCray AT. From French vocabulary to the Unified Medical Language System: a preliminary study. Stud Health Technol Inform 1998;52 Pt 1(0 1):670-674 [FREE Full text] [Medline: 10384539]
138. Le Duff F, Burgun A, Cleret M, Pouliquen B, Barac'h V, Le Beux P. Knowledge acquisition to qualify Unified Medical Language System interconceptual relationships. Proc AMIA Symp 2000:482-486 [FREE Full text] [Medline: 11079930]
139. Merabti T, Abdoune H, Letord C, Sakji S, Joubert M, Darmoni SJ. Mapping the ATC classification to the UMLS metathesaurus: some pragmatic applications. Stud Health Technol Inform 2011;166:206-213. [Medline: 21685626]
140. Delbecque T, Zweigenbaum P. MetaCoDe: A lightweight UMLS mapping tool. In: Artificial Intelligence in Medicine. Berlin: Springer; 2007:242-246.
141. Bousquet C, Souvignet J, Merabti T, Sadou E, Trombert B, Rodrigues J. Method for mapping the French CCAM terminology to the UMLS metathesaurus. Stud Health Technol Inform 2012;180:164-168. [Medline: 22874173]
142. Cossin S, Lebrun L, Lobre G, Loustau R, Jouhet V, Griffier R, et al. Romedi: An open data source about French drugs on the semantic web. Stud Health Technol Inform 2019 Aug 21;264:79-82. [doi: 10.3233/SHTI190187] [Medline: 31437889]
143. Ventura JA. Towards a mixed approach to extract biomedical terms from text corpus. Int J Knowl Disc Bioinfo 2014;4(1):1-15 [FREE Full text] [doi: 10.4018/ijkdb.2014010101]
144. Zweigenbaum P, Baud R, Burgun A, Namer F, Jarrousse E, Grabar N, et al. Towards a unified medical lexicon for French. Stud Health Technol Inform 2003;95:415-420. [Medline: 14664022]
145. Darmoni SJ, Jarrousse E, Zweigenbaum P, Le Beux P, Namer F, Baud R, et al. VUMeF: extending the French involvement in the UMLS Metathesaurus. AMIA Annu Symp Proc 2003:824 [FREE Full text] [Medline: 14728329]
146. Markó K, Schulz S, Hahn U. Automatic lexicon acquisition for a medical cross-language information retrieval system. Stud Health Technol Inform 2005;116:829-834. [Medline: 16160361]
147. Becker M, Böckmann B. Extraction of UMLS® concepts using Apache cTAKES™ for German language. Stud Health Technol Inform 2016;223:71-76. [Medline: 27139387]
148. Weske-Heck G, Zaiss A, Zabel M, Schulz S, Giere W, Schopen M, et al. The German specialist lexicon. Proc AMIA Symp 2002:884-888 [FREE Full text] [Medline: 12463952]
149. Widdows D, Peters S, Cederberg S, Chan C, Steffen D, Buitelaar P. Unsupervised monolingual and bilingual word-sense disambiguation of medical documents using UMLS. In: Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine. 2003 Presented at: ACL 2003 Workshop on Natural Language Processing in Biomedicine; July 11, 2003; Sapporo, Japan p. 9-16. [doi: 10.3115/1118958.1118960]
150. Chiaramello E, Pinciroli F, Bonalumi A, Caroli A, Tognola G. Use of "off-the-shelf" information extraction algorithms in clinical informatics: a feasibility study of MetaMap annotation of Italian medical notes. J Biomed Inform 2016 Oct;63:22-32. [doi: 10.1016/j.jbi.2016.07.017] [Medline: 27444186]
151. Nishimoto N, Terae S, Uesugi M, Ogasawara K, Sakurai T. Development of a medical-text parsing algorithm based on character adjacent probability distribution for Japanese radiology reports. Methods Inf Med 2008;47(6):513-521. [doi: 10.3414/me9127] [Medline: 19057808]
152. Onogi Y, Ohe K, Tanaka M, Nozoe A, Sasaki T, Sato M, et al. Mapping Japanese medical terms to UMLS Metathesaurus. Stud Health Technol Inform 2004;107(Pt 1):406-410. [Medline: 15360844]
153. Nishimoto N, Satoshi T, Jiang G, Uesugi M, Terashita T, Tanikawa T, et al. Semantic distribution study of noun*noun compounds in the Japanese CT clinical reports. AMIA Annu Symp Proc 2006:1048 [FREE Full text] [Medline: 17238667]
154. Han S, Kwak M, Kim S, Yoo S, Park H, Kijoo J, et al. A comparative study on concept representation between the UMLS and the clinical terms in Korean medical records. Stud Health Technol Inform 2004;107(Pt 1):616-620. [Medline: 15360886]
155. Lee KN, Yoon J, Min WK, Lim HS, Song J, Chae SL, et al. Standardization of terminology in laboratory medicine II. J Korean Med Sci 2008 Aug;23(4):711-713. [doi: 10.3346/jkms.2008.23.4.711] [Medline: 18756062]
156. Han S, Choi J. The comparative study on concept representation between the UMLS and the clinical terms in Korean medical records. Int J Med Inform 2005 Jan;74(1):67-76. [doi: 10.1016/j.ijmedinf.2004.09.004] [Medline: 15626637]
157. Park HK, Choi J. Towards chronological summary of medical records. AMIA Annu Symp Proc 2007 Oct 11:911. [Medline: 18694011]
158. Kang B, Kim D, Kim H. Two-Phase chief complaint mapping to the UMLS metathesaurus in Korean electronic medical records. IEEE Trans Inf Technol Biomed 2009 Jan;13(1):78-86. [doi: 10.1109/TITB.2008.2007103] [Medline: 19129026]
159. Ruiz ME, Southwick SB. UB at CLEF 2005: Bilingual CLIR and medical image retrieval tasks. In: Accessing Multilingual Information Repositories. Berlin: Springer; 2006:737-743.
160. Carrero F, Cortizo JC, Gómez JM. Building a Spanish MMTx by using automatic translation and biomedical ontologies. In: Intelligent Data Engineering and Automated Learning. Berlin: Springer; 2008:346-353.

XSL•FO
RenderX

161. Buendía F, Gayoso-Cabada J, Juanes-Méndez J, Martín-Izquierdo M. Cataloguing Spanish medical reports with UMLS terms. In: Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality. 2019 Presented at: The Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality; Oct 16-18, 2019; León, Spain p. 423-430. [doi: 10.1145/3362789.3362878]
162. Carrero F, Cortizo J, Gómez J, de Buenaga M. In the development of a Spanish metamap. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. 2008 Presented at: 17th ACM Conference on Information and Knowledge Management; Oct 26-30, 2008; Napa Valley, California, p. 1465-1466. [doi: 10.1145/1458082.1458335]
163. Markó K, Schulz S, Hahn U. Automatic lexeme acquisition for a multilingual medical subword thesaurus. Int J Med Inform 2007;76(2-3):184-189. [doi: 10.1016/j.ijmedinf.2006.05.032] [Medline: 16839808]
164. Eichmann D, Ruiz M, Srinivasan P. Cross-language information retrieval with the UMLS metathesaurus. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and development in Information Retrieval. 1998 Presented at: 21st Annual International ACM SIGIR Conference on Research and development in Information Retrieval; Aug 24-28, 1998; Melbourne, Australia p. 72-80. [doi: 10.1145/290941.290959]
165. Tringali M, Hole WT, Srinivasan S. Integration of a standard gastrointestinal endoscopy terminology in the UMLS Metathesaurus. Proc AMIA Symp 2002:801-805 [FREE Full text] [Medline: 12463935]
166. Hersh WR, Donohoe LC. SAPHIRE International: a tool for cross-language information retrieval. Proc AMIA Symp 1998:673-677 [FREE Full text] [Medline: 9929304]
167. Göbel G, Andreatta S, Masser J, Pfeiffer KP. A multilingual medical thesaurus browser for patients and medical content managers. Stud Health Technol Inform 2001;84(Pt 1):333-337. [Medline: 11604758]
168. Hellrich J, Hahn U. Fostering Multilinguality in the UMLS: A computational approach to terminology expansion for multiple languages. AMIA Annu Symp Proc 2014;2014:655-660 [FREE Full text] [Medline: 25954371]
169. Déjean H, Gaussier E, Sadat F. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In: Proceedings of the 19th International Conference on Computational Linguistics. 2002 Presented at: 19th International Conference on Computational Linguistics; Aug 24-Sept 1, 2002; Taipei, Taiwan p. 1-7. [doi: 10.3115/1072228.1072394]
170. Hellrich J, Hahn U. Exploiting parallel corpora to scale up multilingual biomedical terminologies. Stud Health Technol Inform 2014;205:575-578. [Medline: 25160251]
171. Hellrich J, Schulz S, Buechel S, Hahn U. JuFiT: A configurable rule engine for filtering and generating new multilingual UMLS terms. AMIA Annu Symp Proc 2015;2015:604-610 [FREE Full text] [Medline: 26958195]
172. Guillén R. Reusing translated terms to expand a multilingual thesaurus. In: Machine Translation and the Information Soup. Berlin: Springer; 1998:374-383.
173. Chen Y, Perl Y, Geller J, Cimino JJ. Analysis of a study of the users, uses, and future agenda of the UMLS. J Am Med Inform Assoc 2007;14(2):221-231 [FREE Full text] [doi: 10.1197/jamia.M2202] [Medline: 17213497]
174. Hoyt R, Yoshihashi A. Health Informatics: Practical Guide for Healthcare and Information Technology Professionals. Morrisville, North Carolina: Lulu Press; 2014:1-533.

Abbreviations

AI: artificial intelligence
MeSH: Medical Subject Headings
NIH: National Institutes of Health
NLP: natural language processing
SNOMED-CT: Systematized Nomenclature of Medicine-Clinical Terms
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 25.05.20; peer-reviewed by A Wang, J Varghese, M Dugas, M Torii; comments to author 28.10.20; revised version received 25.11.20; accepted 02.07.21; published 27.08.21.

Please cite as:
Jing X
The Unified Medical Language System at 30 Years and How It Is Used and Published: Systematic Review and Content Analysis
JMIR Med Inform 2021;9(8):e20675
URL: https://medinform.jmir.org/2021/8/e20675
doi:10.2196/20675
PMID:34236337

XSL•FO
RenderX

©Xia Jing. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 27.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Team Dynamics in Hospital Workflows: An Exploratory Study of a Smartphone Task Manager

Danula Hettiachchi[1], BSc, PhD; Lachie Hayes[2], MBBS; Jorge Goncalves[1], BSc, MSc, PhD; Vassilis Kostakos[1], BSc, PhD

[1]School of Computing and Information Systems, The University of Melbourne, Parkville, Australia

[2]Nothern Hospital Epping, Epping, Australia

Corresponding Author:
Danula Hettiachchi, BSc, PhD
School of Computing and Information Systems
The University of Melbourne
Gratten St
Parkville, 3010
Australia
Phone: 61 474435815
Email: dhettiachchi@student.unimelb.edu.au

Abstract

Background: Although convenient and reliable modern messaging apps like WhatsApp enable efficient communication among hospital staff, hospitals are now pivoting toward purpose-built structured communication apps for various reasons, including security and privacy concerns. However, there is limited understanding of how we can examine and improve hospital workflows using the data collected through such apps as an alternative to costly and challenging research methods like ethnography and patient record analysis.

Objective: We seek to identify whether the structure of the collected communication data provides insights into hospitals' workflows. Our analysis also aims to identify ways in which task management platforms can be improved and designed to better support clinical workflows.

Methods: We present an exploratory analysis of clinical task records collected over 22 months through a smartphone app that enables structured communication between staff to manage and execute clinical workflows. We collected over 300,000 task records between July 2018 and May 2020 completed by staff members including doctors, nurses, and pharmacists across all wards in an Australian hospital.

Results: We show that important insights into how teams function in a clinical setting can be readily drawn from task assignment data. Our analysis indicates that predefined labels such as urgency and task type are important and impact how tasks are accepted and completed. Our results show that both task sent-to-accepted ($P<.001$) and sent-to-completed ($P<.001$) times are significantly higher for routine tasks when compared to urgent tasks. We also show how task acceptance varies across teams and roles and that internal tasks are more efficiently managed than external tasks, possibly due to increased trust among team members. For example, task sent-to-accepted time (minutes) is significantly higher ($P<.001$) for external assignments (mean 22.10, SD 91.45) when compared to internal assignments (mean 19.03, SD 82.66).

Conclusions: Smartphone-based task assignment apps can provide unique insights into team dynamics in clinical settings. These insights can be used to further improve how well these systems support clinical work and staff.

(JMIR Med Inform 2021;9(8):e28245) doi:10.2196/28245

KEYWORDS

task assignment; smartphones; hospital communication; clinical workflows; mobile app; clinical platform; mHealth

Introduction

The free availability, widespread use, reliability, and intuitive nature of modern communication interfaces have led to the use of popular messaging apps like WhatsApp among medical staff [1]. These apps can bring many benefits to clinical teams, such as prompt communication, reduction in interruptions, ability to form groups, and convenient access to other staff members [1,2].

XSL•FO
RenderX

Positive outcomes of using WhatsApp for clinical communication have been highlighted in numerous studies conducted among emergency surgery team members in a hospital in the United Kingdom [3], surgeons of two hospitals in Italy [4], orthopedic team members in a hospital in Ireland [5], and all professionals in medical and emergency teams in a major hospital in Malaysia [2].

However, using personal devices and generic messaging apps poses privacy and security concerns [6]. For example, when handling protected health information (PHI), medical professionals in the United States are required to adhere to communication regulations set out in the Health Insurance Portability and Accountability Act (HIPAA). Nevertheless, there is a lack of awareness and no consensus among medical staff on what apps are considered to be compliant with HIPAA [7]. Other adverse consequences of using regular messaging apps for work include information overload and the impact on the separation between one's work and personal life [2,6].

Although such off-the-shelf messaging apps are ubiquitous and convenient [2], they do not provide structured task management features that would allow medical staff to send, accept, and prioritize tasks. Khanna et al [8] describe how a smartphone-based paging app can potentially bring numerous benefits to medical staff. Integrated interfaces can reduce the effort required to accept or send tasks. Notifications can streamline work by reducing the need to check for updates proactively. Such apps could also learn and initiate routine communications without human intervention to reduce redundant tasks. Similarly, Patel et al [9] report that mobile app–based communication improves efficiency, reduces interruptions, and allows health care professionals to transfer information reliably and clearly compared to using a standard pager system. Therefore, communication tools that overcome the aforementioned security challenges and provide smart task management capabilities are desirable within the health care industry.

In addition, a centralized tool for communication and task assignment can provide greater value for hospitals by providing the opportunity to analyze and understand the team dynamics of medical activity [10,11]. Literature presents such computer-based task management systems and their positive outcomes (eg, a desktop-based system implemented at Middlemore Hospital, Auckland, New Zealand [12], and a system that sends messages to dedicated team smartphones at Toronto General Hospital and Toronto Western Hospital, Toronto, Canada [13]). Similarly, Dock Health is a more modern team collaboration and task management app adopted by medical staff at Boston Children's Hospital, Boston, MA, United States [14].

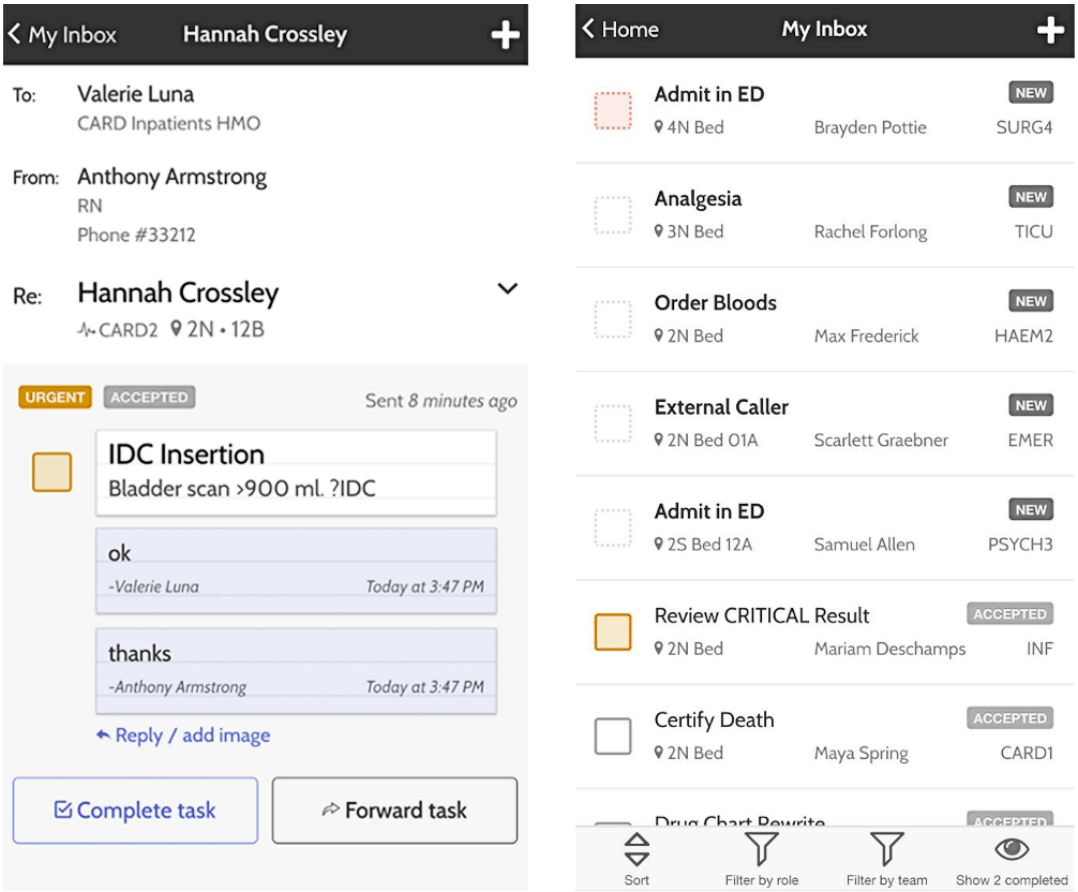
Although hospitals have begun to adopt smartphone-based task management and communication systems, it remains unclear how the task records can be used to gain a better understanding of medical team dynamics and communication patterns. Such insights can be critical in further improving the user experience of the app as well as increasing the efficiency of the task assignment process. Previous attempts to examine hospital workflows and staff communication include ethnographic methods and interviews with clinical staff [11,15], analyzing patient electronic medical records (EMRs) [16], and mapping call data [17]. However, it is costly and challenging to implement such methods at scale, and they fail to provide in-depth and timely data, unlike data collection through task management apps.

In this paper, we present an exploratory analysis of medical task assignment data collected through a mobile app deployed at a hospital in Australia for 22 months. In our study, all staff at a hospital started using a bespoke smartphone platform, MedTasker [18] (Figure 1), to manage and execute clinical workflows. Tasks are defined as work units assigned to a specific staff member through our app and include tasks such as reviewing medications, admitting a patient to a ward, or conducting a medical procedure.

In our exploratory analysis, we seek to identify whether the increased granularity of the collected data provides insights into the hospital's workflows (ie, repeatable patterns of clinical tasks). For example, we are interested in understanding the various individual and team dynamics (ie, factors that influence the direction of a team's behavior and performance) that underpin workflows at the hospital. Furthermore, our analysis also seeks to identify ways in which task management platforms can be improved and designed to better support clinical workflows.

XSL•FO
RenderX

Figure 1. Examples of the MedTasker mobile app interface [18]. Names and task details are fictional.



Methods

MedTasker App

MedTasker [18] is a mobile communication and task management platform built for hospitals. In addition to the typical task management features to create, send, accept, and forward tasks, MedTasker provides a staff directory and supports comments, notes, and attachments. The app includes an escalation process, which can alert additional staff members when the recipient does not accept tasks within a specified period. In addition, hospitals can configure the escalation process to tailor their needs.

The app can also integrate with existing hospital systems such as the Patient Administration System, EMRs, pathology, radiology, paging systems, and Active Directory. All communications using MedTasker use end-to-end encryption. The app also provides a way to share clinical images and files in a secure and privacy-compliant way. Since all tasks are tracked in real time, MedTasker enables visualizations to better manage team workloads and audit task workflows when needed.

Data Collection

MedTasker has been used as the regular task management solution at Northern Hospital, Epping, Victoria, Australia, since 2018. Northern Hospital is the major public health care provider for acute, maternity, subacute, and specialist services in Melbourne's northern suburbs and surrounding regional areas. The hospital has over 5300 dedicated professional staff and treats over 94,000 patients admitted yearly. All staff members including doctors, nurses, and pharmacists use the MedTasker app for general task management across all wards. The app is accessible through desktop and mobile devices (Android and iPhone). Staff typically use the app on their personal smartphones and connect to the internet through the hospital's Wi-Fi network. We collected task assignment data through MedTasker for a period of 22 months starting from July 1, 2018. Data fields include recipient team and level, patient team, sender role, task type, history, urgency, and time. Individual sender and receiver details in the data set were anonymized.

Preprocessing

We adopted several preprocessing steps to ensure the reliability of the data. First, we filtered a number of tasks that were completed immediately after accepting (ie, accepted-to-completed time is 0 minutes). These records mainly correspond to instances where the staff member has already completed the task by the time they mark it as accepted in the system. Second, we removed tasks where sent-to-accepted time exceeded 24 hours. We also excluded all the tasks that were not marked as completed (ie, incomplete tasks) by the end of the time window considered.

Analysis

Our analysis used statistical packages in R (version 3.6.1, R Foundation for Statistical Computing). We used nonparametric tests, the Wilcoxon rank-sum test, and the Kruskal-Wallis rank-sum test when comparing the task acceptance time between different groups and conditions. Finally, our results were discussed in focus groups and interviews with key hospital

XSL•FO
RenderX

members, who reflected on our findings and helped us interpret them. Focus groups and interviews took place at the start of the analysis, halfway through data analysis, and upon completion of the analysis. These sessions lasted about one hour and involved hospital members and the authors of the study.

We mainly used task acceptance time, task completion time, redirection percentage, and task escalation percentage for our analysis. The metric selection was informed by the input provided by the hospital staff; they regularly use these metrics to monitor and evaluate the effectiveness of the task assignment process.

Results

Task Creation

In total, 317,372 tasks were sent and completed between July 2018 and May 2020, with a mean sent-to-accepted time of 15.89 (SD 79.72) minutes. In Figure 2, we observe that the number of tasks sent through the MedTasker app gradually increased within the first year and maintained a consistent level thereafter. On average, 419.88 (SD 121.11) tasks were sent each day for the first 12 months. The average daily task count then increased to 501.83 (SD 123.25) for the remaining duration. Figure 3 shows how the time of day impacts task acceptance and behavior across different types of tasks.

Figure 2. Tasks sent throughout the data collection period.

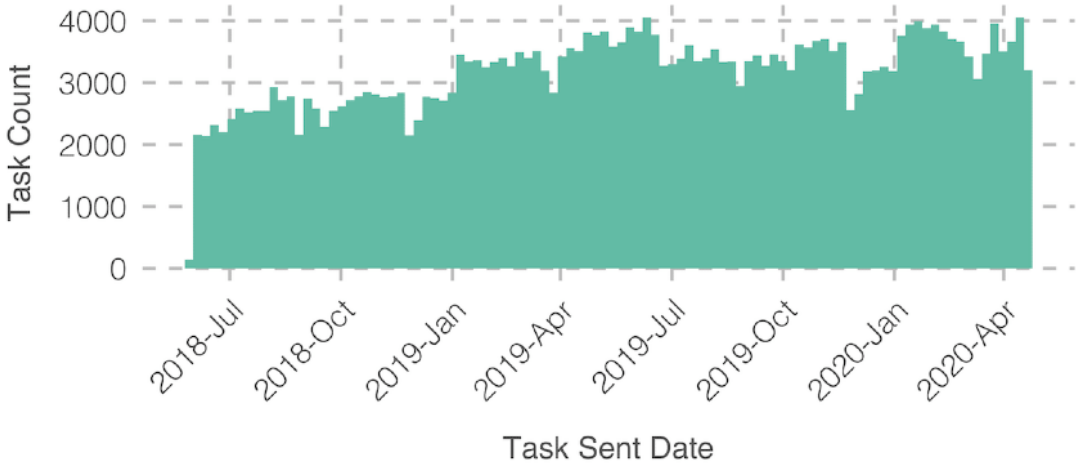
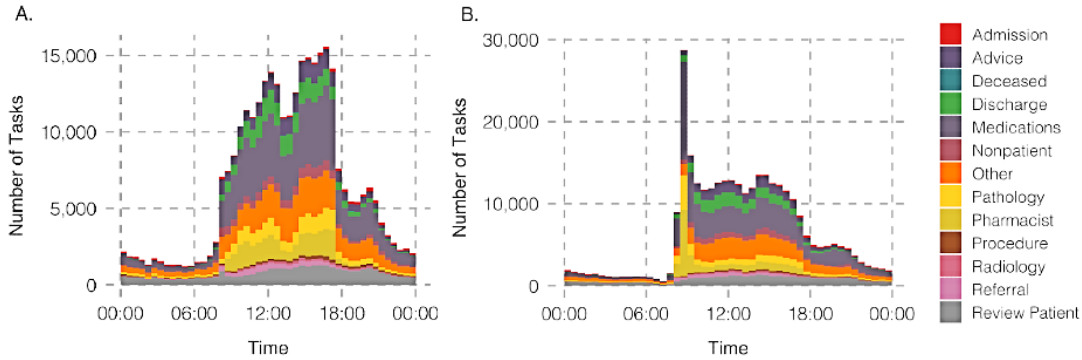


Figure 3. Task acceptance (A) and completion (B) throughout the day.

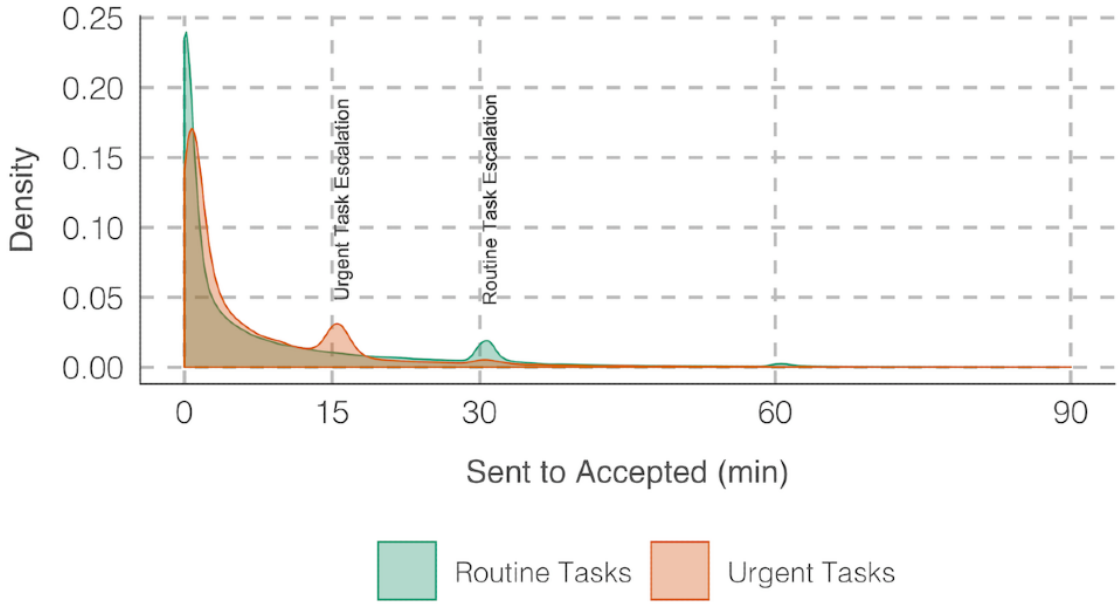


Task Urgency and Escalation

A total of 13,168 of 317,372 (4.1%) tasks were categorized as urgent, and the remaining tasks were routine tasks. The mean sent-to-accepted time was 13.85 (SD 77.81) minutes for urgent tasks and 15.97 (SD 79.80) minutes for routine tasks. A Wilcoxon rank-sum test showed that sent-to-accepted time is significantly higher for routine tasks when compared to urgent tasks (W=1,911,381,996; P<.001), suggesting that recipients accept urgent tasks quicker than routine tasks. Figure 4 shows the impact of task urgency and escalation on sent-to-accepted time. Similarly, the sent-to-completed time (hours) was significantly higher (W=2,651,177,476; P<.001) for routine tasks (mean 14.74, SD 53.25) when compared to urgent tasks (mean 4.10, SD 17.84). The results also indicated that urgent tasks are more likely to be escalated when compared to routine tasks (χ^2_1=453.17; P<.001).

XSL•FO
RenderX

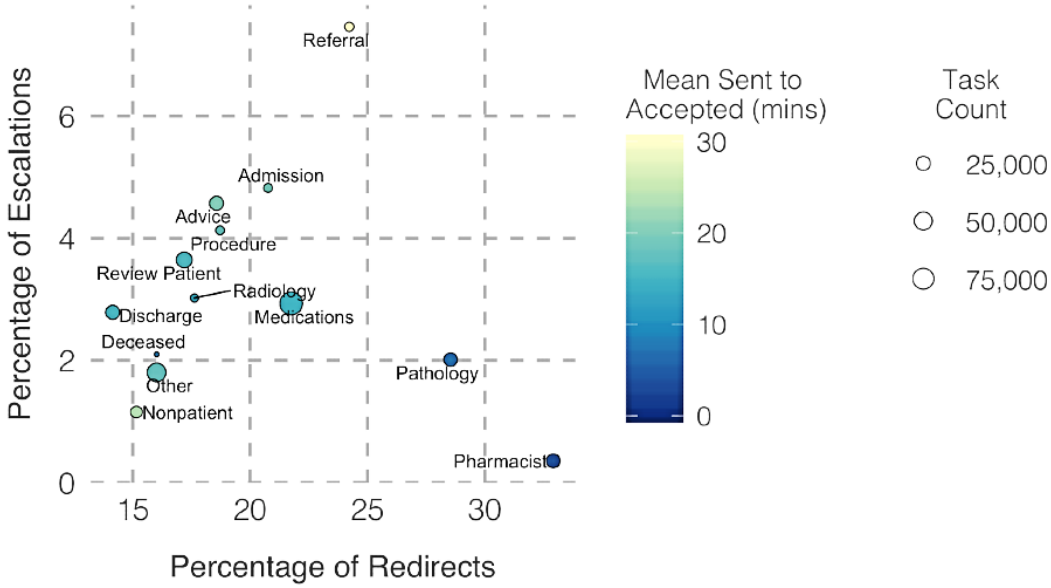
Figure 4. Variation in sent-to-accepted time across routine and urgent tasks.



Task Types

In Figure 5, we explore the variation in the percentage of redirects and escalated tasks against mean sent-to-accepted time using a high-level categorization of tasks. We note key deviations in task types such as Referral, Pathology, and Pharmacist.

Figure 5. Task count, redirects, escalations, and mean sent-to-accepted time for different task types.



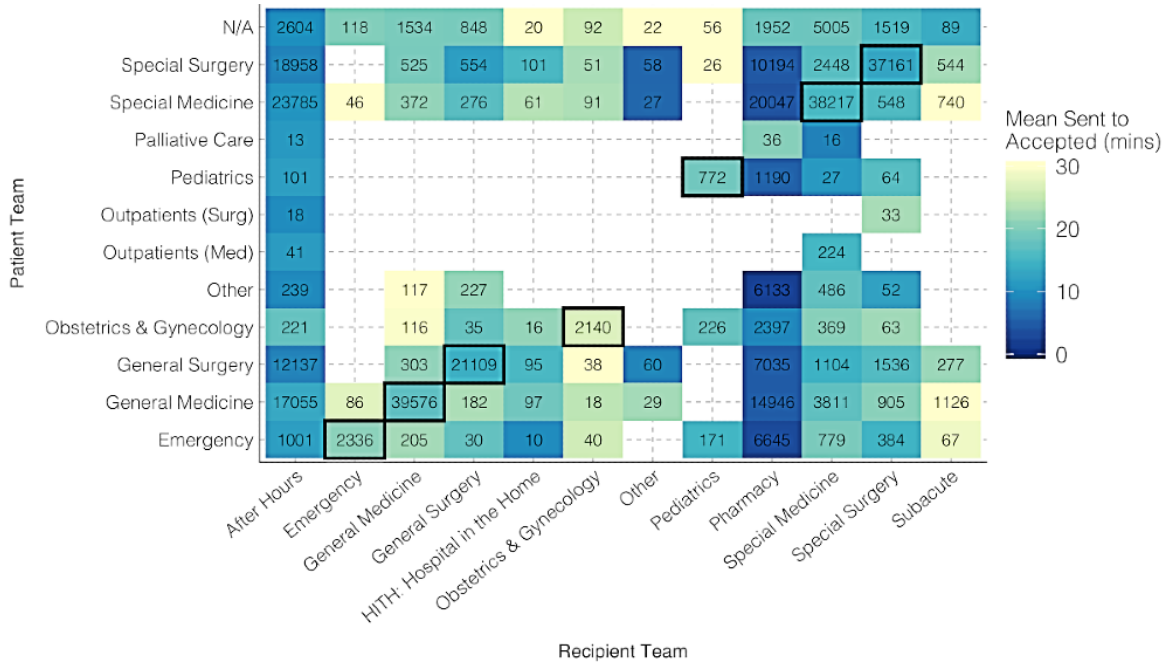
Patient and Recipient Team

To analyze the impact of patient and recipient teams on task acceptance, we created a heat map visualization of task flow, where the total number of tasks are given in each cell (Figure 6). Cells with less than 10 tasks were removed from the graph. We observed that task acceptance times vary depending on the Patient (ie, Sender) and Recipient teams.

To further investigate the team dynamics in the hospital, we categorized tasks as “internal” versus “external” based on surgical and medical wards in the hospital. If the patient team and recipient team were the same for a particular task, then that task was labelled as an internal task. Tasks originating from or received by teams that do not belong to either the medical or surgical categories were excluded from this analysis. A Wilcoxon rank-sum test showed that the sent-to-accepted time (minutes) is significantly higher (W=441,094,646; P<.001) for external assignments (mean 22.10, SD 91.45) when compared to internal assignments (mean 19.03, SD 82.66). Similarly, the sent-to-completed time (hours) was significantly higher (W=457,480,292; P<.001) for external assignments (mean 12.72, SD 42.71) when compared to internal assignments (mean 6.17, SD 22.90). We also observed a task escalation rate of 5.01% in external tasks and 3.45% in internal tasks. A chi-square test showed that external tasks are more likely to be escalated when compared to internal tasks (χ^2_1=38.72; P<.001).

XSL•FO
RenderX

Figure 6. Task flow from patient teams to recipient teams. Med: medical; surg: surgical.



Impact of Roles

The majority of tasks were received by hospital medical officers (HMOs; 119,582/317,372; 37.7%) and interns (94,869/317,372; 29.9%), whereas registrars received a smaller portion of tasks (23,041/317,372; 7.3%). In terms of senders, the majority of tasks were sent by nurses (187,487/317,372; 59.1%). Table 1 shows the variation in task acceptance across different recipient groups.

Table 1. Impact of the recipient level.

Recipient level	Mean sent-to-accepted time (minutes)	Mean sent-to-completed time (hours)	Redirects, %	Escalations, %
Intern	17.91	5.22	15.49	2.98
Hospital medical officer	15.60	4.54	15.83	2.87
Registrar	29.36	16.97	32.63	6.58

Discussion

MedTasker and Task Assignment

Task assignment and management is an important workflow aspect in hospitals that is not well supported by popular communication tools. In our work, we studied the use of MedTasker, a purpose-built task assignment app, and analyzed task assignment data collected over a period of over 22 months. As suggested through the usage trends in Figure 2, hospital staff comfortably adopted the MedTasker app within a year. Task acceptance and completion patterns throughout the day are proportionate to staff availability and reflect the standard hospital schedule where nurses change shifts at 7 AM, 3 PM, and 11 PM. Interestingly, we noted a high number of tasks being accepted at the beginning of the day shift (Figure 3). Although the total number of tasks accepted generally declines through the day, the number of tasks completed increases until the day shift ends.

Our analysis considers the tasks' sent-to-accepted time and the percentage of tasks that are escalated or redirected as key metrics in our evaluation. The analysis indicates significant variations in these important task metrics when we consider task type, task urgency, team, and sender roles. These variations highlight how task assignment data collected through MedTasker reflects the existing operational realities of a clinical environment.

Team Dynamics and Communication in the Hospital

A central theme that emerged from our results and interviews is the trust within and across teams. When examining task types and corresponding metrics in Figure 5, we observe different working patterns. Pharmacist and Pathology tasks have higher task redirection rates, indicating that task senders are not assigning them to the right person in the first instance. However, lower mean sent-to-accepted times suggest that such tasks are quickly redirected to a relevant team member and then accepted. This demonstrates how the specific teams that undertake these tasks function as efficient teams. In contrast, other core medical tasks like Review patient and Procedure are directed to the right person but are not accepted as fast as tasks like Pathology and Pharmacist. Core medical tasks also have relatively high escalation levels. Naturally, it is difficult to sort many core medical tasks into a well-defined task category as they could overlap with multiple categories. These different task acceptance paradigms also highlight the separation between tasks intended for a specific person (eg, Review patient) and tasks intended for a specific team (eg, Pathology). We observe the need for the software to support well-defined tasks and roles, as well as

XSL•FO
RenderX

for the hospital to have closely functioning teams to achieve high efficiency in clinical settings. Referral tasks appear as an outlier with high mean sent-to-accepted time, escalations, and redirections compared to other tasks. Although they are tasks sent among medical staff, they are generally treated as nonurgent tasks.

We obtained further insights about team dynamics by considering the task flow between different patient and recipient teams. Generally, we observed reasonable mean sent-to-accepted times across the majority of the team interactions while certain teams exhibit specific patterns. For example, we observed relatively higher mean sent-to-accepted times for tasks sent and received by the Obstetrics and Gynecology team (Figure 5). This team was slightly underresourced and stationed at a separate ward that is not well-connected with the rest of the hospital, decreasing the trust between teams. In addition, the Pharmacy team stands out with better mean sent-to-accepted times due to their tasks not being generally directed at a specific person. Another key observation is the separation between surgical and medical teams. Our results show that internal tasks or tasks sent among medical or surgical teams have significantly faster task acceptance and completion times and fewer escalations when compared to external tasks or tasks sent across teams. These observations highlight the need to facilitate intrateam and interteam connections and trust for operational efficiency.

Similarly, in terms of sender and recipient roles, we found higher redirect and escalation percentages and longer task acceptance times for tasks received by registrars in comparison to junior staff such as interns and HMOs (Table 1). This observation is expected since Registrars are experienced staff members, and tasks accepted by them are more complex and require expertise. In addition, the findings suggest that HMOs, who received the largest portion of tasks (119,582/317,372; 37.7%), are better organized and more efficient compared to others.

Redesigning Task Management Apps

Computer-based task management systems have the potential to benefit junior medical officers and nurses by improving overall task communication and achieving large reductions in time spent dealing with requests and walking between wards [19]. Early work on desktop computer–based task management systems includes TaskManager [12], a system implemented at Middlemore Hospital, Auckland, New Zealand. Their study shows that having the task management application connected to the hospital's Patient Management System increases the ease of task creation and results in effective communication. Similarly, a communication system deployed at Toronto General Hospital and Toronto Western Hospital, Toronto, Canada, involves a desktop-based physician handover tool and an SMS-based system that sends secure messages to a dedicated team smartphone [13]. A subsequent survey found that clinicians perceive that the system has a positive impact on efficiency and helped speed up daily work tasks. A more modern solution is Dock Health [14], a team collaboration and task management app that has been successfully adopted by medical staff at Boston Children's Hospital, Boston, MA, United States. Their HIPAA-compliant app runs on both mobile devices and web browsers and aims to overcome typical design and user experience issues in health care software.

In our case, our analysis has focused on a particular platform—MedTasker—but nevertheless, our analysis shows more broadly the kinds of meaningful insights that can be derived from data logs of task management apps. These insights can drive policy changes, which in turn can increase productivity in hospitals. Large volumes of historic data can also be used to build smart task management solutions that can optimally schedule tasks and alert when there are resource shortcomings. Additionally, important employee well-being surveys or feedback elements can be easily integrated into the task assignment app. Unlike when using off-the-shelf communication apps such as WhatsApp, by using an app like MedTasker, the hospital administration can have control over how data is governed and avoid security and privacy irregularities [6]. In addition, smartphones apps can be used to effectively communicate with and educate patients [20]. Such apps can be seamlessly integrated with task management apps. These unique advantages make purpose-built task assignment apps like MedTasker very appealing. We also point out that other forms of communication can complement app usage. In our case, hospital staff mentioned using phone calls for extended detailed conversations that mainly involve administrative work, WhatsApp for social collaboration and to notify regarding nonclinical events, and paging systems or speaker announcements for emergencies. Face-to-face communication also regularly occurs within wards but was not captured in this study.

Based on our observations and discussions with hospital staff, we discuss several improvements to MedTasker that we aim to implement in the future. These enhancements are also important to consider when implementing similar task assignment apps for health care. First, our results show that reminders have a strong influence on task acceptance. As opposed to using static time limits, adaptive time limits can be used to send reminders. It is also possible to incorporate workload information, such that reminders are adjusted based on recipients' ongoing workload.

Second, the current practice of creating a task with an urgent or nonurgent label is arbitrary and highly dependent on the individual who creates the task. Since the prespecified task urgency has a significant impact on task acceptance, it is important that this particular label is added appropriately. Future implementations could help task requesters by automatically suggesting the appropriate urgency label based on task information. In addition, we propose user interface improvements that direct attention to urgent tasks when users receive multiple tasks.

Third, our results suggest that trust is important for efficient task assignment. To facilitate trust among teams and individuals, task assignment apps could provide more information regarding users. Contextual information such as current workload information or location can be helpful. Profile photos and other elements are also useful in increasing the levels of image appeal and perceived social presence, which in turn can increase trust [21].

XSL•FO
RenderX

Furthermore, MedTasker is not currently integrated with My Health Record (a major national electronic health record initiative in Australia) or any other external systems such as insurance systems. Although such integrations can be facilitated to further improve patient care, they should be implemented based on policy decisions and guidelines provided by hospitals and regulators.

Limitations

We note several limitations in our work. First, we use task acceptance time and completion time from the data set in our analysis. However, there may be some tasks in which acceptance and completion times recorded through the app do not correspond to actual task times. For example, staff may not immediately indicate task completion when they attend to a series of tasks or experience internet connectivity issues. Second, we acknowledge that different hospitals may operate under different protocols and practices, and as such, it is important to be cautious about extrapolating our findings to other or all hospitals, especially in different countries.

Conclusion

We analyzed hospital task assignment data collected via MedTasker, a dedicated task assignment app deployed at a hospital over 22 months. We show that important insights into how teams function in a clinical setting can be readily drawn from task assignment data. Our analysis shows that predefined labels such as urgency and task type are important and impact how tasks are accepted and completed. We also show how task acceptance varies across teams and roles and highlight that internal tasks are more efficiently managed than external tasks, possibly due to increased trust among team members. Finally, we discuss how smartphone-based task assignment apps can be further improved to support clinical work and staff.

Acknowledgments

This work is partially funded by Australian Research Council Discovery Project DP190102627 and National Health and Medical Research Council grants 1170937 and 2004316.

Authors' Contributions

DH, LH, JG, and VK designed the research. LH collected the data. DH, VK, and JG analyzed the data. LH and VK verified and interpreted the data. DH, JG, and VK contributed to manuscript preparation.

Conflicts of Interest

None declared.

References

1. Nikolic A, Wickramasinghe N, Claydon-Platt D, Balakrishnan V, Smart P. The Use of Communication Apps by Medical Staff in the Australian Health Care System: Survey Study on Prevalence and Use. JMIR Med Inform 2018 Feb 09;6(1):e9 [FREE Full text] [doi: 10.2196/medinform.9526] [Medline: 29426813]
2. Ganasegeran K, Renganathan P, Rashid A, Al-Dubai SAR. The m-Health revolution: Exploring perceived benefits of WhatsApp use in clinical practice. Int J Med Inform 2017 Jan;97:145-151. [doi: 10.1016/j.ijmedinf.2016.10.013] [Medline: 27919374]
3. Johnston MJ, King D, Arora S, Behar N, Athanasiou T, Sevdalis N, et al. Smartphones let surgeons know WhatsApp: an analysis of communication in emergency surgical teams. Am J Surg 2015 Jan;209(1):45-51. [doi: 10.1016/j.amjsurg.2014.08.030] [Medline: 25454952]
4. Nardo B, Cannistrà M, Diaco V, Naso A, Novello M, Zullo A, et al. Optimizing Patient Surgical Management Using WhatsApp Application in the Italian Healthcare System. Telemed J E Health 2016 Sep;22(9):718-725. [doi: 10.1089/tmj.2015.0219] [Medline: 27027211]
5. Ellanti P, Moriarty A, Coughlan F, McCarthy T. The Use of WhatsApp Smartphone Messaging Improves Communication Efficiency within an Orthopaedic Surgery Team. Cureus 2017 Mar 18;9(2):e1040 [FREE Full text] [doi: 10.7759/cureus.1040] [Medline: 28357172]
6. Thomas K. Wanted: a WhatsApp alternative for clinicians. BMJ 2018 Feb 12;360:k622. [doi: 10.1136/bmj.k622] [Medline: 29440047]
7. Freundlich RE, Freundlich KL, Drolet BC. Pagers, Smartphones, and HIPAA: Finding the Best Solution for Electronic Communication of Protected Health Information. J Med Syst 2017 Dec 25;42(1):9. [doi: 10.1007/s10916-017-0870-9] [Medline: 29177600]
8. Khanna RR, Wachter RM, Blum M. Reimagining Electronic Clinical Communication in the Post-Pager, Smartphone Era. JAMA 2016 Jan 05;315(1):21-22. [doi: 10.1001/jama.2015.17025] [Medline: 26746450]
9. Patel B, Johnston M, Cookson N, King D, Arora S, Darzi A. Interprofessional Communication of Clinicians Using a Mobile Phone App: A Randomized Crossover Trial Using Simulated Patients. J Med Internet Res 2016 Apr 06;18(4):e79 [FREE Full text] [doi: 10.2196/jmir.4854] [Medline: 27052694]

XSL•FO
RenderX

10. Morton J, Williams Y, Philpott M. New Zealand's Christchurch Hospital at night: an audit of medical activity from 2230 to 0800 hours. N Z Med J 2006 Mar 31;119(1231):U1916. [Medline: 16582976]
11. Wu R, Rossos P, Quan S, Reeves S, Lo V, Wong B, et al. An evaluation of the use of smartphones to communicate between clinicians: a mixed-methods study. J Med Internet Res 2011 Aug 29;13(3):e59 [FREE Full text] [doi: 10.2196/jmir.1655] [Medline: 21875849]
12. Seddon ME, Hay D. Task Manager: an innovative approach to improving hospital communication after hours. N Z Med J 2010 Oct 15;123(1324):57-66. [Medline: 20953223]
13. Wu R, Lo V, Morra D, Appel E, Arany T, Curiale B, et al. A smartphone-enabled communication system to improve hospital communication: usage and perceptions of medical trainees and nurses on general internal medicine wards. J Hosp Med 2015 Feb;10(2):83-89. [doi: 10.1002/jhm.2278] [Medline: 25352429]
14. Morsy A. Clinician App Tackles Stress of Patient Task Management. IEEE Pulse 2019;10(2):28-30. [doi: 10.1109/MPULS.2019.2899705] [Medline: 31021755]
15. Wu RC, Lo V, Morra D, Wong BM, Sargeant R, Locke K, et al. The intended and unintended consequences of communication systems on general internal medicine inpatient care delivery: a prospective observational case study of five teaching hospitals. J Am Med Inform Assoc 2013;20(4):766-777 [FREE Full text] [doi: 10.1136/amiajnl-2012-001160] [Medline: 23355461]
16. Hribar MR, Read-Brown S, Reznick L, Lombardi L, Parikh M, Yackel TR, et al. Secondary Use of EHR Timestamp data: Validation and Application for Workflow Optimization. In: AMIA Annu Symp Proc. 2015 Presented at: 2015 AMIA Symposium; November 5, 2015; San Francisco, CA, USA p. 1909-1917 URL: http://europepmc.org/abstract/MED/26958290
17. Rucker DW. Using telephony data to facilitate discovery of clinical workflows. Appl Clin Inform 2017 Apr 19;8(2):381-395 [FREE Full text] [doi: 10.4338/ACI-2016-11-RA-0191] [Medline: 28421225]
18. MedTasker. Medtasker - Mobile CommunicationTask Management for Hospitals. URL: https://medtasker.com/ [accessed 2021-01-05]
19. Whitehead NJ, Maharaj ON, Agrez M. A computer-based task management system for junior medical officers during after-hours shifts. Med J Aust 2014 Feb 03;200(2):79-80. [doi: 10.5694/mja13.10884] [Medline: 24484096]
20. Timmers T, Janssen L, van der Weegen W, Das D, Marijnissen W, Hannink G, et al. The Effect of an App for Day-to-Day Postoperative Care Education on Patients With Total Knee Replacement: Randomized Controlled Trial. JMIR mHealth uHealth 2019 Oct 21;7(10):e15323 [FREE Full text] [doi: 10.2196/15323] [Medline: 31638594]
21. Hassanein K, Head M. Manipulating perceived social presence through the web interface and its impact on attitude towards online shopping. International Journal of Human-Computer Studies 2007 Aug;65(8):689-708. [doi: 10.1016/j.ijhcs.2006.11.018]

Abbreviations

EMR: electronic medical record
HIPAA: Health Insurance Portability and Accountability Act
HMO: hospital medical officer
PHI: protected health information

Edited by C Lovis; submitted 25.02.21; peer-reviewed by P Athilingam, S Rostam Niakan Kalhori; comments to author 07.05.21; revised version received 01.07.21; accepted 10.07.21; published 16.08.21.

Please cite as:
Hettiachchi D, Hayes L, Goncalves J, Kostakos V
Team Dynamics in Hospital Workflows: An Exploratory Study of a Smartphone Task Manager
JMIR Med Inform 2021;9(8):e28245
URL: https://medinform.jmir.org/2021/8/e28245
doi:10.2196/28245
PMID:34398797

©Danula Hettiachchi, Lachie Hayes, Jorge Goncalves, Vassilis Kostakos. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 16.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Foodborne Disease Risk Prediction Using Multigraph Structural Long Short-term Memory Networks: Algorithm Design and Validation Study

Yi Du[1,2], PhD; Hanxue Wang[1,2], MEng; Wenjuan Cui[1], PhD; Hengshu Zhu[3], PhD; Yunchang Guo[4], PhD; Fayaz Ali Dharejo[1,2], PhD; Yuanchun Zhou[1,2], PhD

[1]Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
[2]Chinese Academy of Sciences University, Beijing, China
[3]Baidu Inc, Beijing, China
[4]China National Center for Food Safety Risk Assessment, Beijing, China

Corresponding Author:
Yi Du, PhD
Computer Network Information Center
Chinese Academy of Sciences
Information Technology Building of Chinese Academy of Sciences
No. 2 Dongsheng South Road, Zhongguancun, Haidian District
Beijing, 100089
China
Phone: 86 15810134970
Email: duyi@cnic.cn

Abstract

Background: Foodborne disease is a common threat to human health worldwide, leading to millions of deaths every year. Thus, the accurate prediction foodborne disease risk is very urgent and of great importance for public health management.

Objective: We aimed to design a spatial–temporal risk prediction model suitable for predicting foodborne disease risks in various regions, to provide guidance for the prevention and control of foodborne diseases.

Methods: We designed a novel end-to-end framework to predict foodborne disease risk by using a multigraph structural long short-term memory neural network, which can utilize an encoder–decoder to achieve multistep prediction. In particular, to capture multiple spatial correlations, we divided regions by administrative area and constructed adjacent graphs with metrics that included region proximity, historical data similarity, regional function similarity, and exposure food similarity. We also integrated an attention mechanism in both spatial and temporal dimensions, as well as external factors, to refine prediction accuracy. We validated our model with a long-term real-world foodborne disease data set, comprising data from 2015 to 2019 from multiple provinces in China.

Results: Our model can achieve F1 scores of 0.822, 0.679, 0.709, and 0.720 for single-month forecasts for the provinces of Beijing, Zhejiang, Shanxi and Hebei, respectively, and the highest F1 score was 20% higher than the best results of the other models. The experimental results clearly demonstrated that our approach can outperform other state-of-the-art models, with a margin.

Conclusions: The spatial–temporal risk prediction model can take into account the spatial–temporal characteristics of foodborne disease data and accurately determine future disease spatial–temporal risks, thereby providing support for the prevention and risk assessment of foodborne disease.

(JMIR Med Inform 2021;9(8):e29433) doi:10.2196/29433

KEYWORDS

foodborne disease; risk; prediction; spatial–temporal data

XSL•FO
RenderX

Introduction

Foodborne disease is caused by pathogenic bacteria that enter the body due to ingestion of contaminated food, resulting in symptoms such as diarrhea and abdominal pain [1]. According to the World Health Organization, more than 600 million people worldwide suffer from diseases caused by contaminated food every year, of whom 4.2 million die of foodborne illness [2]. The high incidence of foodborne diseases seriously threatens health and social economy. Most existing research efforts on foodborne disease have mostly been concentrated in the fields of medical science and food safety [3-6]; however, researchers have turned their attention to exploiting machine learning technologies to address foodborne disease–related topics, such as analyzing the correlation between foodborne diseases and food [7], discovering foodborne disease outbreak locations using social media [8-10], analyzing foodborne disease pathogens [11,12], and predicting foodborne disease outbreaks [13-15]. While considerable efforts have been made, an open challenge remains—accurately predicting foodborne disease risk by mining spatial–temporal patterns in historical disease records, using similar methods to those used for flu prediction [16-18], which is of great significance for public health management. By providing estimates of the trends of foodborne disease in future periods, accurate foodborne disease risk prediction can support effective guidance for government epidemic prevention policies. Because foodborne disease risk usually follows a certain spatial–temporal pattern—for example, the incidence in summer is higher than those in autumn and winter, and risk of foodborne diseases in a region is similar to those in regions with similar weather or urban functional structure—the prediction of foodborne disease risk can be solved as a spatial–temporal data modeling problem.

In the literature, a variety of methods for spatial–temporal data modeling have been proposed, including traditional statistical models [19,20] and deep learning methods, such as recurrent neural network [21], long short-term memory (LSTM) [22], convolutional neural network [23], graph convolutional network [24], temporal graph convolutional network [25], and structural recurrent neural network [26]. To solve the problem of spatial–temporal data modeling, structural recurrent neural networks use recurrent neural networks to model temporal dependence and model spatial dependence with structural recurrent neural networks on spatial–temporal graphs. Such models possess scalability; however, models are limited to static representations of spatial dependence by region proximity (ie, the models lack dynamic spatial correlation representation).

Compared with COVID-19 [27], influenza [16-18], and other infectious diseases [28], foodborne disease is spread through food rather than people. Therefore, the data characteristics of foodborne disease outbreaks are quite different from those related to infectious diseases, for example, sparse data increase the difficulty of predicting foodborne disease risk. Foodborne disease risk prediction also differs from traffic prediction [25,29-33]. Traffic problems require short-term prediction, while foodborne disease risk problems require long-term prediction.

To address these challenges, in this paper, we propose the use of a multigraph structural LSTM based spatial–temporal prediction model to determine the risk of foodborne disease in different regions in future periods, which considers various spatial dependencies and uses a dynamic fusion method, with multistep prediction using a encoder–decoder structure, to support future disease prevention and control, and with attention mechanisms in spatial and temporal dimensions, as well as external features, to further improve performance. To the best of our knowledge, this is the first study to focus on spatial–temporal foodborne disease risk prediction and report validation results using real-world data sets.

We propose a multistep spatial–temporal data prediction model based on encoder–decoder structure and composed entirely of LSTM modules, to address the problem of spatial–temporal foodborne disease risk prediction; we propose a dynamic fusion method to fuse region proximity, historical trend similarity, regional function similarity and food exposure similarity, with a spatial–temporal attention mechanism and external feature embedding; and we validated our model with extensive experiments on a long-term real-world foodborne disease data set, with data from 2015 to 2019 in multiple provinces of China; experimental results clearly demonstrated that our approach can outperform other state-of-the-art methods, with a margin.

Methods

Problem Definition

Region Graph

We divide each city or region into irregular subregions by administrative areas and organized them into an undirected graph $\boldsymbol{G}=(\boldsymbol{v}, \boldsymbol{e}, A)$, where v is a set of nodes and each node corresponds to a subregion, e is a set of edges with each edge connecting 2 subregions defined by some rules, and A represents the adjacency matrix of $\boldsymbol{G}$. In particular, each v_i in $v=(v_1, v_2,...v_n)$ is the minimal spatial unit, where N is the total number of spatial units, and e_{ij} is the edge that connects v_i and v_j.

Historical Data Sequence

To represent the historical data sequence, we calculated the number of disease records at each prediction period, that is, given a subregion v_i, we defined the sequence of counts to denote the historical data sequence in subregion v_i during the time window T.

Spatial–Temporal Graph

To represent spatial–temporal data characteristics, we organized the historical data sequence and the spatial graph into spatial–temporal graphs. Foodborne disease data at timestep t in a subregion is represented as graph signal , and the entire spatial–temporal graph is represented as

Disease Risk

To evaluate the predicted disease risk intuitively, we divided each region's disease record count into 2 classes using a ratio, which we determined by consulting domain experts: when the

XSL•FO
RenderX

disease record count in a region at any given timestep exceeds 70% of the historical sequence of this region, the risk at that timestep for that region is considered high risk or low risk.

Disease Risk Prediction

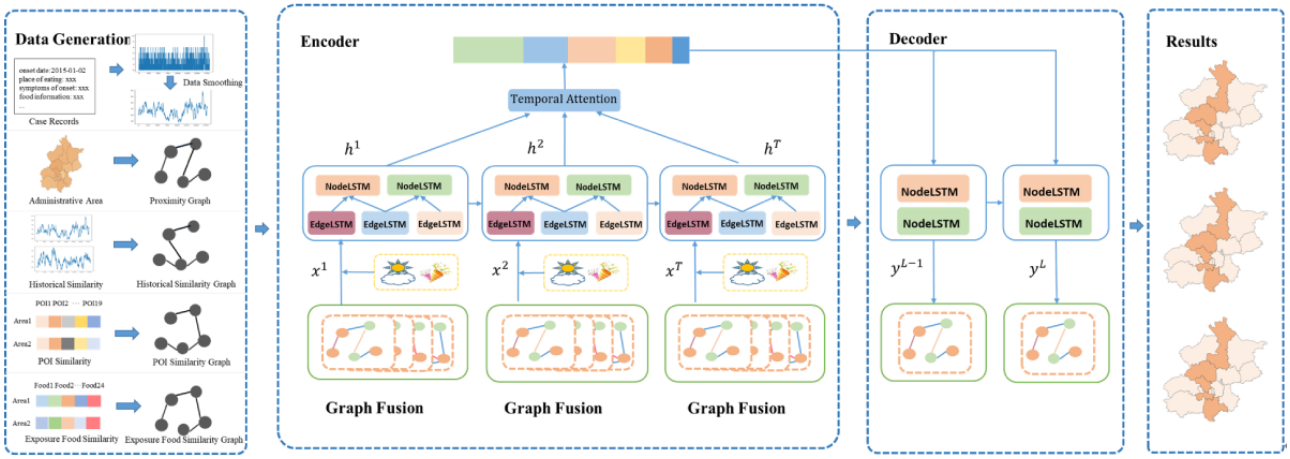
The risk of foodborne disease in a region is affected by its historical data and by the risk of surrounding area and is, therefore, a spatial–temporal prediction problem. Given the historical disease record data from subregions v during time period $\boldsymbol{T}$, our task was to determine the unknown disease risk level for each subregion in future time slots $\boldsymbol{L}$. Formally, our aim was to compute the following:

Model Framework

Model Overview

Our model is an encoder–decoder multigraph structural LSTM (Figure 1). This model consists of 5 modules. The *Data Generation* module comprises temporal sequence and multiple spatial graph (geographic proximity, historical data similarity, regional functional similarity, and foodborne disease exposure food similarity) data processing. The *Multigraph Fusion* module takes into account multiple spatial correlations and merges them dynamically. The *Encoder–Decoder* module uses LSTM networks to model temporal dependence and spatial dependence of foodborne disease risk by using the edge LSTM and the node LSTM, respectively, simultaneously in the encoder. In the decoder, the node LSTM is used to predict foodborne disease risk in each region in the 1 or more future timesteps. The *Spatial–Temporal Attention* module takes spatial–temporal relationship complexity into account and assigns temporal importance values to timesteps and spatial importance values to adjacent edges of nodes. The *External Feature Embedding* module combines various external features (eg, holidays, temperature) and merges external features into the encoder at each timestep.

Figure 1. Foodborne disease spatial–temporal risk prediction model framework. LSTM: long short-term memory; POI: point of interest.



Data Generation

This module performs data processing of temporal sequence and multiple spatial graph data (geographic proximity, historical data similarity, regional functional similarity, and foodborne disease exposure food similarity).

Temporal sequence data were collected from historical foodborne disease records, from which disease record counts were calculated. Due to the sparseness of data, we performed data augmentation, with a sliding 1-month window by moving the start of the unnatural month, which resulted in an expansion of the data. Temporal sequence data were normalized (range 0-1), using minimum–maximum normalization.

Data were characterized by regional proximity because, intuitively, adjacent regions will have similarity risks of disease due to climate and geography, as well as from population movement between regions. For graph $\boldsymbol{G}$=($\boldsymbol{v}$, $\boldsymbol{e}$, A), if v_i and v_j are spatially adjacent, then is 1, otherwise is 0.

For each region, disease risk trends will follow a relatively fixed pattern, and regions with similar historical disease risk trends will have similar disease risk trends in future periods. We used historical data sequence to calculate the pairwise historical similarities between regions using Pearson correlation coefficients. We set a threshold; the adjacency value between 2 nodes v_i and v_j with a similarity less than the threshold is 0. The threshold is used to control the sparsity of edges.

Regions with similar urban functions will have similar population and business structures, and thus, similar foodborne disease risk. We used point-of-interest (POI) data from each region to characterize this feature. POI can be divided by function into 19 categories, the term frequency–inverse document frequency can be used to embed these data as vectors for every region, and the similarity between of POI vectors for regions can be evaluated [34].

XSL•FO
RenderX

Exposure food, the transmission medium of foodborne disease, plays an important role in the prediction of foodborne disease risk. Intuitively, exposure to foodborne diseases at different timesteps and in different regions are different, and the impact on the risk of foodborne diseases is also different. Therefore, we counted the number of exposures for each food category (23 categories) in different regions at different timesteps, which were represented as vectors using term frequency–inverse document frequency. Similarities between exposure vectors for regions at each timestep were calculated, representing spatial correlations.

Multigraph Fusion

Our dynamic fusion method, for multiple spatial graphs constructed by different spatial correlations, was designed to merge adjacent matrices $\{A^1, A^2...A^m\}$, where m represents the number of constructed graphs. We defined 4 parameters, W_1, W_2, W_3, W_4, and to obtain the dynamic merged graph, element-wise products between the parameters and adjacent matrices tare calculated to adjust the weights of the geographic proximity, historical data similarity, region functional similarity, and exposure food similarity graphs.

The parameters are continuously adjusted, through network learning, to control the influences of multiple spatial dependencies on the final inputs.

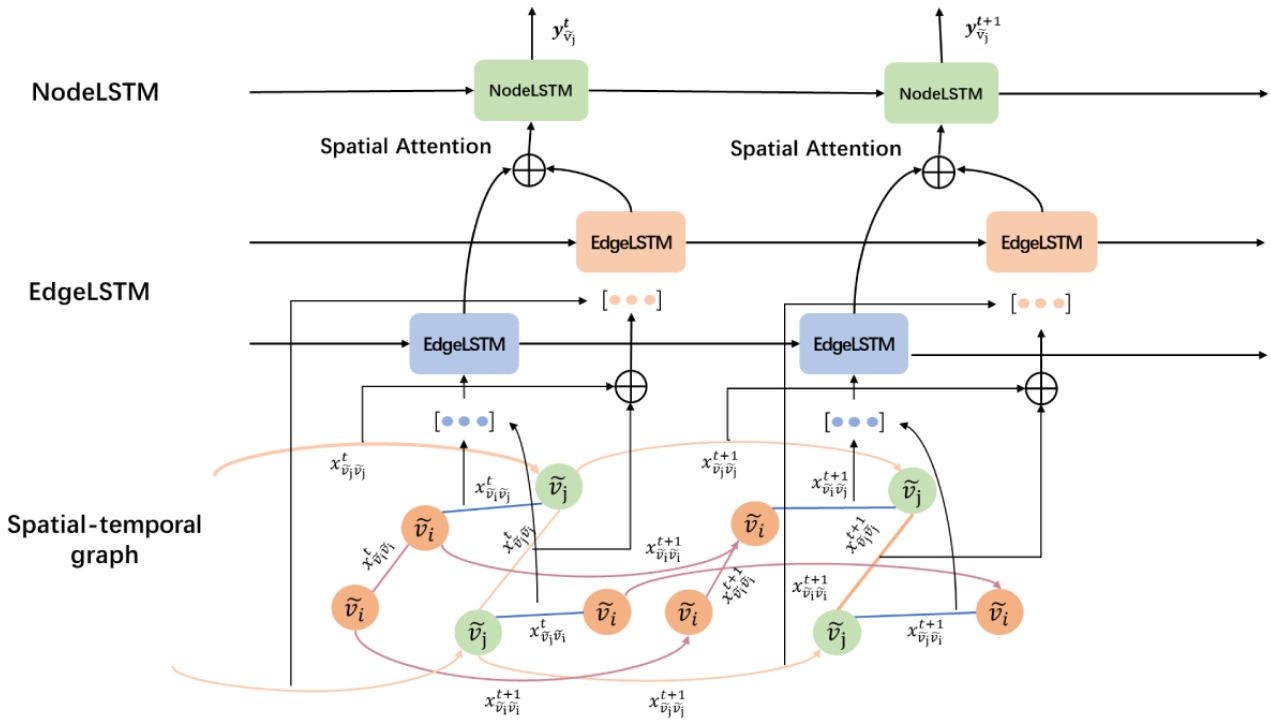
Encoder–Decoder

In order to model spatial dependence and temporal dependence simultaneously and conduct multistep prediction, we organize the historical temporal sequence data and the fused spatial graph into the structure of spatial–temporal graph and construct a graph structural LSTM model of encoder–decoder architecture inspired by the structural recurrent neural network architecture [26].

In the encoder, a structural LSTM network (Figure 2) was constructed with node LSTMs and edge LSTMs to model temporal dependence and spatial dependence. We divide nodes $v=(v_1, v_2,...v_n)$ on the spatial graph into 2 categories in a ratio according to the sum of values of each node at all timesteps in the temporal dimension. The edges between nodes were divided into 3 categories, according to connected nodes. Then, we constructed node LSTMs and edge LSTMs for each category of nodes and each category of edges (Figure 3). For each edge LSTM, the input at each timestep was the concatenation of the current node values connected by the edges of its category, and for each node LSTM, the input at each timestep was the fusion of the current outputs of edge LSTMs related to its node category. It not only contained the information of the current category of nodes but also contained the information of adjacent node categories to model spatial dependence. The current state of the node LSTM and edge LSTM was not only influenced by the current input, but also by the previous timesteps, to model temporal dependence.

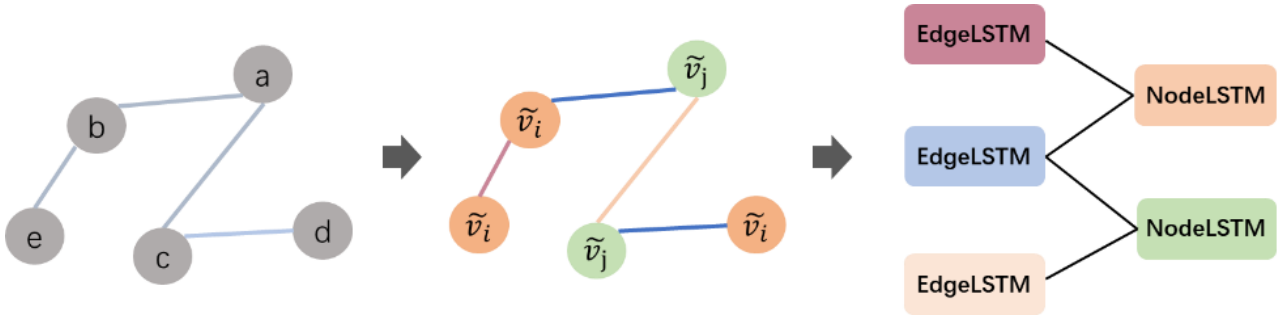
In the decoder, for each node LSTM, we used the context vector learned from the encoder to predict the value of 1 or more timesteps in the future.

Figure 2. Structural long short-term memory (LSTM) details.



XSL•FO
RenderX

Figure 3. Grouping nodes and edges in long short-term memory (LSTM) networks.



Spatial-Temporal Attention

In order to eliminate the influence of distance on temporal dependence, and to fully consider temporal and spatial correlations, we applied a spatial-temporal attention mechanism. In the temporal dimension, we calculate the score between hidden states with current spatial–temporal state, transformed into a normalized value with softmax operation, then apply a weighted summarization as

In the spatial dimension, we calculate the score of each edge LSTM, normalized by softmax to assign different weight to different edge LSTM every timestep.

External Feature Embedding

The risk of foodborne disease may be influenced by the change of external factors (for example, people eating out on holidays more often than working days, or high temperature and humid weather being more likely to cause food spoilage). Therefore, to incorporate external features into our model, we first preprocess temperature data by filling the missing value and computing the mean value for a month. For the holiday feature, we calculated the number of holidays per month, which was represented as a series of fixed-length vectors and concatenated with the input sequence of node LSTMs in previous timesteps to predict the future disease risk.

Model Validation

Data Set

We validated our model using a real-world data set (China National Center for Food Safety Risk Assessment [35]), which consisted of foodborne disease records reported by sentinel hospitals in almost all provinces in China. Each record contains information such as time of onset, place of eating, place of living, symptoms of onset, and food information. We selected all the records in the 4 provinces with best-quality data from 2015 to 2019—Beijing, Zhejiang, Shanxi, and Hebei. Due to data acquisition limitations, we only obtain the POI information for Beijing. Therefore, only 3 spatial dependencies were used for Zhejiang, Shanxi, and Hebei. We collected temperature data and holiday data from 2015 to 2019 to simulate the impact of weather and holiday on the foodborne disease risk.

Comparison Models and Evaluation Metrics

We compared our model with historical average, autoregressive, ARIMA (autoregressive integrated moving average), LSTM, and spatial–temporal graph convolutional network models. Historical average models estimate future results by computing the average value of historical data, which is too simple to model spatial-temporal dependence. Autoregressive models are statistical time-series models that use a linear combination of the values of several previous timesteps to describe future values. ARIMA models, which as the name implies, use autoregressive terms and moving average terms. Data must be processed before applying the ARIMA model to ensure that data are stationary. LSTM networks are mostly used for natural language processing problems [22]. LSTM networks can learn sequence dependence due to its chain structure. We applied LSTM to every node of the graph and evaluated the model by merging the results of all nodes. Spatial–temporal graph convolutional network models are based on convolutional neural networks but use graph convolutional networks instead of traditional convolutional neural networks for spatial dimensions and temporal convolutional neural network instead of recurrent neural networks for temporal dimensions. Spatial–temporal graph convolutional network models have achieved outstanding results in traffic prediction [31].

Given that we used a binary definition of disease risk, to avoid the effect of imbalances between 2 classes, we used

to evaluate model performance. In order to avoid the effect of parameter initialization on the results, we performed 5 trials for each model and averaged the results.

Results

Performance Comparison

Comparison With Other Methods

Table 1 and Figure 4 summarize foodborne disease risk prediction performance results for 1, 2, and 3 months in each of the 4 provinces. Our proposed model outperformed all other models for all 4 provinces and achieved the highest F1 score for every forecast period. Traditional statistical models, such as autoregressive and ARIMA models, performed worse than deep learning models for most provinces, indicating that traditional methods were too simple to solve complex nonlinear spatiotemporal problems. LSTM networks modeled the temporal

XSL•FO
RenderX

dependence of each node on the spatial–temporal graph independently and ignored the dynamic spatial correlation between nodes, resulting in relatively poor performance. The spatial–temporal graph convolutional network model used convolution neural networks to model temporal dependence as well as spatial dependence, with better performance than that of the LSTM model for most provinces. Our proposed method with a single graph (that is, a regional proximity graph) simulated temporal dependence and spatial dependence simultaneously with a reasonable attention mechanism, resulting in better performance than those of the other methods. At most timesteps, it had the second-best prediction results. By accounting for rich spatial dependencies, our multigraph model exhibited better performance than that of the single-graph model for all 4 provinces, achieving the best results. The highest F1 score was 20% higher than the best results of the other models.

Table 1. Performance of different models using data from 4 provinces.

Province and forecast period	Model						
	Historical average	AR[a]	ARIMA[b]	LSTM[c]	ST-GCN[d]	Ours (single graph)	Ours (multigraph)
	F1 score	F1 score	F1 score	F1 score, mean (SD)	F1 score, mean (SD)	F1 score, mean (SD)	F1 score, mean (SD)
Beijing							
1-month prediction	0.679	0.742	0.734	0.750 (0.007)	0.777 (0.034)	0.811 (0.014)	0.822 (0.011)
2-month prediction	0.675	0.741	0.720	0.744 (0.012)	0.737 (0.023)	0.785 (0.007)	0.812 (0.017)
3-month prediction	0.674	0.733	0.664	0.743 (0.019)	0.724 (0.041)	0.768 (0.011)	0.805 (0.021)
Zhejiang							
1-month prediction	0.484	0.597	0.558	0.551 (0.021)	0.651 (0.026)	0.648 (0.021)	0.679 (0.009)
2-month prediction	0.471	0.562	0.474	0.501 (0.017)	0.604 (0.031)	0.630 (0.019)	0.660 (0.012)
3-month prediction	0.457	0.531	0.404	0.441 (0.015)	0.544 (0.029)	0.603 (0.020)	0.645 (0.008)
Shanxi							
1-month prediction	0.373	0.559	0.390	0.550 (0.022)	0.582 (0.045)	0.677 (0.011)	0.709 (0.013)
2-month prediction	0.369	0.548	0.314	0.549 (0.027)	0.583 (0.039)	0.684 (0.015)	0.699 (0.019)
3-month prediction	0.366	0.541	0.246	0.542 (0.017)	0.585 (0.043)	0.683 (0.012)	0.695 (0.017)
Hebei							
1-month prediction	0.682	0.632	0.531	0.553 (0.018)	0.449 (0.027)	0.692 (0.005)	0.720 (0.006)
2-month prediction	0.675	0.616	0.494	0.532 (0.016)	0.445 (0.048)	0.683 (0.012)	0.703 (0.010)
3-month prediction	0.666	0.593	0.452	0.513 (0.020)	0.392 (0.033)	0.668 (0.007)	0.698 (0.012)

[a]AR: autoregressive.

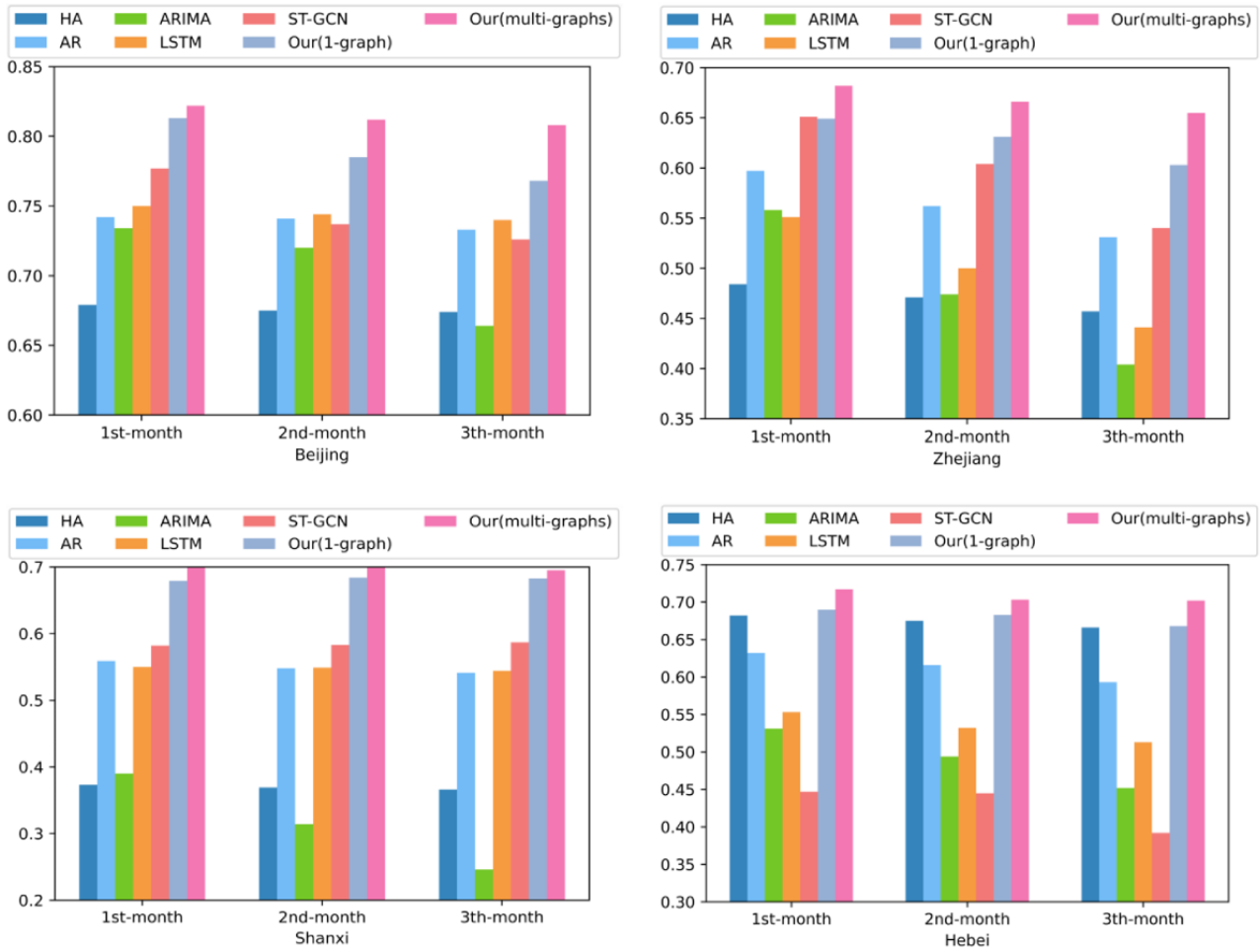
[b]ARIMA: autoregressive integrated moving average.

[c]LSTM: long short-term memory.

[d]ST-GCN: spatial–temporal graph convolutional network.

XSL•FO
RenderX

Figure 4. Performance in 4 provinces. AR: autoregressive; ARIMA: autoregressive integrated moving average; HA: historical average; LSTM: long short-term memory; ST-GCN: spatial–temporal graph convolutional network.



Effect of Spatial Dependence

The results of the Beijing data set, using 4 different spatial graphs to represent spatial dependence between regions and multiple spatial graph fusion (Table 2), demonstrate that different spatial dependence affects prediction: single spatial dependence is not as effective as the fusion of multiple dependencies.

Table 2. Performance of models with different spatial dependencies.

Model type	F1 score		
	1-month prediction	2-month prediction	3-month prediction
Single-graph			
Proximity	0.813	0.785	0.768
Time series similarity	0.800	0.776	0.732
POI similarity	0.797	0.705	0.741
Exposure food similarity	0.813	0.756	0.743
Multigraph	0.822	0.812	0.805

Effect of External Features

Using the Beijing data set, the performance of models with external features is slightly better than those of models without external features for 1-, 2-, and 3-month predictions (Table 3), which demonstrates that the external features affect the trend of foodborne disease to some extent.

XSL•FO
RenderX

Table 3. Performance of models with or without external features.

Model type	F1 score		
	1-month prediction	2-month prediction	3-month prediction
External features	0.818	0.810	0.803
No external features	0.822	0.812	0.805

Effect of Attention Mechanism

For the Beijing data set, the removal of the attention mechanism in the spatial dimension or in the temporal dimension reduced the effectiveness of the model (Table 4). With the removal of the attention mechanism in the temporal dimension, as the prediction range increased, model performance decreased. This also confirms that, in the multistep prediction, the use of an attention mechanism can solve the distance problem in sequence dependence.

Table 4. Performance of models with or without an attention mechanism.

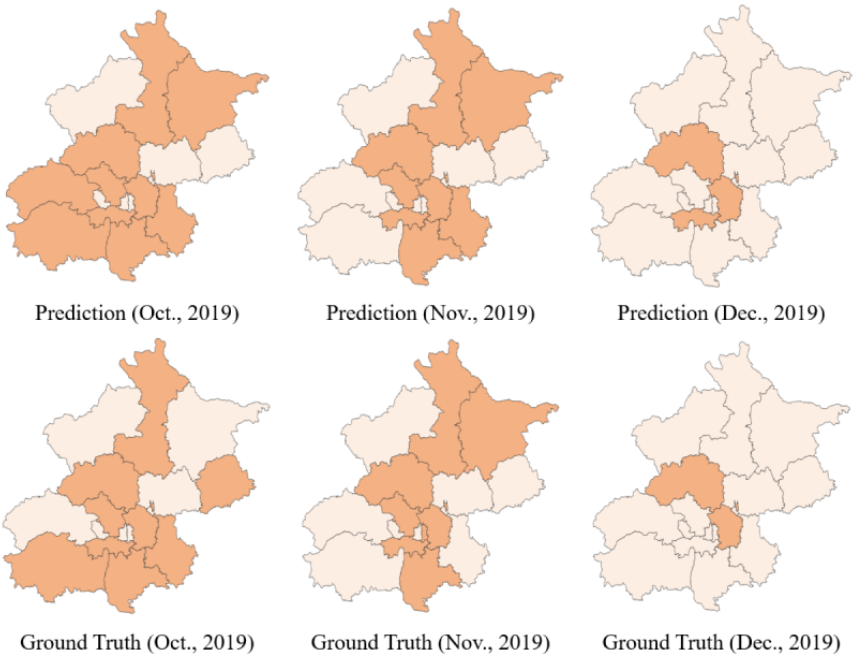
Model type	F1 score		
	1-month prediction	2-month prediction	3-month prediction
Spatial attention only	0.815	0.801	0.788
Temporal attention only	0.807	0.805	0.798
With attention mechanism	0.822	0.812	0.805

Mapped Results

We selected 3 consecutive months in the Beijing data set (October, November, and December 2019), for which we mapped the predicted values and the ground truths (Figure 5). Disease risks in most regions were correctly predicted, and only 1 or 2 regions had incorrect predictions for each prediction range. Incorrect predictions were often affected by the value of the surrounding region, which is also consistent with clustered outbreak characteristics of foodborne diseases. To a certain extent, this case suggests that our model is able to capture the spatial–temporal correlations between data and can provide accurate multistep prediction.

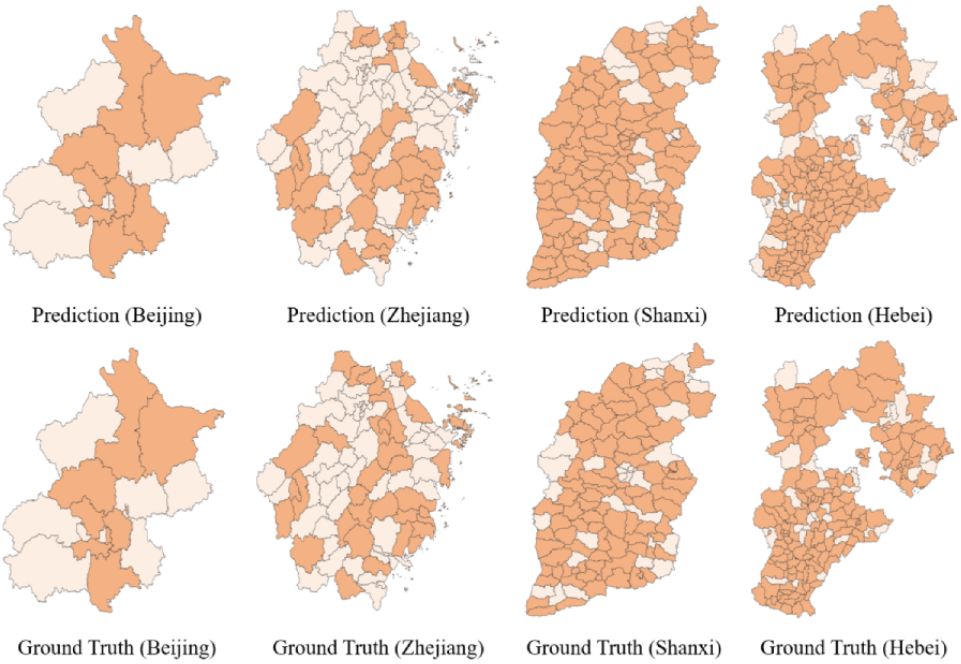
We use the same method to display the results of each province in November 2019 (Figure 6), demonstrating that our model can correctly predict disease risk in these 4 provinces to a large extent. Due to the difference in the number of counties and cities in each province, model prediction accuracies differed. Provinces with more subregions had more incorrect predictions. As in the previous case, most regions with incorrect predictions were the values of surrounding regions. In general, our model can achieve good results in predicting spatial–temporal foodborne disease risk and has a certain degree of robustness. It can achieve multistep disease risk prediction, which can provide more information for the prevention and control of foodborne disease.

Figure 5. Case study 1: The first row displays the predictions and the second row displays ground truths for Beijing in October, November, and December in 2019.



XSL•FO
RenderX

Figure 6. Case study 2: The first row displays predictions for Beijing, Zhejiang, Shanxi, and Hebei in November 2019, and the second row displays the ground truths for Beijing, Zhejiang, Shanxi, and Hebei in November 2019.



Discussion

Principal Results

Our proposed model utilizes structural LSTM to model spatial dependence and temporal dependence in data and takes into account multiple spatial correlations rather than the single spatial proximity. We also incorporated external features and spatial–temporal attention mechanisms to refine the model. The model was validated using the real-world foodborne disease data sets.

The results demonstrate that our model performs better than other models, for the 4 provinces that we selected, in determining future foodborne disease risk. Our model with multiple spatial graphs achieved the best prediction results for all provinces and prediction ranges, and our model with a single graph achieved the second-best prediction results in most cases, which shows that compared to other prediction models, including statistical models and deep learning models, our method can model temporal and spatial dependence better.

We have a better understanding of the influence of each module of the model on prediction from experiments with spatial dependence, including external features, and including an attention mechanism. Each spatial dependence has a different effect on model prediction, and models that only use a single spatial dependence are not as effective as models that use multiple spatial dependencies. Models with external features will have more accurate risk prediction results; we also use the same method to conduct experiments to verify the influence of spatial–temporal attention on the model, and the spatial–temporal attention mechanism had a positive effect on the model. Mapped results demonstrate that our model is accurate, with long-term prediction advantages, and that our model is robust, meaning that it can be used for nationwide foodborne disease risk prediction. We found that most incorrect predictions are clustered (and predicted to be the value of a nearby area).

Limitations

This study has certain limitations. First, due to the difficulty in obtaining multisource data and because model training takes a long time, we only selected 4 provinces (those with the best-quality data) to conduct experiments. Therefore, the experimental results may not be representative of all provinces in the country. In the future, we will conduct more experiments in more provinces to validate the model. Second, our model takes 4 spatial correlations into account, but real spatial correlations may be more complicated. Therefore, in the future, we will further analyze foodborne disease data and correlations with other data, to refine our model. Third, our model uses month as the temporal unit. Month-based risk prediction can better estimate long-term disease risk; however, the use of finer time-granularity disease risk prediction can provide more precise guidance for disease risk prevention and control disease risk prediction that uses smaller units can provide more comprehensive support for the prevention of foodborne diseases.

Conclusions

We focused on foodborne disease risk prediction and proposed a multigraph structural LSTM spatial–temporal prediction model based on an encoder–decoder structure. Disease risk in each region in the future was considered to be influenced by the historical disease records as well as by disease risk in surrounding areas. Moreover, in addition to proximity in space, other spatial correlations that affect disease risk prediction were taken into account by using an adaptive multigraph fusion method to adjust the effect of spatial dependencies in different circumstances. We also added a spatial–temporal attention mechanism and external features to refine the model.

Applied to a real-world foodborne disease data set from Beijing, Zhejiang, Shanxi, and Hebei, the model's performance was

XSL•FO
RenderX

better than those of other models, and highest F1 score was 20% higher than the best results of the other models. Our model can better predict the risk of foodborne diseases in the future and can provide supporting data for risk assessment, prevention, and control of foodborne diseases.

In the future, we will evaluate our model in more provinces, consider more spatial correlations, with finer time granularity, and construct an interactive foodborne disease risk prediction system that can provide more intuitive and convenient supporting data for the prevention of foodborne diseases.

Acknowledgments

This research is supported by the National Key Research and Development Plan (grant 2017YFC1601504) and the Natural Science Foundation of China (grant 61836013).

Conflicts of Interest

None declared.

References

1. Cliver D, Riemann H. Foodborne Diseases. California, US: Academic Press; 2002.
2. Oliver SP. Foodborne pathogens and disease special issue on the national and international pulsenet network. Foodborne Pathog Dis 2019 Jul;16(7):439-440. [doi: 10.1089/fpd.2019.29012.int] [Medline: 31259613]
3. Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg Infect Dis 2001 Jun;7(3):382-389. [doi: 10.3201/eid0703.017303] [Medline: 11384513]
4. McCabe-Sellers BJ, Beattie SE. Food safety: emerging trends in foodborne illness surveillance and prevention. J Am Diet Assoc 2004 Nov;104(11):1708-1717. [doi: 10.1016/j.jada.2004.08.028] [Medline: 15499359]
5. Li W, Pires SM, Liu Z, Ma X, Liang J, Jiang Y, et al. Surveillance of foodborne disease outbreaks in China, 2003–2017. Food Control 2020 Dec;118:107359. [doi: 10.1016/j.foodcont.2020.107359]
6. Gallo M, Ferrara L, Calogero A, Montesano D, Naviglio D. Relationships between food and diseases: what to know to ensure food safety. Food Res Int 2020 Nov;137:109414. [doi: 10.1016/j.foodres.2020.109414] [Medline: 33233102]
7. Thakur M, Olafsson S, Lee J, Hurburgh CR. Data mining for recognizing patterns in foodborne disease outbreaks. J Food Eng 2010 Mar;97(2):213-227. [doi: 10.1016/j.jfoodeng.2009.10.012]
8. Sadilek A, Kautz H, Silenzio V. Predicting disease transmission from geo-tagged micro-blog data. 2012 Presented at: AAAI Conference on Artificial Intelligence; July 22-26; Toronto, Canada.
9. Sadilek A, Kautz H, DiPrete L, Labus B, Portman E, Teitel J, et al. Deploying Nemesis: preventing foodborne illness by data mining social media. AIMag 2017 Mar 31;38(1):37-48. [doi: 10.1609/aimag.v38i1.2711]
10. Effland T, Lawson A, Balter S, Devinney K, Reddy V, Waechter H, et al. Discovering foodborne illness in online restaurant reviews. J Am Med Inform Assoc 2018 Dec 01;25(12):1586-1592 [FREE Full text] [doi: 10.1093/jamia/ocx093] [Medline: 29329402]
11. Vilne B, Meistere I, Grantiņa-Ieviņa L, Ķibilds J. Machine learning approaches for epidemiological investigations of food-borne disease outbreaks. Front Microbiol 2019 Aug 6;10:1722 [FREE Full text] [doi: 10.3389/fmicb.2019.01722] [Medline: 31447800]
12. Pan W, Zhao J, Chen Q. Classification of foodborne pathogens using near infrared (NIR) laser scatter imaging system with multivariate calibration. Sci Rep 2015 Apr 10;5(1):9524 [FREE Full text] [doi: 10.1038/srep09524] [Medline: 25860918]
13. Xiao X, Ge Y, Guo Y. Automated detection for probable homologous foodborne disease outbreaks. 2015 Presented at: Pacific-Asia Conference on Knowledge Discovery and Data Mining; May 19-22; Ho Chi Minh, Vietnam. [doi: 10.1007/978-3-319-18038-0_44]
14. Neill D, Moore A. Rapid detection of significant spatial clusters. 2004 Presented at: Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 22-25; Seattle, Washington, USA p. 256-265. [doi: 10.1145/1014052.1014082]
15. Nogueira M, Greis N. Rule-based complex event processing for food safety and public health. In: Bassiliades N, Governatori G, Paschke A, editors. Rule-Based Reasoning, Programming, and Applications. Berlin: Springer; 2011.
16. Wu Y, Yang Y, Nishiura H. Deep learning for epidemiological predictions. 2018 Presented at: 41st International ACM SIGIR Conference on Research & Development in Information Retrieval; July 8-12; Ann Arbor, MI, USA p. 1085-1088. [doi: 10.1145/3209978.3210077]
17. Wang L, Chen J, Marathe M. DEFSI: deep learning based epidemic forecasting with synthetic information. 2019 Jul 17 Presented at: AAAI Conference on Artificial Intelligence; January 27-February 1; Honolulu, Hawaii, USA p. 9607-9612. [doi: 10.1609/aaai.v33i01.33019607]

XSL•FO
RenderX

18. Aramaki E, Maskawa S, Morita M. Twitter catches the flu: detecting influenza epidemics using twitter. : Association for Computational Linguistics; 2011 Presented at: In Proceedings of the conference on empirical methods in natural language processing; 2011 July 27-31; Edinburgh, Scotland.
19. Akaike H. Fitting autoregressive models for prediction. Ann Inst Stat Math 1969 Dec;21(1):243-247. [doi: 10.1007/bf02532251]
20. Box GEP, Pierce DA. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J Am Stat Assoc 1970 Dec;65(332):1509-1526. [doi: 10.1080/01621459.1970.10481180]
21. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986 Oct 9;323(6088):533-536. [doi: 10.1038/323533a0]
22. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
23. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput 1989 Dec;1(4):541-551. [doi: 10.1162/neco.1989.1.4.541]
24. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. In: Proceedings of the 30th Annual Advances in Neural Information Processing Systems. 2016 Presented at: Advances in Neural Information Processing Systems 29; December 5-10; Barcelona, Spain p. 3844-3852.
25. Zhao L, Song YJ, Zhang C, Liu Y, Wang P, Lin T, et al. T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction. IEEE Trans Intell Transport Syst 2020 Sep;21(9):3848-3858. [doi: 10.1109/tits.2019.2935152]
26. Jain A, Zamir R, Savarese S. Structural-RNN: deep learning on spatio-temporal graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; 2016 June 26-July 1; Las Vegas, Nevada, USA p. 5308-5317. [doi: 10.1109/cvpr.2016.573]
27. Chen D, Yang Y, Zhang Y, Yu W. Prediction of COVID-19 spread by sliding mSEIR observer. Sci China Inf Sci 2020 Nov 12;63(12):1-13. [doi: 10.1007/s11432-020-3034-y]
28. Li Y, Zou X. Identifying disease modules and components of viral infections based on multi-layer networks. Sci China Inf Sci 2016 Jun 7;59(070102):1-15. [doi: 10.1007/s11432-016-5580-2]
29. Zhang J, Zheng Y, Qi D, Li R, Yi X, Li T. Predicting citywide crowd flows using deep spatio-temporal residual networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2017 Presented at: AAAI Conference on Artificial Intelligence; Feb 4-9; San Francisco, California, USA p. 1655-1661. [doi: 10.1016/j.artint.2018.03.002]
30. Guo S, Lin Y, Feng N, Song C, Wan H. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In: AAAI Conference on Artificial Intelligence. 2019 Presented at: Proceedings of the AAAI Conference on Artificial Intelligence; 2019 Jan 27- Feb 1; Honolulu, Hawaii, USA p. 922-929. [doi: 10.1609/aaai.v33i01.3301922]
31. Yu B, Yin H, Zhu Z. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018 Presented at: 27th International Joint Conference on Artificial Intelligence; 2018 Feb 2-7; New Orleans, Louisiana, USA p. 3634-3640. [doi: 10.24963/ijcai.2018/505]
32. Wang B, Luo X, Zhang F. Graph-based deep modeling and real time forecasting of sparse spatio-temporal data. 2018 Presented at: ACM SIGKDD Conference on Knowledge Discovery and Data Mining; Aug 19-23; London,England.
33. Kim Y, Wang P, Mihaylova L. Structural recurrent neural network for traffic speed prediction. 2019 Presented at: IEEE International Conference on Acoustics, SpeechSignal Processing; May 12-17; Brighton, USA p. 5207-5211. [doi: 10.1109/icassp.2019.8683670]
34. Ramos J. Using TF-IDF to determine word relevance in document queries. Semantic Scholar. 2003. URL: https://www.semanticscholar.org/paper/Using-TF-IDF-to-Determine-Word-Relevance-in-Queries-Ramos/b3bf6373ff41a115197cb5b30e57830c16130c2c [accessed 2021-06-16]
35. Foodborne disease surveillance and reporting system. China National Center for Food Safety Risk Assessment. URL: https://foodnet.cfsa.net.cn/ [accessed 2019-01-01]

Abbreviations

COVID-19: coronavirus disease 2019
LSTM: long short-term memory
POI: point of interest

XSL•FO
RenderX

Edited by G Eysenbach; submitted 07.04.21; peer-reviewed by L Min, G Liu; comments to author 29.04.21; revised version received 11.05.21; accepted 19.05.21; published 02.08.21.

Please cite as:
Du Y, Wang H, Cui W, Zhu H, Guo Y, Dharejo FA, Zhou Y
Foodborne Disease Risk Prediction Using Multigraph Structural Long Short-term Memory Networks: Algorithm Design and Validation Study
JMIR Med Inform 2021;9(8):e29433
URL: https://medinform.jmir.org/2021/8/e29433
doi:10.2196/29433
PMID:34338648

©Yi Du, Hanxue Wang, Wenjuan Cui, Hengshu Zhu, Yunchang Guo, Fayaz Ali Dharejo, Yuanchun Zhou. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 02.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Deep Neural Network for Estimating Low-Density Lipoprotein Cholesterol From Electronic Health Records: Real-Time Routine Clinical Application

Sangwon Hwang[1], PhD; Chanwoo Gwon[2], MS; Dong Min Seo[3], BS; Jooyoung Cho[4], MD, PhD; Jang-Young Kim[5], MD, PhD; Young Uh[4], MD, PhD

[1]Artificial Intelligence Bigdata Medical Center, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea
[2]Wonju Industry-Academic Cooperation Foundation, Yonsei University Mirae Campus, Wonju, Republic of Korea
[3]Department of Medical Information, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea
[4]Department of Laboratory Medicine, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea
[5]Department of Internal Medicine, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea

Corresponding Author:
Young Uh, MD, PhD
Department of Laboratory Medicine
Yonsei University Wonju College of Medicine
20, Ilsan- ro, Wonju, Gangwon-do
Wonju, 26426
Republic of Korea
Phone: 82 33 741 1592
Fax: 82 33 731 0506
Email: u931018@yonsei.ac.kr

Abstract

Background: Previously, we constructed a deep neural network (DNN) model to estimate low-density lipoprotein cholesterol (LDL-C).

Objective: To routinely provide estimated LDL-C levels, we applied the aforementioned DNN model to an electronic health record (EHR) system in real time (deep LDL-EHR).

Methods: The Korea National Health and Nutrition Examination Survey and the Wonju Severance Christian Hospital (WSCH) datasets were used as training and testing datasets, respectively. We measured our proposed model's performance by using 5 indices, including bias, root mean-square error, P10-P30, concordance, and correlation coefficient. For transfer learning (TL), we pretrained the DNN model using a training dataset and fine-tuned it using 30% of the testing dataset.

Results: Based on 5 accuracy criteria, deep LDL-EHR generated inaccurate results compared with other methods for LDL-C estimation. By comparing the training and testing datasets, we found an overfitting problem. We then revised the DNN model using the TL algorithms and randomly selected subdata from the WSCH dataset. Therefore, the revised model (DNN+TL) exhibited the best performance among all methods.

Conclusions: Our DNN+TL is expected to be suitable for routine real-time clinical application for LDL-C estimation in a clinical laboratory.

(JMIR Med Inform 2021;9(8):e29331) doi:10.2196/29331

KEYWORDS

low-density lipoprotein cholesterol; deep neural network; transfer learning; real-time clinical application

Introduction

Low-density lipoprotein cholesterol (LDL-C) is a major marker of cardiovascular disease (CVD) because of its role in the pathophysiology of atherosclerosis [1]. The contemporary reference measurement procedure for LDL-C is ultracentrifugation [2]. However, owing to the difficulty in applying this in a clinical setting, LDL-C levels have mostly been estimated by other means [3-6].

XSL•FO
RenderX

Friedewald et al [3] observed that most plasma samples are comprised of chylomicrons and that most triglycerides (TGs) in plasma are present in very low-density lipoprotein cholesterol (VLDL-C) at a ratio of 5:1, while the chylomicrons are undetectable. This observation led to the 1972 Friedewald (FW) equation, which is used to estimate LDL-C [3]. Martin et al [4] showed in 2014 that VLDL-C levels estimated by simply dividing the TG level by 5 may inaccurately predict LDL-C levels, specifically in hypertriglyceridemia. They divided subjects according to the levels of TG and non–high-density lipoprotein cholesterol (non-HDL-C), yielding 180 groups (clusters) [4]. For those, 180 equations were established and integrated into the novel estimation method. More recently, Sampson et al [5] used the interaction between TG and non-HDL-C and a correction factor (TG^2) to estimate LDL-C, resulting in the National Institutes of Health (NIH) method.

Deep learning techniques, specifically deep neural networks (DNNs), provide multilayer stacks of simple networks (eg, perceptrons or modules) with nonlinear functions applied between each layer [7]. The numerous perceptrons and the nonlinearity between them allow researchers to represent complex real data in a way that solves a variety of challenging tasks such as classification and regression. We previously established a deep learning model to estimate LDL-C, including 180 perceptrons [6], motivated by the model of Martin et al [4]. This yielded accurate results for LDL-C estimation.

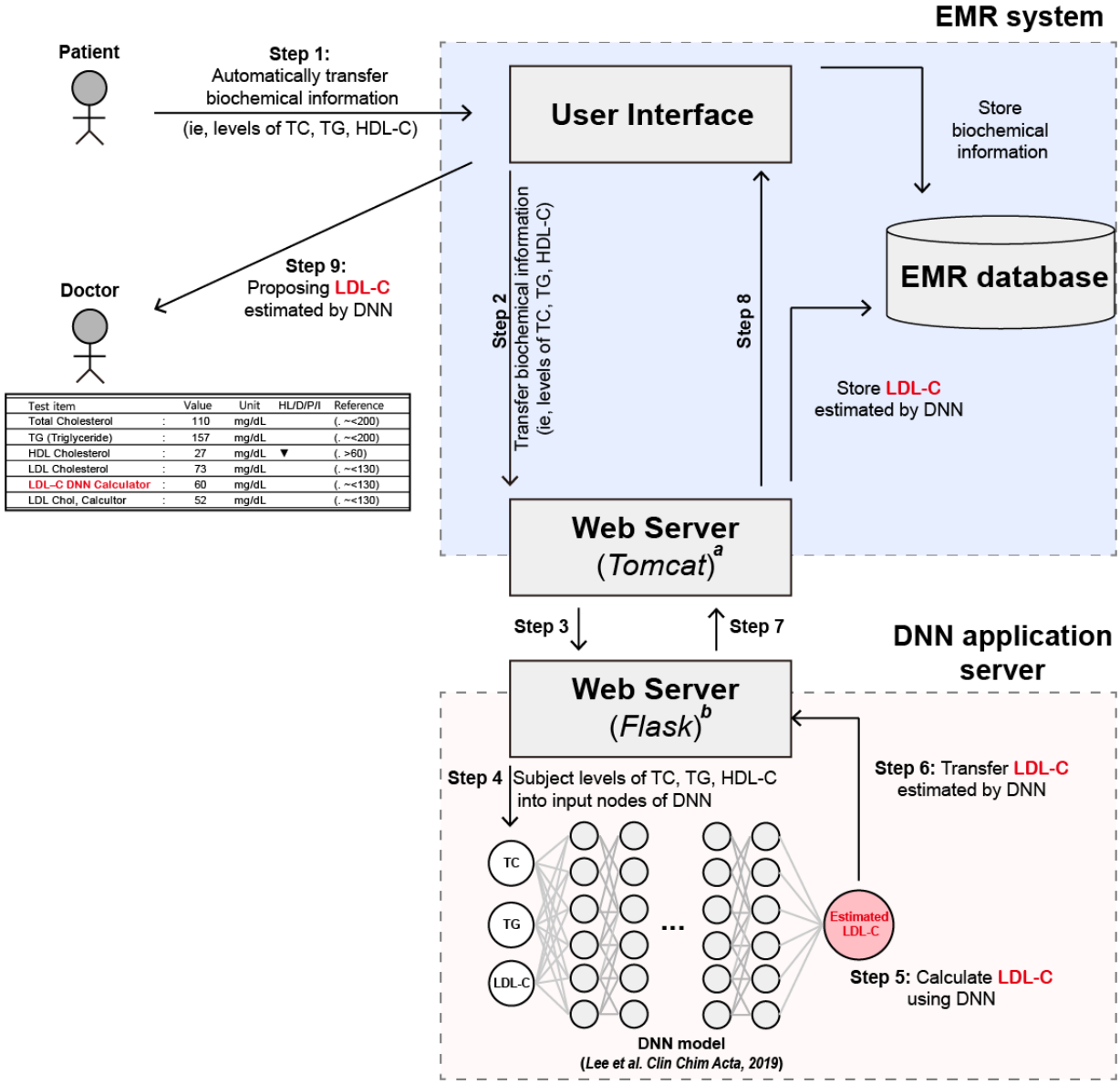
Additionally, DNNs are easy to apply in clinical settings and hospital databases. Several studies have adopted linear regression to estimate LDL-C using fewer than 5 trained weights (parameters) [8,9]. With such a low number, it is possible to adapt the linear model–based LDL estimator to a hospital database without having to rebuild the system. With the DNN proposed by Lee et al [6], approximately 4600 trained weights were established as a matrix. Although it had many weights, it was applicable to clinical settings and hospital databases using matrix calculation. Moreover, if the independent DNN application server is present, it is easy to apply and upgrade without rebuilding the system.

Transfer learning (TL) is a method of transferring knowledge from a previously trained task to a new but related one [10]. In a clinical setting, it is enormously difficult to collect real patient data and preprocess them to analyzable forms (structured data). Moreover, for these analyses, a great deal of effort is needed to resolve ethical issues and receive board approval for data collection. The difficulty of preparing an analyzable dataset presents an enormous obstacle for training because it typically requires an enormous dataset to train numerous perceptrons [7]. However, TL adopts a pretrained model learned from publicly available or large-scale datasets. Hence, it is considered to be a powerful method when it comes to small-scale dataset training requirements.

Over the past decade, enormous volumes of medical data have been stored in electronic health records (EHRs) (ie, electronic medical records [EMRs]) from which many studies have compiled patient information for secondary use for health care tasks and medical decisions (eg, disease prediction). Shickel et al [11] reviewed the current research that applied deep learning to EHRs. Although there have been many studies that constructed models using data obtained from EHR data, very few were found to have performed real-time clinical applications of the established model [12]. This study aimed to remedy this by applying previously constructed models to an EHR system. Hence, we performed the following 3 tasks for this study. First, we applied the DNN model from Lee et al [6] to the Wonju Severance Christian Hospital (WSCH) EHR system to generate real-time results for estimated LDL-C (deep LDL-EHR; Figure 1). Second, we measured performance based on several accuracy indices for the estimated LDL-C levels provided by the real-time application of our DNN model (deep LDL-EHR) and compared them to those of other LDL estimation methods. Third, we revised the DNN model by using TL, a multitask learning algorithm (Figure 2).

XSL•FO
RenderX

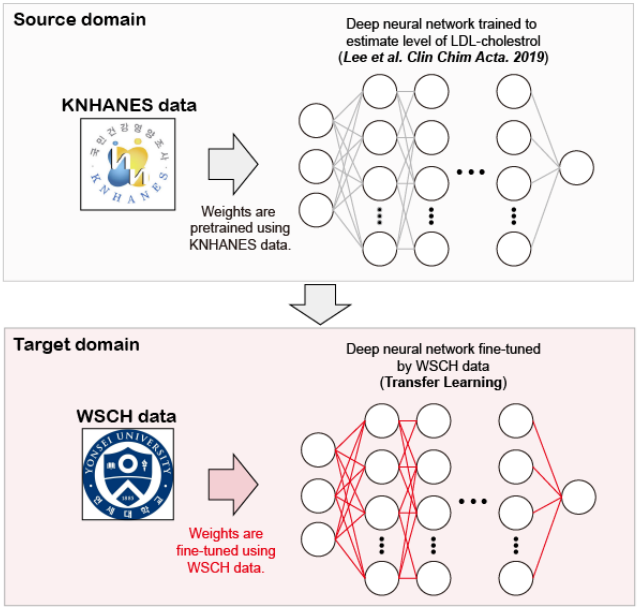
Figure 1. Overall workflow of deep LDL-EHR: Steps 3, 7, and 8 provide input- or output-value transfers between 2 platforms; the (Tomcat)[a] web server was established using Apache Tomcat [13] on a JAVA server page and servlet application; the (Flask)[b] web server was established using the Flask framework [14], a lightweight web application framework based on TensorFlow and Keras in Python. DNN: deep neural network; EMR: electronic medical record; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; TC: total cholesterol; TG: triglyceride.



Test item		Value	Unit	HL/D/P/I	Reference
Total Cholesterol	:	110	mg/dL		(. ~<200)
TG (Triglyceride)	:	157	mg/dL		(. ~<200)
HDL Cholesterol	:	27	mg/dL	▼	(. >60)
LDL Cholesterol	:	73	mg/dL		(. ~<130)
LDL-C DNN Calculator	:	60	mg/dL		(. ~<130)
LDL Chol, Calcultor	:	52	mg/dL		(. ~<130)

XSL•FO
RenderX

Figure 2. Transfer learning: For the task in the source domain, the deep neural network (DNN) model has the same structure and data as those previously trained by Lee et al [6], while ours is trained and saved on the DNN application server. For the task in the target domain, the DNN model saved in the DNN application server is loaded and retrained (fine-tuned) using Wonju Severance Christian Hospital (WSCH) data (30% randomly selected subjects) on a local computer. KNHANES: Korea National Health and Nutrition Examination Survey; LDL: low-density lipoprotein.



Methods

Application of Our DNN Model in a Clinical Laboratory

Experts in various fields (ie, clinical pathologists, database administrators, cardiologists, and computer scientists) have collaborated to construct a deep LDL-EHR model that we are using to provide LDL-C estimations for hospital patients. The application of our DNN model (ie, the deep LDL-EHR) in a clinical laboratory consists of 2 main subsystems: the EMR and a DNN application server. The EMR system is responsible for receiving and storing patient medical data (eg, levels of total cholesterol [TC], HDL-C, and TG) and transferring them to the DNN application server. The following core components are part of the EMR system: a user interface that receives data from users and stores them in the EMR database; a web server that hosts the application that permits users to see laboratory results and estimates via a web browser; a database that stores all data, including laboratory markers (input data) and results estimated by deep learning; and a physical server that runs these software components. The web service was developed using JAVA Server Pages (JSP) and a servlet application [15], and the user interface is based on the hypertext markup language, cascade style sheets, and JavaScript [16]. The web server was established in Apache Tomcat [13] based on JSP and servlets. We used a Sybase relational database management system for its construction [17].

The DNN application server hosts the DNN application, which is built upon a Python environment running separately from the EMR system. It is responsible for performing the estimation of LDL-C values based on the received data (TC, HDL-C, TG) from the EMR system and for transferring the estimated values of LDL-C back to the EMR system (Figure 1). This application server is comprised of several core components, including a flask-based web server [14] built using the flask framework (ie, a lightweight web application framework on Python), which receives data from the EMR system and transfers estimated LDL-C values back to the EMR system. It is also comprised of an application that calculates LDL-C values using the data received from the EMR system, a TensorFlow [18] framework that provides various Python application programming interfaces (APIs) that execute high-performance DNN analysis, a Keras [19] neural network library installed atop a Microsoft cognitive toolkit, TensorFlow, and Theano, which provides high-level easy-to-use APIs for creating neural networks. Although the 2 libraries are technically separate, TensorFlow and Keras are typically used in a unified manner.

Note that the optimization of weights or parameters is performed on a local computer and is saved in the form of a matrix; the DNN application server processes only the matrix operations using previously trained weights in the local computer.

Data Collection

From July 2020 to December 2020, we obtained 11,125 estimated LDL results from a real-time system. Because these results were obtained from inpatients and outpatients from all departments (eg, cardiology, gastroenterology, endocrinology, oncology, and health check-up centers) in real time, we could not trace whether examinations were performed before or after fasting. The TC, TG, HDL-C, and LDL-C data were analyzed using the modular Diagnostic de Performance Énergétique system (Roche Diagnostics, Basel, Switzerland).

We collected 2009-2015 Korea National Health and Nutrition Examination Survey (KNHANES) datasets to replicate the DNN model of Lee et al [6] Note that results in Multimedia Appendix 1 refer to the DNN model of Lee et al [6], and those in Figure 4 refer to the replicated DNN model. Subjects missing TC, HDL-C, TG, and LDL-C data were excluded. Therefore, data for 15,074 subjects were analyzed for this study, nearly the same as the number used in the previous study [6]. All participants were tested for lipid profiles after at least 12 hours of fasting.

XSL•FO
RenderX

Lipid profiles (ie, TC, HDL-C, TG, and LDL-C) were measured using the Hitachi 7600 analyzer (Hitachi, Tokyo, Japan).

Other LDL-C Estimation Methods

There have been numerous studies on the estimation of LDL-C, and they largely used linear regression methods [20,21]. Among them, we empirically selected some representative methods, including FW, Novel, and NIH methods [3-5]. The FW method estimates LDL-C by subtracting levels of HDL-C and TG/5 from TC. The Novel method integrates clustering and linear regression, initially arranging a sample into one of 180 subgroups previously determined by TG and non-HDL-C levels. Afterward, a case of 180 linear regression equations is applied to the sample. The NIH method uses TC, HDL-C, TG, and their combinations, including the square of TG (TG^2) and a multiplication value between TG and non-HDL-C. The source code for these equations is available at our GitHub homepage [22].

DNN and TL

The DNN model included 6 hidden layers with 30 hidden nodes in each. We used a rectified linear unit as an activation function to implement nonlinearity between the hidden layers. The details of this model are described in the study by Lee et al [6].

We used TL [10] to upgrade this DNN model [6]. TL includes a source domain that is typically a large-scale dataset alongside a small-scale target domain that contains more specific data compared with those of the source domain [10]. As described in Figure 2, from the source task (ie, KNHANES dataset), we extracted the desired information (ie, trained weights). From the target task (ie, subset of the WSCH dataset), we retrained (fine-tuned) the DNN. The source code for the DNN+TL is available at our GitHub homepage [22].

Performance Measurement

To assess and compare the accuracy of each LDL-C estimation method, we measured the following 5 indices: bias (estimated LDL-C [eLDL-C] – measured LDL-C [mLDL-C]), root mean square error (RMSE), P10 to P30, concordance, and correlation coefficient.

Jeong et al [23] implemented the one-sample *t* test to compare the average bias between true and estimated values from a regression task. Motivated by this, we used the one-sample *t* test to measure the degree of average bias of each estimation method differing from zero.

Numerous studies have implemented RMSE to measure the degree of accuracy for LDL-C estimation methods [4-6,23]. Hence, we decided to use the RMSE for the estimation accuracies of each method as follows.

P30 has been implemented to measure the clinical accuracy of estimation methods for glomerular filtration rate [23]. This study used P10 and P30, and we expanded these indices as *Pn* (*n* = 10, 15, 20, 25, and 30), measured as the ratio of samples from which LDL-C was estimated using each method within mLDL-C ± n% divided by all samples.

In studies that provided the estimation method for LDL-C [4,5], concordance has been used to examine the classification accuracy between mLDL-C and eLDL-C. In detail, both mLDL-C and eLDL-C values are categorized as 6 subgroups based on the National Cholesterol Education Program (NCEP) Adult Treatment III guideline cutoffs that other studies used [24,25]. Concordance was measured as follows:

where *A* are samples with mLDL-C within a specific range and *B* are samples with eLDL-C in the same interval as mLDL-C.

Several methods of correlation have been used to measure the degree of consistency between true and estimated values (ie, mLDL-C and eLDL-C) [5,23]. Specifically, we used Pearson correlation coefficient, a normalized measurement of the covariance of 2 lists of values (ie, mLDL-C and eLDL-C) divided by the product of their standard deviation.

Jacob and Speed [26] suggested that the selected features and their predictive performances should be examined based on a random sampling perspective for generalization. In other words, the samples selected for the training model (ie, DNN+TL) greatly affect its performance. Therefore, we performed the following tasks considering the random sampling perspective. In step 1, we made a pair of random sample datasets, including training and testing, which were randomly divided at a ratio of 0.3 and 0.7, respectively. In step 2, we established a DNN+TL model using the randomly selected training set and measured the *t* value and RMSE of the DNN+TL model for the testing set. We also measured the *t* value and RMSE of other models (ie, FW, Novel, NIH, and DNN) for the testing set. In step 3, we iterated Steps 1 to 2 at 1000 times, and 2 matrices consisting of 5 columns (5 LDL-C estimation methods) and 1000 rows (# of iterations) were generated, including the *t* value and RMSE. We compared 2 indices (ie, *t* value and RMSE) among the 5 methods based on one-way analysis of variance and performed multiple comparisons using the Bonferroni post hoc test.

Variance Importance

We implemented permutation importance [27] and Shapley addictive explanations (SHAP) [28] to identify the contribution of each feature (ie, TC, HDL-C, and TL) to the final output of the DNN model. Permutation importance is a heuristic method used to measure normalized feature importance by measuring the decrease in a model's performance when a feature is permuted [27]. SHAP is an addictive feature attribution method used to determine feature importance by measuring a weighted average value of all possible differences between 2 sets of outputs that are resulted from models with and without the feature [28]. The permutation importance was measured using the *permutation_importance* function in the sklearn package [29], and the SHAP was calculated using the *DeepExplainer* function in the SHAP package [28].

XSL•FO
RenderX

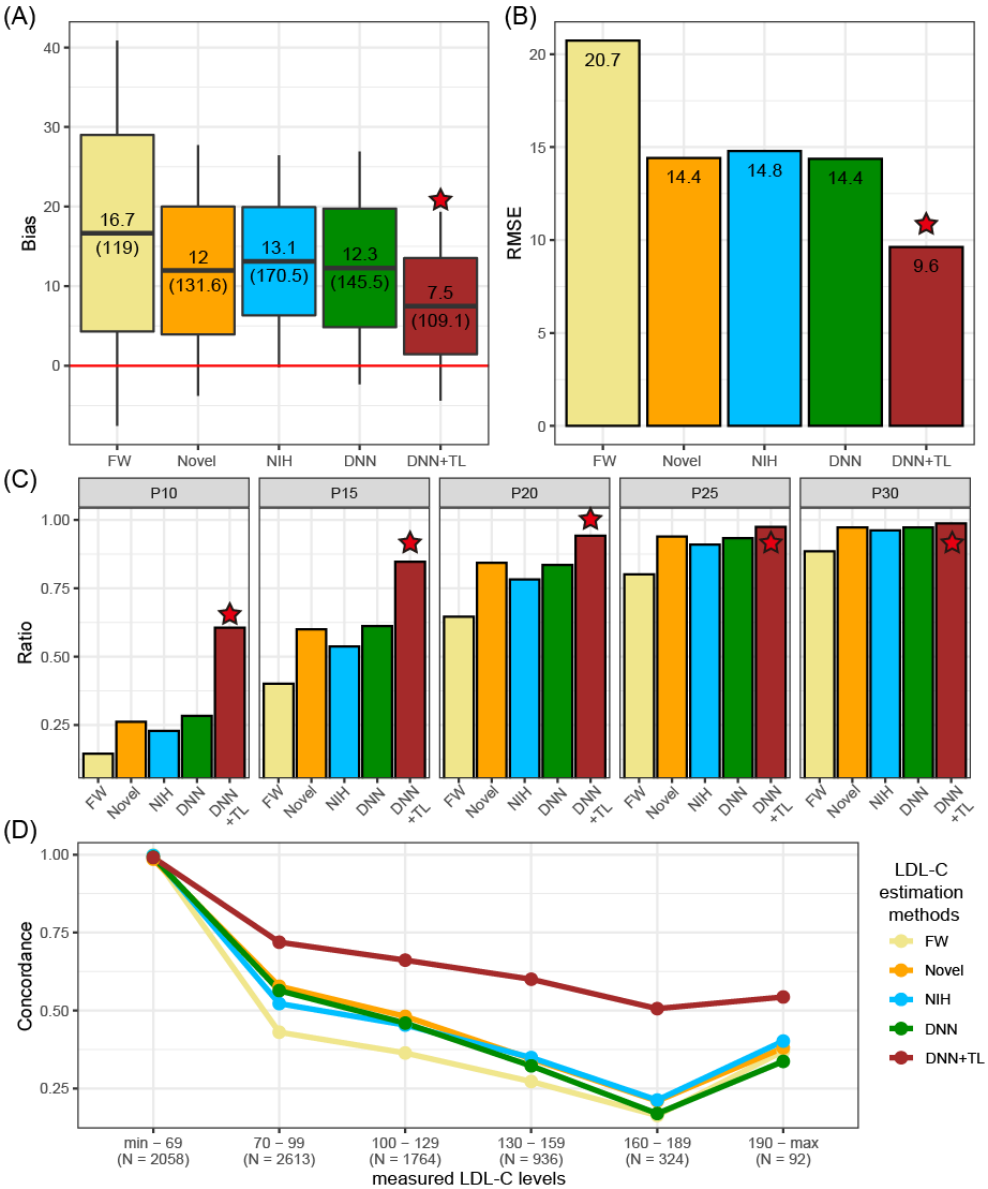
Statistics

Statistical analyses were performed using the *R* programming language (v.3.6.4). For a comparison of continuous variables based on 2 groups, we used the *t* test and the Mann Whitney *U* test. For categorical variables, we used the Chi-squared test, and a P value of <.05 was considered to be statistically significant.

Results

From the real-time application (Figure 1), we obtained 11,125 LDL values estimated using the DNN model. The distribution of bias (box plot) and RMSE (bar plot) of each LDL estimation method are illustrated in Multimedia Appendix 1. The estimated LDL-C values using the Novel method differed least from zero, and the values using the FW equation method were biased the most from zero. The eLDL-C levels using the DNN application system had, from among the 4 methods, the second most biased distribution from zero among the difference values between eLDL-C and mLDL-C (Multimedia Appendix 1). When comparing the RMSE of each method, the FW method resulted in the highest RMSE, followed by the DNN application system. In all the P10 to P30, the FW method showed the lowest ratio, and the DNN application system showed the second lowest ratio (Figure 3C; Multimedia Appendix 1). We compared concordances between groups stratified by mLDL-C and eLDL-C levels obtained from the 4 methods (Figure 3D). Therefore, the novel method showed the highest concordance from 70 to 129 of the mLDL-C levels, and the NIH method showed the highest concordance from 130 to the maximum mLDL-C levels (Multimedia Appendix 1). Collectively, the DNN application generated inaccurate results compared with the others.

Figure 3. Performance of 5 LDL estimation methods: (A) upper and lower numbers indicate the average and one-sample *t* value, respectively, while the black bars, upper or lower margins, and maximum or minimum lines for each boxplot indicate 1 SD and 1.96 SDs, respectively; (B) numbers in bar plots indicate real values of RMSE; (C) P10 to P30; (D) concordance of each LDL-C estimation method. Stars in each plot indicate the model with the best performance. Note that the deep neural network (DNN) method was the replicated model for the DNN model. FW: Friedewald equation; NIH: National Institutes of Health; RMSE: root mean square error; TL: transfer learning.



XSL•FO
RenderX

We compared the lipid profiles of the KNHANES dataset with those of the WSCH dataset (Table 1). All 4 variables differed significantly between the 2 datasets. We concluded that differential characteristics between the training set (KNHANES) and the testing set (WSCH) triggered inaccurate results from the DNN application system. In other words, an overfitting problem existed in the deep LDL-EHR model. To overcome this limitation, we adopted the TL method [10]. Using the 2009-2015 KNHANES datasets, we trained the DNN model using the same structure and hyperparameters as those of the model proposed by Lee et al [6], yielding a pretrained DNN model. Next, we randomly selected 30% of the WSCH dataset, which was used to fine-tune the pretrained DNN model (Figure 2).

Table 1. General characteristics of and comparisons between the Korea National Health and Nutrition Examination Survey (KNHANES) and Wonju Severance Christian Hospital (WSCH) datasets.

Variable	KNHANES (n=15,074)	WSCH (n=11,125)	P value
Age (years), mean (SD)	45.5 (18.2)	59.4 (15.5)	<.001[a]
Age (years), median (IQR)	46 (32-60)	60 (51-70)	<.001[b]
Male, n (%)	7507 (49.8)	6435 (57.8)	<.001[c]
Total cholesterol (mg/dL), mean (SD)	188.8 (37.7)	156.4 (41.6)	<.001[a]
Total cholesterol (mg/dL), median (IQR)	186 (162-212)	152 (128-182)	<.001[b]
HDL[d] cholesterol (mg/dL), mean (SD)	48.7 (12.1)	50.2 (14.2)	<.001[a]
HDL cholesterol (mg/dL), median (IQR)	47.3 (40.1-55.7)	48 (40-58)	<.001[b]
Triglyceride (mg/dL), mean (SD)	160.2 (135.6)	139.7 (126.2)	<.001[a]
Triglyceride (mg/dL), median (IQR)	120 (76-211)	114 (83-163)	<.001[b]
Measured LDL[e] cholesterol (mg/dL), mean (SD)	112 (32.3)	94.8 (35.9)	<.001[a]
Measured LDL cholesterol (mg/dL), median (IQR)	109 (89-132)	90 (68-117)	<.001[b]

[a]Determined using a *t* test.

[b]Determined using a Mann-Whitney *U* test.

[c]Determined using a Chi-squared test.

[d]HDL: high-density lipoprotein.

[e]LDL: low-density lipoprotein.

We compared the performances of the 5 methods, including the aforementioned 4 and DNN+TL methods (Figure 3). Based on the bias and RMSE, the DNN+TL was biased least from zero (mean 7.5; t_{7786}=109.1) and had the lowest RMSE (Figures 3A and 3B). In all of P10 to P30, the DNN+TL method had the highest ratio among the other methods. Particularly in P10, the superior performance of the DNN+TL method was notable (Figure 3C). Regarding the concordance of the LDL-C estimation methods, the DNN+TL method had the highest ratio through most of the LDL-C range except for a section of LDL-C from the minimum to 69 mg/dL (Figure 3D).

We illustrated correlation plots describing the distribution of eLDL-C values and the matched LDL-C levels estimated by the 5 methods, including FW, Novel, and DNN (Figure 4). In DNN+TL, the LDL-C level is the most accurately estimated among the other 4 methods based on the Pearson correlation coefficient (Figure 4).

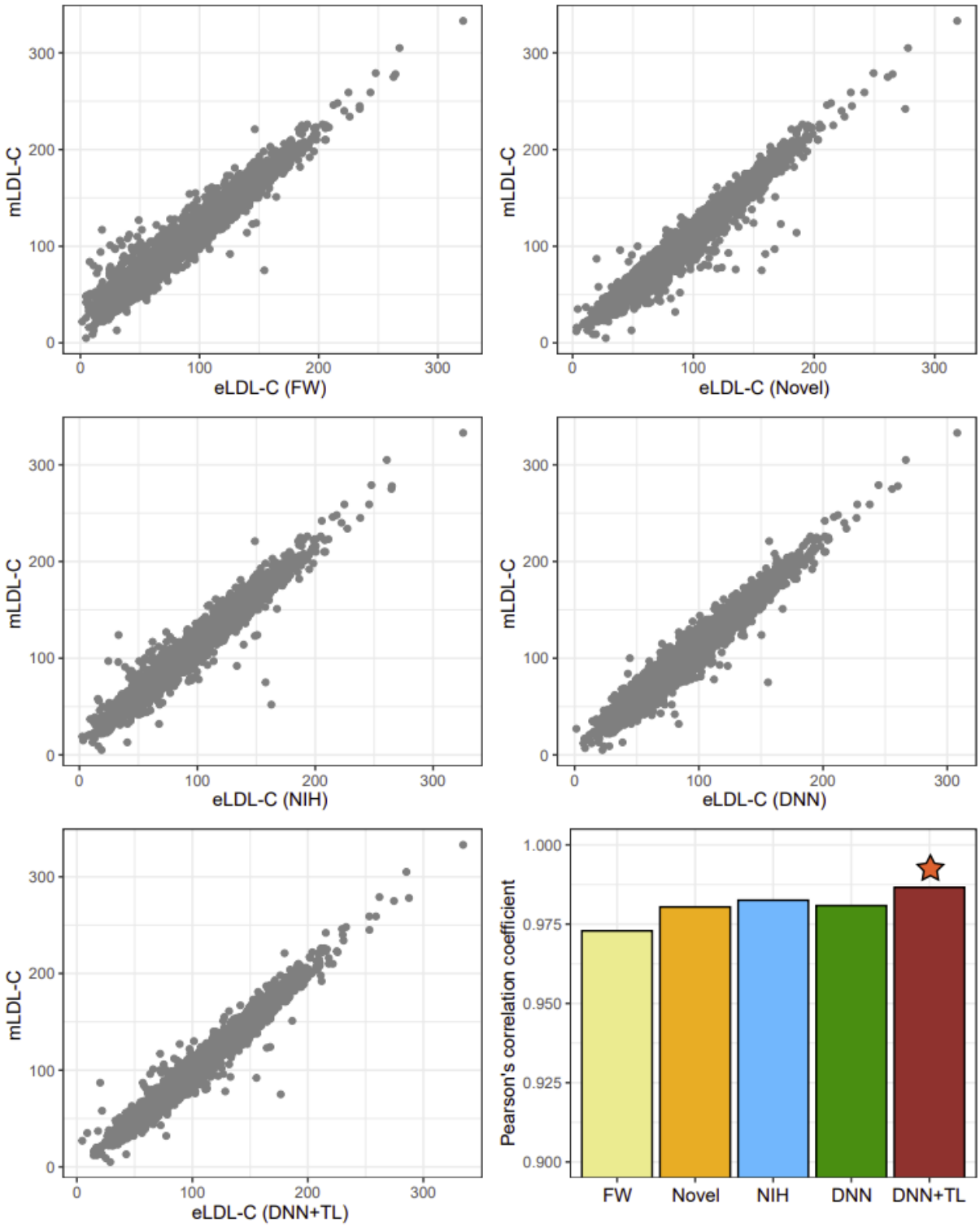
For the 5 LDL-C estimation methods, we generated distributions of *t* values and RMSE, separately, by iterating the random selection of training set at 1000 times (Figure 5). As a result, DNN+TL exhibited the best performance for both bias from zero (*t* value, Bonferroni-corrected *P*<.001 for DNN+TL vs other methods) and absolute error (RMSE, Bonferroni-corrected *P*<.001).

For input features (ie, TC, HDL-C, and TG) and their deep learning models (ie, DNN and DNN+TL), we measured the variance (global) importance by using permutation importance and SHAP (Figure 6). In both DNN and DNN+TL, TC was the best crucial feature based on 2 indices of the variance importance. Moreover, TG and HDL-C comprised the second-most important variable based on permutation importance and SHAP, respectively (Figure 6A). In DNN+TL, the second important feature was TG, based on all indices of the variance importance (Figure 6B). Moreover, we illustrated the distribution of the ratio of TG to VLDL-C in relation to TG levels (Multimedia Appendix 2). VLDL-C, as analyzed in our study, is not a measured value, but is instead the result calculated by subtracting the values of HDL-C and eLDL-C (by the 5 methods) from TC. We found that the TG to VLDL-C ratio estimated by 3 models had large variance at high TG levels (Multimedia Appendix 2), which was similar with the results in the study by Martin et al [4]. The distribution of the TG to VLDL-C ratio estimated by the DNN+TL model looked like a mixture between the ratios by mLDL-C and DNN (Multimedia Appendix 2), indicating that the DNN+TL had fine-tuned the previous DNN model [6] to represent the characteristics of the

XSL•FO
RenderX

WSCH dataset by importantly considering the TG variable (Figure 6).

Figure 4. Correlation plots and coefficients between measured low-density lipoprotein cholesterol (mLDL-C) and estimated LDL-C (eLDL-C) calculated by 5 methods. The points on the scatterplots indicate the individual samples. A star indicates the highest Pearson correlation coefficient. DNN: deep neural network; FW: Friedewald method; NIH: National Institutes of Health; TF: transfer learning.



XSL•FO
RenderX

Figure 5. Comparison of performance based on a random sample perspective. A one-sample *t* test was used. DNN: deep neural network; FW: Friedewald method; NIH: National Institutes of Health; TL: transfer learning.

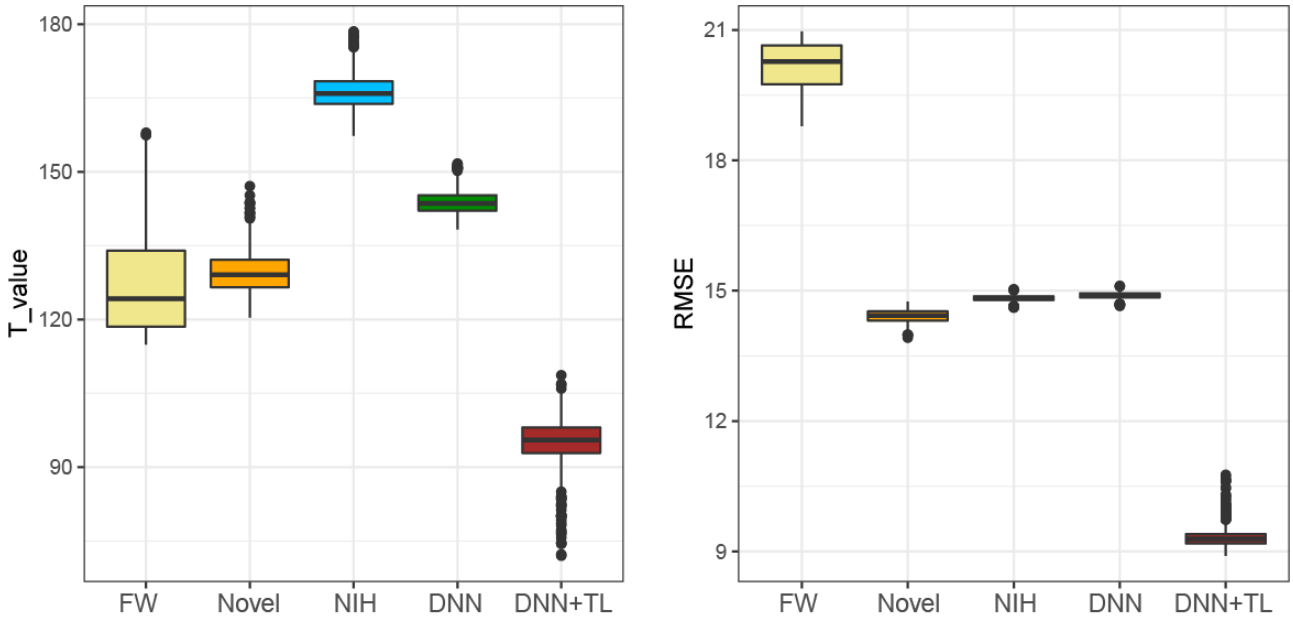
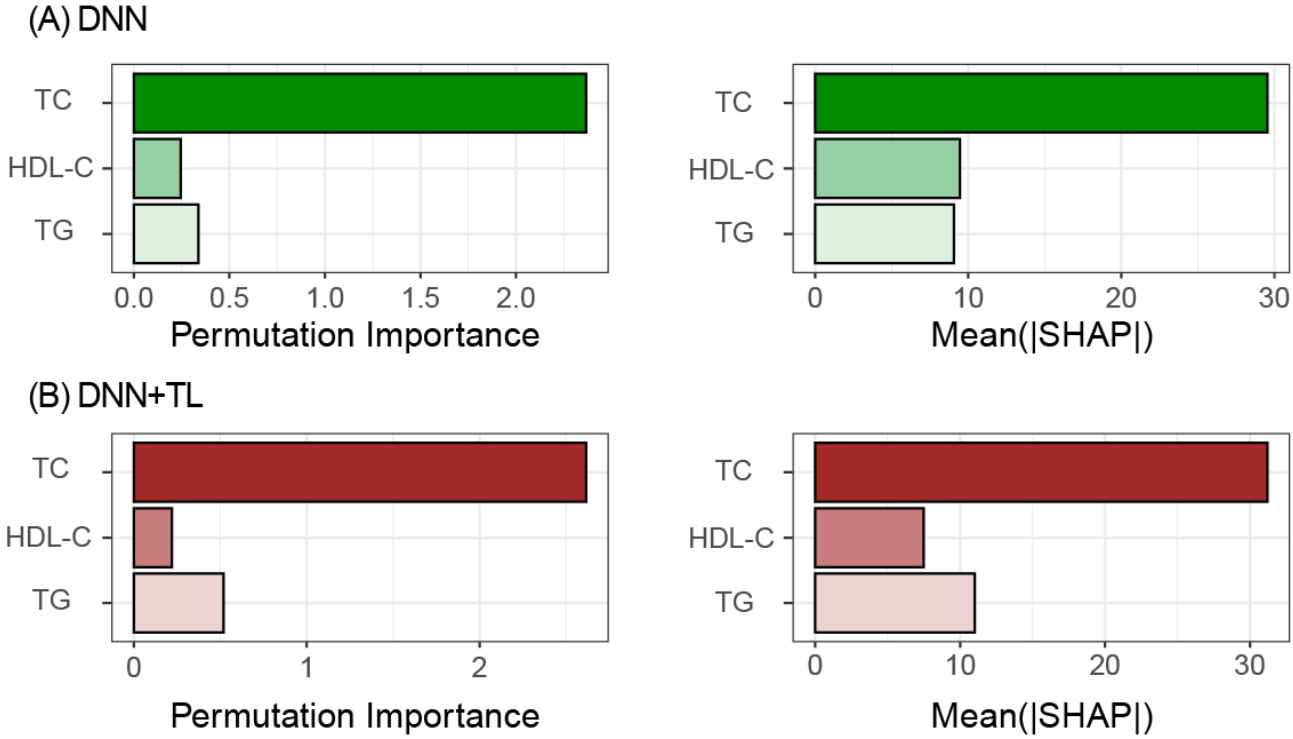


Figure 6. Variance importance based on permutation importance and Shapley addictive explanations (SHAP). DNN: deep neural network; HDL-C: high-density lipoprotein cholesterol; TC: total cholesterol; TG: triglyceride; TL: transfer learning.



Discussion

Principal Findings

We applied the DNN model for LDL-C estimation from EHR (deep LDL-EMR) data to generate real-time results. However, we found that our original deep LDL-EMR generated inaccurate results compared with other LDL estimation methods. We hypothesized that these inaccuracies may have been caused by the batch effect between the 2 different datasets. We therefore adopted a TL method to fine-tune the DNN model using local data-specific characteristics. Therefore, the DNN+TL method resulted in the most accurate results of all methods.

Approximately 15,000 subjects (KNHANES) were used to construct the DNN, and about 3300 WSCH LDL-C results were used for fine-tuning it. Martin et al [4] assigned approximately 900,000 subjects to develop the Novel method. Meeusen et al [25] enrolled 23,055 individuals from the Mayo Clinic and externally validated the Novel method. In 2020, Sampson et al [5] used approximately 9000 LDL-C test results to develop the NIH method while internally and externally validating it through approximately 9000 LDL-C results and those of another 4

XSL•FO
RenderX

databases. Our DNN model was established using approximately 18,000 LDL-C results obtained from 2 different institutions, and validation was established using approximately 77,000 LDL-C results, which was comparable to the validation in other studies.

In the study by Martin et al [4] (the Novel method), the median TG distribution was 115 (IQR 82-166). Research by Meeusen et al [25] resulted in a median TG distribution of 131 (IQR 89-196). In a study by Sampson et al [5] (NIH method), the median TG distribution was 149 (IQR 98-253). Our derivation dataset (KNHANES) had a median TG of 120 (IQR 76-211), and our validation dataset had a median TG of 114 (IQR 83-163). Although data from the Novel method had a TG distribution more similar to our validation dataset than the TG distribution from the NIH method, the performances obtained from these methods were almost identical. However, we found that our deep LDL-EHR model generated extremely accurate results for the derivation set and comparably inaccurate results for the testing dataset. In other words, an overfitting problem occurred in our deep LDL-EHR model. Therefore, we adopted a TL method to fine-tune (overall retainment with little change in trained parameters) the deep LDL-EHR (DNN+TL) model, yielding the best performance among all the methods.

Limitations and Future Work

The most important limitation of the present study is the referenced homogenous method used to measure LDL-C. Representative methods for estimating LDL-C [3-5] use the heterogeneous method of ultracentrifugation (eg, beta-quantification) [30,31]. Besides, we implemented the homogeneous precipitation-based (direct) method as the reference for establishing an LDL-C regression model. Nauck et al [30] suggested that the homogenous method satisfied the NCEP requirements and proposed accurate LDL-C results with a coefficient of variation less than 4% and a bias less than 4%. Moreover, the homogenous method seems to have better classified subjects into NCEP criteria than the FW method [30]. The homogenous method does not require the preliminary lipoprotein fractionation step (eg, ultracentrifugation). In other words, it is easy to use and often provides improved precision; therefore, it has gained rapid acceptance worldwide [31]. However, for high-risk CVD patients or groups, future studies should analyze both beta-quantifications and direct methods to provide more accurate and generalized estimates for decreasing CVD-related mortality.

In future studies, we plan to update the trained weights in the LDL-EHR model with optimized parameters using TL. Another study is needed to evaluate the performance of an updated version of the LDL-EHR (DNN+TL) model for the newly selected samples. Furthermore, as suggested by other studies [6,32], it is crucial to develop an LDL-C estimation method that considers demographic, medical, anthropometric, and laboratory phenotypes, such as age, obesity, chronic disease, and liver profiles.

Conclusion

We applied a real-time deep learning model to estimate LDL-C using EHR system data. However, we encountered several unforeseen problems. When applying the DNN model to real patients, our tool could not outperform the other LDL-C estimation methods (ie, Novel and NIH). We overcame this by upgrading our DNN using a TL algorithm (DNN+TL), resulting in superior LDL-C estimation performance compared with the other methods. Our study suggests that the revised version of our deep LDL-EHR (DNN+TL) may contribute to future accurate estimations for LDL-C in real clinical settings.

Acknowledgments

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI19C1035).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Performances of four LDL estimation methods.
[PDF File (Adobe PDF File), 369 KB - medinform_v9i8e29331_app1.pdf]

Multimedia Appendix 2
The distribution of TG:VLDL-C in relation to TG.
[PDF File (Adobe PDF File), 288 KB - medinform_v9i8e29331_app2.pdf]

References

1. Ference B, Ginsberg H, Graham I, Ray K, Packard C, Bruckert E, et al. Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel. Eur Heart J 2017 Aug 21;38(32):2459-2472 [FREE Full text] [doi: 10.1093/eurheartj/ehx144] [Medline: 28444290]

XSL•FO
RenderX

2. Miller W, Myers G, Sakurabayashi I, Bachmann L, Caudill S, Dziekonski A, et al. Seven direct methods for measuring HDL and LDL cholesterol compared with ultracentrifugation reference measurement procedures. Clin Chem 2010 Jun;56(6):977-986 [FREE Full text] [doi: 10.1373/clinchem.2009.142810] [Medline: 20378768]
3. Friedewald W, Levy R, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clin Chem 1972 Jun;18(6):499-502. [Medline: 4337382]
4. Martin SS, Blaha MJ, Elshazly MB, Toth PP, Kwiterovich PO, Blumenthal RS, et al. Comparison of a novel method vs the Friedewald equation for estimating low-density lipoprotein cholesterol levels from the standard lipid profile. JAMA 2013 Nov 20;310(19):2061-2068 [FREE Full text] [doi: 10.1001/jama.2013.280532] [Medline: 24240933]
5. Sampson M, Ling C, Sun Q, Harb R, Ashmaig M, Warnick R, et al. A New Equation for Calculation of Low-Density Lipoprotein Cholesterol in Patients With Normolipidemia and/or Hypertriglyceridemia. JAMA Cardiol 2020 May 01;5(5):540-548 [FREE Full text] [doi: 10.1001/jamacardio.2020.0013] [Medline: 32101259]
6. Lee T, Kim J, Uh Y, Lee H. Deep neural network for estimating low density lipoprotein cholesterol. Clin Chim Acta 2019 Feb;489:35-40. [doi: 10.1016/j.cca.2018.11.022] [Medline: 30448282]
7. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
8. Chen Y, Zhang X, Pan B, Jin X, Yao H, Chen B, et al. A modified formula for calculating low-density lipoprotein cholesterol values. Lipids Health Dis 2010 May 21;9:52 [FREE Full text] [doi: 10.1186/1476-511X-9-52] [Medline: 20487572]
9. de Cordova CMM, de Cordova MM. A new accurate, simple formula for LDL-cholesterol estimation based on directly measured blood lipids from a large cohort. Ann Clin Biochem 2013 Jan 29;50(Pt 1):13-19. [doi: 10.1258/acb.2012.011259] [Medline: 23108766]
10. Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Trans. Knowl. Data Eng 2010 Oct;22(10):1345-1359. [doi: 10.1109/tkde.2009.191]
11. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform 2018 Sep;22(5):1589-1604 [FREE Full text] [doi: 10.1109/JBHI.2017.2767063] [Medline: 29989977]
12. Kim YY, Oh SJ, Chun YS, Lee WK, Park HK. Gene expression assay and Watson for Oncology for optimization of treatment in ER-positive, HER2-negative breast cancer. PLoS One 2018 Jul 6;13(7):e0200100 [FREE Full text] [doi: 10.1371/journal.pone.0200100] [Medline: 29979736]
13. Vukotic A, Goodwill J. Apache Tomcat 7. Cham, Switzerland: Springer Publishing Company; 2011.
14. Ronacher A. Flask: web development, one drop at a time. URL: https://flask-doc.readthedocs.io/en/latest/ [accessed 2021-07-23]
15. Wolf D, Henley AJ. Java EE Web Application Primer: Building Bullhorn: A Messaging App with JSP, Servlets, JavaScript, Bootstrap and Oracle. New York, NY: Apress; 2017.
16. McGrath M. HTML, CSS & JavaScript in easy steps. Leamington Spa, UK: In Easy Steps Limited; 2020.
17. Hadjigeorgiou C. RDBMS vs NoSQL: Performance and scaling comparison. University of Edinburgh. 2013. URL: https://static.epcc.ed.ac.uk/dissertations/hpc-msc/2012-2013/RDBMS%20vs%20NoSQL%20-%20Performance%20and%20Scaling%20Comparison.pdf [accessed 2021-07-23]
18. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A System for Large-Scale Machine Learning. 2016 Presented at: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16); November 2-4, 2016; Savannah, GA.
19. Manaswi NK. Understanding and Working with Keras. In: Deep Learning with Applications Using Python. Berkeley, CA: Apress; 2018:31-43.
20. Wadhwa N, Krishnaswamy R. Comparison of LDL-Cholesterol Estimate using Various Formulae with Directly Measured LDL-Cholesterol in Indian Population. J Clin Diagn Res 2016 Dec;10(12):BC11-BC13 [FREE Full text] [doi: 10.7860/JCDR/2016/22272.9018] [Medline: 28208843]
21. Piani F, Cicero AF, Ventura F, Dormi A, Fogacci F, Patrono D, BLIP Study Group. Evaluation of twelve formulas for LDL-C estimation in a large, blinded, random Italian population. Int J Cardiol 2021 May 01;330:221-227. [doi: 10.1016/j.ijcard.2021.02.009] [Medline: 33581176]
22. WCH-AI-LAB: DNN-TL. GitHub. URL: https://github.com/WCH-AI-LAB/DNN-TL [accessed 2021-07-25]
23. Jeong T, Cho E, Lee W, Chun S, Hong K, Min W. Accuracy Assessment of Five Equations Used for Estimating the Glomerular Filtration Rate in Korean Adults. Ann Lab Med 2017 Sep 01;37(5):371-380 [FREE Full text] [doi: 10.3343/alm.2017.37.5.371] [Medline: 28643485]
24. National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation,Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. Circulation 2002 Dec 17;106(25):3143-3421. [Medline: 12485966]
25. Meeusen J, Lueke A, Jaffe A, Saenger AK. Validation of a proposed novel equation for estimating LDL cholesterol. Clin Chem 2014 Dec;60(12):1519-1523. [doi: 10.1373/clinchem.2014.227710] [Medline: 25336719]

XSL•FO
RenderX

26. Jacob L, Speed TP. The healthy ageing gene expression signature for Alzheimer's disease diagnosis: a random sampling perspective. Genome Biol 2018 Jul 25;19(1):97 [FREE Full text] [doi: 10.1186/s13059-018-1481-6] [Medline: 30045771]
27. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics 2010 May 15;26(10):1340-1347. [doi: 10.1093/bioinformatics/btq134] [Medline: 20385727]
28. Lundberg S, Lee SI. A unified approach to interpreting model predictions. Cornell University. URL: https://arxiv.org/abs/1705.07874 [accessed 2021-07-23]
29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825-2830.
30. Nauck M, Warnick GR, Rifai N. Methods for measurement of LDL-cholesterol: a critical assessment of direct measurement by homogeneous assays versus calculation. Clin Chem 2002 Feb;48(2):236-254. [Medline: 11805004]
31. Warnick GR, Kimberly MM, Waymack PP, Leary ET, Myers GL. Standardization of Measurements for Cholesterol, Triglycerides, and Major Lipoproteins. Laboratory Medicine 2008 Jul 17;39(8):481-490. [doi: 10.1309/6ul9rhjh1jffu4py]
32. Chakraborty M, Tudu B. Comparison of ANN models to predict LDL level in Diabetes Mellitus type 2. InInternational Conference on Systems in Medicine and Biology; 2011 Presented at: International Conference on Systems in Medicine and Biology; December 16-18, 2010; Kharagpur, India. [doi: 10.1109/icsmb.2010.5735410]

Abbreviations

API: application programming interface
CVD: cardiovascular disease
DNN: deep neural network
EHR: electronic health record
eLDL-C: estimated low-density lipoprotein cholesterol
EMR: electronic medical record
HDL-C: high-density lipoprotein cholesterol
JSP: JAVA Server Pages
KNHANES: Korea National Health and Nutrition Examination Survey
LDL-C: low-density lipoprotein cholesterol
mLDL-C: measured low-density lipoprotein cholesterol
NCEP: National Cholesterol Education Program
NIH: National Institutes of Health
RMSE: root mean square error
SHAP: Shapley addictive explanations
TC: total cholesterol
TG: triglyceride
TL: transfer learning
VLDL-C: very low-density lipoprotein cholesterol
WSCH: Wonju Severance Christian Hospital

Edited by G Eysenbach; submitted 02.04.21; peer-reviewed by T Lee, Z Ren, H Li, PP Zhao; comments to author 23.04.21; revised version received 18.06.21; accepted 05.07.21; published 03.08.21.

Please cite as:
Hwang S, Gwon C, Seo DM, Cho J, Kim JY, Uh Y
A Deep Neural Network for Estimating Low-Density Lipoprotein Cholesterol From Electronic Health Records: Real-Time Routine Clinical Application
JMIR Med Inform 2021;9(8):e29331
URL: https://medinform.jmir.org/2021/8/e29331
doi:10.2196/29331
PMID:34342586

©Sangwon Hwang, Chanwoo Gwon, Dong Min Seo, Jooyoung Cho, Jang-Young Kim, Young Uh. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 03.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Ranking Rule-Based Automatic Explanations for Machine Learning Predictions on Asthma Hospital Encounters in Patients With Asthma: Retrospective Cohort Study

Xiaoyi Zhang[1], MSc; 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 encounters impose a heavy burden on the health care system. To improve preventive care and outcomes for patients with asthma, we recently developed a black-box machine learning model to predict whether a patient with asthma will have one or more asthma hospital encounters in the succeeding 12 months. Our model is more accurate than previous models. However, black-box machine learning models do not explain their predictions, which forms a barrier to widespread clinical adoption. To solve this issue, we previously developed a method to automatically provide rule-based explanations for the model's predictions and to suggest tailored interventions without sacrificing model performance. For an average patient correctly predicted by our model to have future asthma hospital encounters, our explanation method generated over 5000 rule-based explanations, if any. However, the user of the automated explanation function, often a busy clinician, will want to quickly obtain the most useful information for a patient by viewing only the top few explanations. Therefore, a methodology is required to appropriately rank the explanations generated for a patient. However, this is currently an open problem.

Objective: The aim of this study is to develop a method to appropriately rank the rule-based explanations that our automated explanation method generates for a patient.

Methods: We developed a ranking method that struck a balance among multiple factors. Through a secondary analysis of 82,888 data instances of adults with asthma from the University of Washington Medicine between 2011 and 2018, we demonstrated our ranking method on the test case of predicting asthma hospital encounters in patients with asthma.

Results: For each patient predicted to have asthma hospital encounters in the succeeding 12 months, the top few explanations returned by our ranking method typically have high quality and low redundancy. Many top-ranked explanations provide useful insights on the various aspects of the patient's situation, which cannot be easily obtained by viewing the patient's data in the current electronic health record system.

Conclusions: The explanation ranking module is an essential component of the automated explanation function, and it addresses the interpretability issue that deters the widespread adoption of machine learning predictive models in clinical practice. In the next few years, we plan to test our explanation ranking method on predictive modeling problems addressing other diseases as well as on data from other health care systems.

International Registered Report Identifier (IRRID): RR2-10.2196/5039

(JMIR Med Inform 2021;9(8):e28287) doi:10.2196/28287

KEYWORDS

asthma; clinical decision support; machine learning; patient care management; forecasting

XSL•FO
RenderX

Introduction

Background

Approximately 7.7% of Americans and over 339 million people worldwide have asthma [1,2]. Asthma incurs a total medical cost of US $50 billion [3], 1,564,440 emergency department (ED) visits, and 182,620 inpatient stays annually in the United States [1]. A primary goal of asthma management is to decrease the number of asthma hospital encounters, namely, ED visits and inpatient stays. The state-of-the-art approach for achieving this goal is to deploy a predictive model to identify patients at high risk of having poor outcomes in the future. Once identified, the patient is placed into a care management program. The program will assign a care manager to regularly contact the patient to assess asthma control status, adjust asthma medications when needed, and help schedule appointments for health and other relevant services. Many health plans, including those in 9 of 12 metropolitan communities [4], and many health care systems, such as the University of Washington Medicine (UWM), Intermountain Healthcare, and Kaiser Permanente Northern California, currently use this approach [5]. When used correctly, this approach prevents up to 40% of future asthma hospital encounters [4,6-9].

Due to limited capacity, a care management program can serve at most 3% of patients [10]. To maximize the effectiveness of these programs, an accurate predictive model should be used to identify the highest-risk patients. For this purpose, we recently developed a machine learning model powered by extreme gradient boosting (XGBoost) [11] on UWM data to predict which patients with asthma will have asthma hospital encounters in the succeeding 12 months [12]. Compared with previous models [5,13-26], this model is more accurate and improves the area under the receiver operating characteristic curve by ≥0.09. In addition, we previously developed a method to automatically explain the model's predictions in the form of rules and to suggest tailored interventions without sacrificing model performance [27,28]. Our method works for any black-box machine learning predictive model built on tabular data and addresses the interpretability issue that deters the widespread adoption of machine learning predictive models in clinical practice. Among all the published automated explanation methods for machine learning predictions [29,30], only our method can automatically recommend tailored interventions. For an average patient whom our UWM model correctly predicted to have future asthma hospital encounters, our method generated over 5000 rule-based explanations, if any [27]. The amount of nonredundant information in these explanations is usually two orders of magnitude less than the number of explanations, as multiple explanations often share some common components. The user of the automatic explanation function wants to quickly obtain the most useful information for a patient by viewing only the top few explanations. Therefore, we need to appropriately rank the explanations generated for each patient. Currently an open problem, procedures for appropriately ranking explanations are particularly important for the adoption of our automated explanation method in a busy clinical environment.

Objectives

To fill this gap, the aim of this study is to develop a method to appropriately rank the rule-based explanations generated by our automated explanation method [27,28] for a patient. We demonstrated our explanation ranking method in a test case that predicts asthma hospital encounters in patients with asthma.

Methods

Items Reused From Our Previous Papers

We reused the following items from our previous papers [12,27]: patient cohort, prediction target (ie, the dependent variable), features (ie, independent variables), data set, data preprocessing method, predictive model, cutoff threshold for binary classification, and automated explanation method. A list of symbols used in this paper is provided in Textbox 1.

XSL•FO
RenderX

Textbox 1. List of symbols.

List of Symbols

- C_r: confidence of the association rule r
- d: decay constant
- $f(d, p_i, r)$: exponential decay function computed for the feature-value pair item p_i on the left-hand side of the association rule r
- f: feature
- m: number of feature-value pair items on the left-hand side of an association rule
- $\max(v_r(x))$: maximum value of the variable $v_r(x)$ across all the rules found for the patient
- $\text{mean}(f(r))$: mean of $f(d, p_i, r)$ over all the feature-value pair items on the left-hand side of the association rule r
- $\min(v_r(x))$: minimum value of the variable $v_r(x)$ across all the rules found for the patient
- n: maximum number of top-ranked explanations that are allowed to be displayed initially
- norm(): normalization function
- N_r: number of feature-value pair items on the left-hand side of the association rule r
- p: feature-value pair item
- p_i: the i-th feature-value pair item on the left-hand side of an association rule
- q: number of association rules generated by our automated explanation method for the patient
- r: association rule
- $score_p$: ranking score of the feature-value pair item p
- $score_r$: ranking score of the association rule r
- S_r: commonality of the association rule r
- t, t_i: number of times that a feature-value pair item appears in the higher-ranked rules
- u: a value or a range
- v: outcome value
- $v_r(x)$: variable whose value on the association rule r is x
- w_a: weight for the term $\delta_{actionable}(r)$ in the rule scoring function
- w_b: weight for the term $\delta_{actionable}(p)$ in the item scoring function
- w_c: weight for the term $\text{norm}(C_r)$ in the rule scoring function
- w_d: weight for the term $\text{mean}(f(r))$ in the rule scoring function
- w_g: weight for the term $\exp(-d \cdot t)$ in the item scoring function
- w_n: weight for the term $\text{norm}(N_r)$ in the rule scoring function
- w_s: weight for the term $\text{norm}(\log_{10} S_r)$ in the rule scoring function
- x: value
- $\delta_{actionable}(p)$: indicator function for whether the feature-value pair item p is actionable
- $\delta_{actionable}(r)$: indicator function for whether the association rule r is actionable

Ethics Approval

The institutional review board of the UWM approved this secondary analysis retrospective cohort study.

Patient Cohort

In Washington State, the UWM is the largest academic health care system. Its enterprise data warehouse stores clinical and administrative data from 3 hospitals and 12 clinics for adults. The patient cohort included all adult patients with asthma (aged ≥18 years) who received care at any of these UWM facilities between 2011 and 2018. In a specific year, a patient was considered asthmatic if the patient had one or more asthma diagnosis codes (International Classification of Diseases [ICD], Tenth Revision: J45.x; ICD, Ninth Revision: 493.0x, 493.1x, 493.8x, 493.9x) documented in the encounter billing database during the year [13,31,32]. We excluded the patients who died during that year.

XSL•FO
RenderX

Prediction Target

Given a patient deemed asthmatic in an index year, we wanted to predict whether the patient would experience any asthma hospital encounter at the UWM in the succeeding 12 months, that is, any ED visit or inpatient stay at the UWM with asthma (ICD-10: J45.x; ICD-9: 493.0x, 493.1x, 493.8x, 493.9x) as its principal diagnosis. In predictive model training and testing, the patient's outcome in the succeeding 12 months was predicted using the patient's data until the end of the year.

Data Set

We used a structured administrative and clinical data set retrieved from the UWM's enterprise data warehouse. This data set contained information recorded for the visits by the patient cohort to the 12 clinics and 3 hospitals of the UWM over the 9-year span of 2011-2019. As the prediction target was for the following 12 months, the effective data in the data set spanned across the 8-year period of 2011-2018.

The Training and Test Set Split

We used the data from 2011 to 2017 as the training set to train the predictive model and to mine the association rules used by our automated explanation method. We used the data of 2018 as the test set to demonstrate our ranking method for the rule-based explanations generated by our automated explanation method.

Predictive Model and Features

Our UWM model used the XGBoost classification algorithm [11] and 71 features to predict the prediction target. As our UWM model was built on a single computer whose memory could hold the entire data set, the exact greedy algorithm was used to find the best split for tree learning in XGBoost [11]. These 71 features are listed in Table S2 in Multimedia Appendix 1 of our previous paper [12]. They were constructed based on the structured attributes in our data set and described various aspects of the patient's situation, such as demographics, encounters, diagnoses, laboratory tests, procedures, vital signs, and medications. An example feature is the patient's mean length of stay for an ED visit in the past year. Every input data instance to our predictive model includes these 71 features. Features that are the same as or similar to these 71 features were formerly used to predict asthma hospital encounters in patients with asthma and to provide automatic explanations on Intermountain Healthcare data as well as on Kaiser Permanente Southern California data [28,33-35]. For binary classification, we set the cutoff threshold at the top 10% of patients predicted to be at the highest risk. Our previous study [12] showed that on the test set, our model reached an area under the receiver operating characteristic curve of 0.902, an accuracy of 90.6% (13,268/14,644), a sensitivity of 70.2% (153/218), a specificity of 90.91% (13,115/14,426), a positive predictive value of 10.45% (153/1464), and a negative predictive value of 99.51% (13,115/13,180).

Review of Our Automated Explanation Method

Success Stories

Our automated explanation method [27,28] was designed as a general method that works for any machine learning predictive model built on tabular data. We initially demonstrated our method for predicting the diagnosis of type 2 diabetes [36]. Later, we successfully applied our method to predict asthma hospital encounters in patients with asthma on Intermountain Healthcare data [28], UWM data [27], and Kaiser Permanente Southern California data [34]. Other researchers have also successfully applied our method to project lung transplantation or death in patients with cystic fibrosis [37]; to project cardiac death in patients with cancer; and to use projections to manage heart transplant waiting list, posttransplant follow-ups, and preventive care in patients with cardiovascular diseases [38].

Main Idea

Our automated explanation method [27,28] uses class-based association rules [39,40] mined from historical data to explain a model's predictions and to recommend tailored interventions. As shown in Figure 1, the association rules are constructed separately from the predictive model and are used solely to provide explanations rather than to make predictions. Thus, our automated explanation method can work with any machine learning predictive model built on tabular data with no performance penalty. That is, our method falls into the category of model-agnostic explanation methods, which are widely used to automatically explain machine learning predictions [29,30].

Before rule mining starts, an automated discretizing method based on the minimum description length principle [40,41] is first applied to the training set to convert continuous features into categorical features. The association rules are then mined from the training set using a standard method, such as Apriori [39]. Each rule shows that a feature pattern is linked to an outcome value and has the form

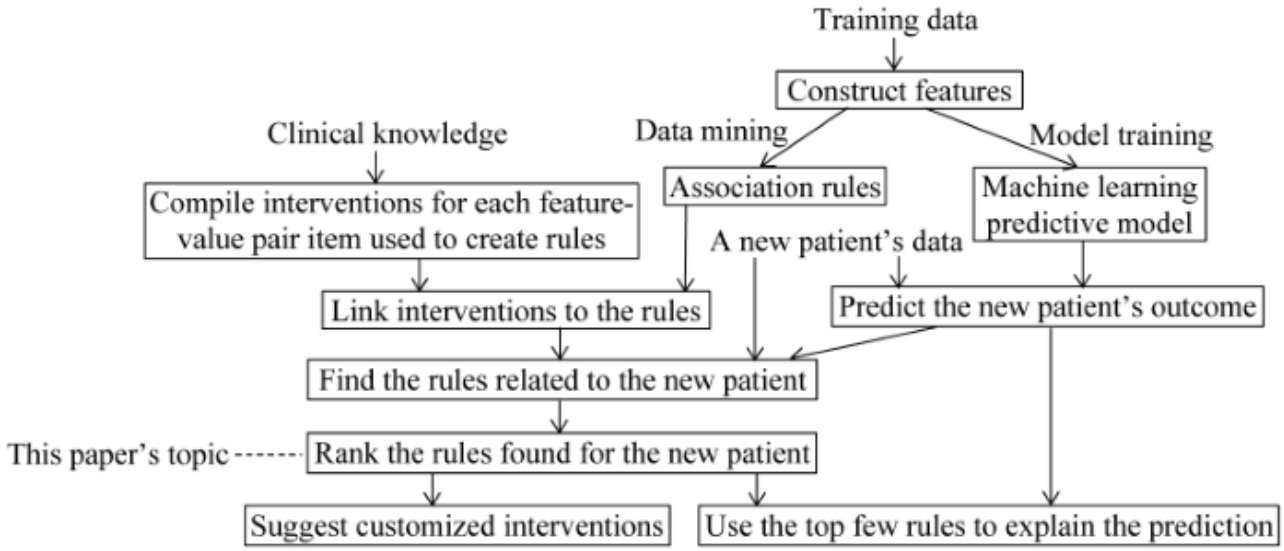
$$p_1 \text{ AND } p_2 \text{ AND } \ldots \text{AND } p_m \rightarrow v \quad \textbf{(1)}$$

Here, each item p_i ($1 \leq i \leq m$) is a feature-value pair (f, u). u is either the specific value of feature f or a range in which the value of f falls. For binary classification of a good versus a poor outcome, v is the poor outcome value; for example, the patient will have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months. For a patient fulfilling all of p_1, p_2, ..., and p_m, the rule indicates that the patient's outcome is likely to be v. An example rule is given below:

The patient had ≥13 ED visits in the past year AND the patient had ≥4 systemic corticosteroid prescriptions in the past year → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.

XSL•FO
RenderX

Figure 1. The flow diagram of our automated explanation method coupled with our explanation ranking method.



Constraints Put on the Association Rules

Our automated explanation method imposes several constraints on the association rules used by it. In this section, we review some of the constraints that are relevant to our explanation ranking method. For an association rule

$$p_1 \text{ AND } p_2 \text{ AND } ...\text{AND } p_m \rightarrow v, \quad \textbf{(2)}$$

commonality measures its coverage in the context of v; among all of the data instances linking to v, commonality is the percentage of data instances fulfilling p_1, p_2, ..., and p_m. Meanwhile, *confidence* measures its precision; among all of the data instances fulfilling p_1, p_2, ..., and p_m, the confidence is the percentage of data instances linking to v. For every association rule used by our automated explanation method, we require its commonality to be greater than or equal to a given minimum commonality threshold, such as 1%; its confidence to be greater than or equal to a given minimum confidence threshold, such as 50%; and its left-hand side to have no more than a given number (eg, 5) of feature-value pair items. As detailed in our previous papers [27,28], by setting the thresholds to these values, we can fulfill three goals concurrently. First, explanations can be given to most patients whom our UWM model correctly predicts as having ≥1 asthma hospital encounter in the succeeding 12 months. Second, the rule has sufficiently high confidence for the user of the automated explanation function to trust the rule. Third, no rule is overly complex.

The Explanation Method

For each feature-value pair item used to create association rules, a clinician in the development team of the automated explanation function precompiles 0 or more interventions. An item linking to at least one intervention is called actionable. The interventions related to the actionable items on the left-hand side of a rule are automatically linked to that rule. A rule linking to at least one intervention is called actionable.

For each patient predicted to have a poor outcome by the predictive model, the prediction is explained by the related association rules. For each such rule, the patient satisfies all of the feature-value pair items on its left-hand side. The poor outcome value appears on its right-hand side. Each rule delineates a reason for the patient's predicted poor outcome. Every actionable rule is displayed along with its linked interventions. The user of the automated explanation function can choose from these tailored interventions for the patient. The rules mined from the training set typically cover common reasons for having poor outcomes. Nonetheless, some patients could have poor outcomes due to rare reasons, such as the patient was prescribed between three and seven asthma medications during the past year AND the patient was prescribed ≥11 distinct medications during the past year AND the patient has some drug or material allergy AND the patient had ≥1 active problem in the problem list during the past year. Hence, our explanation method usually explains the predictions for most, though not all, of the patients correctly predicted by the model to have poor outcomes.

Ranking the Rule-Based Explanations Generated by Our Automated Explanation Method

Overview

For an average patient whom the predictive model predicts to have a poor outcome, our automated explanation method finds many related association rules, if any. Multiple rules often share some common feature-value pair items on their left-hand sides. To avoid overwhelming the user of the automated explanation function and to enable the user to quickly obtain the most useful information by viewing only the top few rules, we need to appropriately rank the rules found for a patient. As a rule often has a long description, a standard computer screen can show only a few rules simultaneously. To reduce the burden on the user, we present the rules in a manner similar to how a web search engine presents its search results for a keyword query. We chose a small number n, such as 3. The user can opt to change the value of n, for example, based on the size of the computer screen. If ≤n rules are found for the patient, we display all of these rules. Otherwise, if >n rules are found for the patient, we display the top n rules by default. If desired, the user can request to see more rules, for example, by dragging a vertical scroll bar or by clicking the *next page* button.

XSL•FO
RenderX

The main idea of our association rule ranking method is to consider multiple factors in the ranking process. The procedure incorporates these factors into a rule scoring function that strikes a balance among them and then ranks the rules found for a patient based on the scores computed for the rules in an iterative manner. In each iteration, the scores of the remaining rules are recomputed, and then, a rule is chosen from them. In the following, we describe our rule ranking method in detail.

Factors Considered in the Association Rule Ranking Process

When ranking the association rules found for a patient, we consider five factors:

1. *Factor 1*: All else being equal, a rule with a higher confidence is more precise and should rank higher.
2. *Factor 2*: All else being equal, a rule with a higher commonality covers a larger portion of patients with poor outcomes and should rank higher.
3. *Factor 3*: All else being equal, a rule with fewer feature-value pair items on its left-hand side is easier to comprehend and should rank higher.
4. *Factor 4*: In information retrieval, search engine users want to see diversified search results [42-44]. Similarly, the user of the automated explanation function wants to see diversified information in the top-ranked rules. Hence, all else being equal, a rule whose left-hand side has more items appearing in the higher-ranked rules should rank lower. The more times the items on the left-hand side of this rule appear in those rules, the lower this rule should rank.
5. *Factor 5*: The user of the automated explanation function wants to find suitable interventions for the patient. Thus, all else being equal, an actionable rule should rank higher than a nonactionable rule.

The Rule Scoring Function

We incorporate the five factors listed above into a rule scoring function to strike a balance among them. For an association rule

$$r\text{: } p_1 \text{ AND } p_2 \text{ AND } \ldots \text{AND } p_m \rightarrow v, \quad \textbf{(3)}$$

its ranking score is a linear combination of five terms, one per factor:

$$score_r = w_c \cdot \text{norm}(C_r) + w_s \cdot \text{norm}(\log_{10} S_r) - w_n \cdot \text{norm}(N_r) + w_d \cdot \text{mean}(f(r)) + w_a \cdot \delta_{actionable}(r) \quad \textbf{(4)}$$

At a high level,

1. C_r denotes r's confidence. The term norm(C_r) has a weight w_c>0 and addresses factor 1.
2. S_r denotes r's commonality. The term norm($\log_{10}S_r$) has a weight w_s>0 and addresses factor 2.
3. N_r denotes the number of feature-value pair items on r's left-hand side. The term norm(N_r) has a weight w_n>0 and addresses factor 3.
4. The term mean($f(r)$) has a weight w_d>0 and addresses factor 4. For each i ($1 \le i \le m$), the function $f(d, p_i, r)$ is computed based on the number of times the item p_i appears in the higher-ranked rules. The value of $f(d, p_i, r)$ is always between 0 and 1. Consequently, the value of mean($f(r)$) is always between 0 and 1.
5. The term $\delta_{actionable}(r)$ is the indicator function for whether r is actionable, has a weight w_a>0, and addresses factor 5.

Let $v_r(x)$ denote the variable, such as confidence, whose value on the association rule r is x. min($v_r(x)$) and max($v_r(x)$) denote the minimum and maximum values of $v_r(x)$ across all the rules found for the patient, respectively. If max($v_r(x)$)≠min($v_r(x)$), the function norm(x) [x−min($v_r(x)$)]/[max($v_r(x)$)−min($v_r(x)$)] normalizes x to a value between 0 and 1. If max($v_r(x)$)=min($v_r(x)$), all of the rules found for the patient have the same value of $v_r(x)$, and thus, there is no need to consider $v_r(x)$ in ranking these rules. In this case, norm(x) is set to 0.

C_r, $\log_{10}S_r$, and N_r have different value ranges. To make C_r, $\log_{10}S_r$, and N_r comparable with each other, we use norm() to put them into the same range of 0 to 1. mean($f(r)$) and $\delta_{actionable}(r)$ also fall within this range. To reflect that factors 1, 2, and 3 are equally important, we set the default values of w_c, w_s, and w_n to 1. To encourage the top-ranked rules to include diversified feature-value pair items, we wanted w_d's value to be >1 and set w_d's default value to 50. To strongly push the actionable rules to rank higher than the nonactionable rules, we wanted w_a's value to be ≫1 and set w_a's default value to 100. The value of w_a does not impact the score differences and, hence, the relative rankings among the actionable rules. When w_a is $>w_c+w_s+w_n+w_d$, the actionable rules always have larger scores than the nonactionable rules because norm(C_r), norm($\log_{10}S_r$), norm(N_r), and mean($f(r)$) are all between 0 and 1.

Detailed Description of the Five Terms Used in the Rule Scoring Function

In this section, we sequentially describe the five terms used in the rule scoring function in detail.

As norm() is a monotonically increasing function, all else being equal, the term norm(C_r) gives a larger ranking score to an association rule with a higher confidence C_r.

As shown in Figure 2, the commonality values for the association rules used by our automated explanation method have a skewed distribution. Most of the commonality values are clustered in the lower-value range. The commonality values of the rules generated by our automated explanation method for a patient are a sample from this distribution. We want the same weight w_s to work for different patients, regardless of how the sample is taken from this distribution. Thus, for every patient, we want the variance of the terms computed on the corresponding rules' commonality values to have approximately the same scale. For this purpose, we use the $\log_{10}$() function to transform the commonality values so that the resulting values are distributed more evenly than the raw values. As both norm() and $\log_{10}$() are monotonically increasing functions, norm($\log_{10}$()) is also a monotonically increasing function. All else being equal,

XSL•FO
RenderX

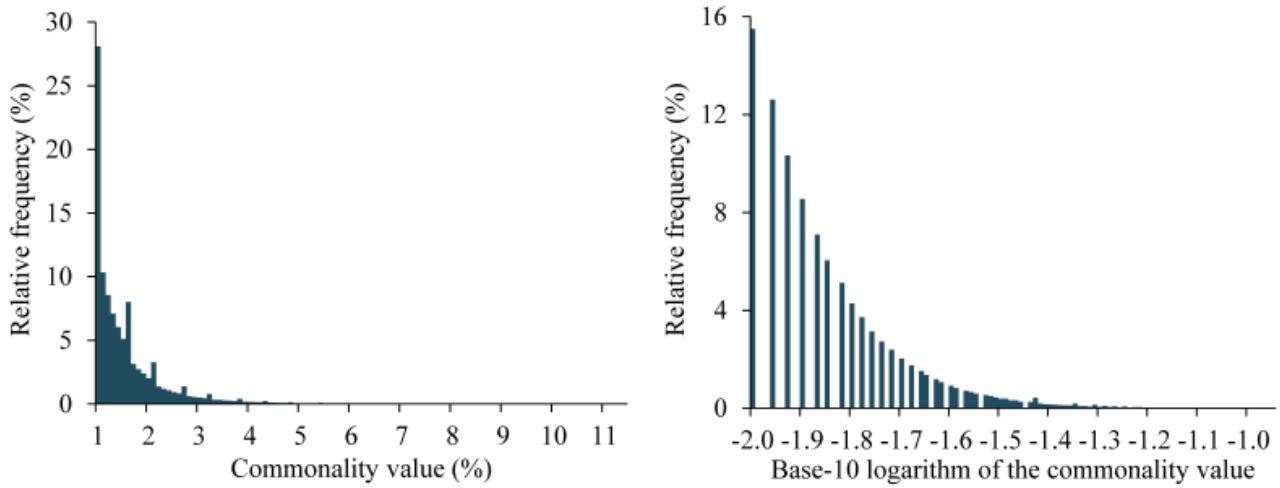
the term norm($\log_{10}S_r$) gives a larger ranking score to a rule with a higher commonality S_r.

As −norm() is a monotonically decreasing function, all else being equal, the term −norm(N_r) assigns a larger ranking score to an association rule with a smaller number N_r of feature-value pair items on its left-hand side.

In the k-th iteration of the association rule ranking process, the top k−1 rules have already been determined. We work on identifying the k-th ranked rule. For each feature-value pair item p_i on the left-hand side of a rule r that is found for the patient and whose rank has not yet been decided, we compute the exponential decay function $f(d, p_i, r)$ exp($-d \cdot t_i$). Here, d>0 is the decay constant, with a default value of 5. t_i is the number of times p_i appears in the top k−1 rules. A larger value of t_i results in a smaller value of $f(d, p_i, r)$. Recall that the term mean($f(r)$) is the mean of $f(d, p_i, r)$ over all the items on r's left-hand side. All else being equal, mean($f(r)$) assigns a smaller ranking score to a rule whose left-hand side has more items appearing in the top k−1 rules.

$\delta_{actionable}(r)$ is equal to 1 if the association rule r is actionable and is equal to 0 if r is nonactionable. All else being equal, the term $\delta_{actionable}(r)$ assigns a larger ranking score to an actionable rule compared with that of a nonactionable rule.

Figure 2. The distribution of the commonality values of all of the association rules used by our automated explanation method for predicting asthma hospital encounters in patients with asthma at the University of Washington Medicine.



The Iterative Association Rule Ranking Process

If only one association rule is found for a patient, there is no need to rank the rule. If ≥2 rules are found for the patient, we rank these rules iteratively. In the k-th iteration, we compute the ranking score for every rule r that is found for the patient and whose rank has not yet been determined. Compared with the case in the previous iteration, the score needs to be updated if and only if the value of mean($f(r)$) changes, that is, if and only if any feature-value pair item on r's left-hand side also appears on the left-hand side of the (k−1)-th ranked rule. Among all the rules that are found for the patient and whose ranks have not yet been determined, we select the rule with the highest score as the k-th ranked rule. If ≥2 of these rules have the same highest score, we choose one of them randomly as the k-th ranked rule.

For Each Association Rule on Display, Sort the Feature-Value Pair Items on Its Left-Hand Side

The same feature-value pair item could appear on the left-hand side of ≥2 top-ranked association rules. The user of the automated explanation function tends to read both the rules and the items on the left-hand side of a rule in the display order. To help the user obtain the most useful information as quickly as possible, for each rule on display, we need to appropriately rank the items on its left-hand side. For this purpose, we considered two factors:

1. *Factor 6*: The user wants to see new information as quickly as possible. Hence, all else being equal, an item for a rule that already appears in the higher-ranked rules should rank lower. As the number of times the item appears in higher-ranked rules increases, the rank of the item should decrease.
2. *Factor 7*: The user wants to find suitable interventions for the patient. Thus, all else being equal, an actionable item should rank higher than a nonactionable item.

We incorporate the two factors listed above into an item scoring function to strike a balance between them. Consider the k-th ranked association rule. For each feature-value pair item p on its left-hand side, p's ranking score is a linear combination of two terms, one per factor:

$$score_p = w_g \cdot \exp(-d \cdot t) + w_b \cdot \delta_{actionable}(p) \quad \textbf{(5)}$$

The terms in the equation above are further explained below:

1. In the equation for $score_p$ above, d is the same decay constant used in $f(d, p_i, r)$ in the rule scoring function. t is the number of times p appears in the top k−1 rules. The larger the value of t, the smaller the value of the exponential decay function exp($-d \cdot t$). Hence, all else being equal, the exp($-d \cdot t$) term assigns a smaller ranking score to an item that appears more times in the top k−1 rules. This addresses factor 6.

XSL•FO
RenderX

2. The term $\delta_{actionable}(p)$ is an indicator function for whether p is actionable. The term $\delta_{actionable}(p)$ is equal to 1 if p is actionable and is equal to 0 if p is nonactionable. All else being equal, the $\delta_{actionable}(p)$ term causes an actionable item to have a higher ranking score than that of a nonactionable item. This addresses factor 7.

Both $\exp(-d \cdot t)$ and $\delta_{actionable}(p)$ are between 0 and 1. For the weight $w_g>0$ of the term $\exp(-d \cdot t)$, we set its default value to 1. For the weight $w_b>0$ of the term $\delta_{actionable}(p)$, we set its default value to 2, which is >1. The value of w_b has no impact on the score differences and, hence, the relative ranking among the actionable items on the left-hand side of the association rule. When w_b is $>w_g$, the actionable items always have larger scores than those of the nonactionable items because $\exp(-d \cdot t)$ is between 0 and 1.

When the rank of an association rule is decided, we compute the ranking score for each feature-value pair item on the rule's left-hand side. We then sort these items in descending order of their scores. Items with the same score are randomly prescribed and given consecutive ranks.

Computer Coding Implementation

We used the R programming language to implement our explanation ranking method.

Providing Informative Examples of the Explanation Ranking Results

We want to demonstrate various aspects of the results produced by our explanation ranking method. For this purpose, we chose 8 patients with asthma in the test set, each of whom our UWM model correctly predicted to have ≥1 asthma hospital encounter in 2019, and our automated explanation method could explain this prediction. For each patient, we show the top three explanations produced by our explanation ranking method. Each patient satisfied one or more of the following conditions and was an informative case:

1. *Condition 1*: The patient had numerous encounters, laboratory tests, or medication prescriptions in 2018, reflecting a complex condition. In this case, we want to show how well the top three explanations capture and summarize the patient's key information related to asthma outcome prediction.
2. *Condition 2*: All or most of the asthma-related encounters that the patient had in 2018 were ED visits. Such a patient often had poor asthma control because of poor treatment adherence. In this case, we want to show how well the interventions linking to the top three explanations address the poor asthma control.
3. *Condition 3*: For each of the top three association rules produced for the patient, the rule's confidence value is close to the minimum confidence threshold. The rule's commonality value is close to the minimum commonality threshold. In this case, we want to illustrate these *borderline* rules. Recall that below either threshold, a rule will not be used by our automated explanation method.
4. *Condition 4*: The top three rules produced for the patient share several common feature-value pair items on their left-hand sides. This could happen, for example, when our automated explanation method finds only a few rules for the patient because the patient had only a small amount of information recorded in the electronic health record (EHR) system during the past year. In this case, we want to demonstrate the information redundancy in these rules.
5. *Condition 5*: A patient at high risk for future asthma hospital encounters often had ≥1 hospital encounter related to asthma during the past year. The patient being examined does not fall into this category. The patient had several feature values correlated with future asthma hospital encounters but no hospital encounter related to asthma during the past year. In this case, we want to show how well the top three explanations capture these feature values.

Sensitivity Analysis of the Parameters Used in the Rule Scoring Function

The rule scoring function uses six parameters whose default values are as follows: w_c=1, w_s=1, w_n=1, w_d=50, d=5, and w_a=100. To assess the impact of the five parameters w_c, w_s, w_n, w_d, and d on the association rule ranking results, we performed five experiments. In each experiment, we changed the value of one of these five parameters and kept the other parameters at their default values. In comparison with the case of all parameters taking their default values, we measured the average percentage change in the unique feature-value pair items contained in the top min(3, q) rules for a patient, where q denotes the number of rules generated by our automated explanation method for the patient. The percentage change in the unique items was defined as 100×the number of changed unique items divided by the number of unique items in the top min(3, q) rules. The average was taken over all patients in the test set, each of whom was predicted to have ≥1 asthma hospital encounter in 2019 and had at least one applicable rule (ie, q≥1). Multiple rules often differ from each other by only one item on their left-hand sides. In addition, switching items among the top few rules for a patient has little impact on the total amount of information that the user of the automated explanation function obtains from these rules. Thus, we measured the number of changed unique items in the top few rules per patient instead of the number of changed top rules per patient or the number of changed items per top rule.

As explained before, when w_a is $>w_c+w_s+w_n+w_d$, the actionable rules always rank higher than the nonactionable rules. Meanwhile, the concrete value of w_a has no impact on the ranking of the actionable rules. All the rules that our automated explanation method used on the UWM data set were actionable [27]. Thus, we did not perform a sensitivity analysis on w_a. For a similar reason, we did not perform a sensitivity analysis on the weights w_g and w_b used in the item scoring function.

Results

The Demographic and Clinical Characteristics of Our Patient Cohort

Each UWM data instance used in this study corresponds to a distinct patient and index year pair and is used to predict the

XSL•FO
RenderX

patient's outcome in the succeeding 12 months. Tables S1 and S2 in Multimedia Appendix 1 show our patient cohort's demographic and clinical characteristics during 2011-2017 and 2018 separately. These two sets of characteristics were similar to each other. During 2011-2017, 1.74% (1184/68,244) of data instances were linked to asthma hospital encounters in the succeeding 12 months. During 2018, 1.49% (218/14,644) of data instances were linked to asthma hospital encounters in the succeeding 12 months. A detailed comparison of these two sets of characteristics is presented in our previous paper [12].

Execution Time

For an average patient with asthma, our explanation ranking method took <0.01 seconds to produce the top three explanations. This is sufficiently fast for providing real-time clinical decision support.

Informative Examples of the Explanation Ranking Results

The Top Three Association Rules That Our Explanation Ranking Method Produced in Each Informative Example

The test set included 134 patients with asthma, each of whom our UWM model correctly predicted to have ≥1 asthma hospital encounter in 2019, and our automated explanation method could explain this prediction. To show the reader various aspects of the results produced by our explanation ranking method, we chose 8 of these patients who were informative cases. Tables 1-8 present the top three association rules that our explanation ranking method produced for each of the eight patients. For each of the top three rules produced for the seventh selected patient, Table 9 lists the interventions linked to the rule.

Table 1. The top three association rules that our explanation ranking method produced for the first selected patient (patient 1). This patient satisfied condition 1.

Rank	Association rule	Confidence of the rule		Commonality of the rule (n=1184), n (%)
		Total, n	Value, n (%)	
1	• The patient had 2 or 3 ED[a] visits related to asthma during the past year • AND the patient was prescribed between 7 and 11 distinct asthma medications during the past year • AND the patient was prescribed between 5 and 7 distinct asthma relievers during the past year • AND the patient had ≥1 active problem in the problem list during the past year • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	46	24 (52.17)	24 (2.03)
2	• The patient's mean length of stay of an ED visit during the past year was >0.205 day • AND the patient was prescribed ≥4 systemic corticosteroids during the past year • AND the patient's most recent ED visit related to asthma occurred no less than 26 days ago and no more than 100 days ago • AND the patient was prescribed 2 distinct nebulizer medications during the past year • AND the patient is not a White patient • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	28	14 (50)	14 (1.18)
3	• The patient was prescribed nebulizer medications ≥8 times during the past year • AND the patient had ≥5 no shows during the past year • AND the patient had 2 or 3 ED visits related to asthma during the past year • AND the patient's mean temperature during the past year was ≤98.09 Fahrenheit • AND the patient is ≤54 years old • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	32	18 (56.25)	18 (1.52)

[a]ED: emergency department.

XSL•FO
RenderX

Table 2. The top three association rules that our explanation ranking method produced for the second selected patient (patient 2). This patient satisfied condition 1.

Rank	Association rule	Confidence of the rule		Commonality of the rule (n=1184), n (%)
		Total, n	Value, n (%)	
1	• The patient's most recent diagnosis of asthma with acute exacerbation or status asthmaticus was from ≤110 days ago • AND the patient was prescribed ≥10 short-acting β-2 agonists during the past year • AND the patient had no outpatient visit during the past year • AND the patient's first encounter related to asthma was from ≥1 year ago • → The patient will likely have ≥1 inpatient stay or ED[a] visit for asthma in the succeeding 12 months.	87	54 (62.07)	54 (4.56)
2	• The patient was prescribed asthma medications ≥16 times during the past year • AND the patient's mean respiratory rate during the past year was >16.89 breaths per minute • AND the patient's most recent visit was an ED visit • AND the patient is a Black or an African American patient • AND the patient was totally allowed between 1 and 33 medication refills during the past year • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	32	18 (56.25)	18 (1.52)
3	• The patient had between 8 and 16 asthma diagnoses during the past year • AND the patient's lowest SpO_2[b] level during the past year was between 8.0% and 94.5% • AND the patient's most recent ED visit related to asthma occurred no less than 26 days ago and no more than 100 days ago • AND the patient is not a White patient • AND the patient had ≤6 encounters during the past year • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	35	18 (51.43)	18 (1.52)

[a]ED: emergency department.

[b]SpO_2: peripheral capillary oxygen saturation.

XSL•FO
RenderX

Table 3. The top three association rules that our explanation ranking method produced for the third selected patient (patient 3). This patient satisfied condition 1.

Rank	Association rule	Confidence of the rule		Commonality of the rule (n=1184), n (%)
		Total, n	Value, n (%)	
1	• The patient's most recent diagnosis of asthma with acute exacerbation or status asthmaticus was from ≤110 days ago • AND the patient's most recent visit was an ED[a] visit • AND the patient had between 9 and 17 primary or principal asthma diagnoses during the past year • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	127	79 (62.2)	79 (6.67)
2	• The patient had between 17 and 27 asthma diagnoses during the past year • AND the patient's most recent visit was an ED visit • AND the patient had no visit to the primary care provider during the past year • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	68	38 (55.88)	38 (3.21)
3	• The patient was prescribed ≥10 short-acting β-2 agonists during the past year • AND the highest severity of all asthma diagnoses of the patient during the past year was moderate or severe persistent asthma • AND the patient was allowed ≥34 medication refills during the past year • AND the patient is ≤54 years old • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	40	20 (50)	20 (1.69)

[a]ED: emergency department.

Table 4. The top three association rules that our explanation ranking method produced for the fourth selected patient (patient 4). This patient satisfied condition 2.

Rank	Association rule	Confidence of the rule		Commonality of the rule (n=1184), n (%)
		Total, n	Value, n (%)	
1	• The patient had ≥7 ED[a] visits related to asthma during the past year • AND the patient is single • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	37	34 (91.89)	34 (2.87)
2	• The patient had between 9 and 17 primary or principal asthma diagnoses during the past year • AND the patient's most recent outpatient visit related to asthma was from ≥365 days ago • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	105	66 (62.86)	66 (5.57)
3	• The patient had ≥28 asthma diagnoses during the past year • AND the patient had no outpatient visit during the past year • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	19	16 (84.21)	16 (1.35)

[a]ED: emergency department.

XSL•FO
RenderX

Table 5. The top three association rules that our explanation ranking method produced for the fifth selected patient (patient 5). This patient satisfied condition 5.

Rank	Association rule	Confidence of the rule		Commonality of the rule (n=1184), n (%)
		Total, n	Value, n (%)	
1	• The patient had ≥20 diagnoses of asthma with acute exacerbation during the past year • AND the patient was prescribed ≥10 short-acting β-2 agonists during the past year • → The patient will likely have ≥1 inpatient stay or ED[a] visit for asthma in the succeeding 12 months.	82	48 (58.54)	48 (4.05)
2	• The patient had ≥28 asthma diagnoses during the past year • AND the patient was prescribed nebulizer medications ≥8 times during the past year • AND the patient had no outpatient visit to the primary care provider during the past year • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	55	37 (67.27)	37 (3.13)
3	• The patient had ≥18 primary or principal asthma diagnoses during the past year • AND the patient was prescribed ≥8 distinct asthma relievers during the past year • AND the patient's mean heart rate during the past year was >80 beats per minute • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	116	58 (50)	58 (4.9)

[a]ED: emergency department.

Table 6. The top three association rules that our explanation ranking method produced for the sixth selected patient (patient 6). This patient satisfied conditions 3 and 4.

Rank	Association rule	Confidence of the rule		Commonality of the rule (n=1184), n (%)
		Total, n	Value, n (%)	
1	• The patient had 2 or 3 ED[a] visits related to asthma during the past year • AND the patient's most recent outpatient visit related to asthma was from ≤104 days ago • AND the patient was prescribed ≤2 inhaled corticosteroids during the past year • AND the patient is ≤54 years old • AND the patient's relative change of weight during the past year was ≤3% • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	40	22 (55)	22 (1.86)
2	• The patient had between 3 and 8 diagnoses of asthma with (acute) exacerbation during the past year • AND the patient had 2 or 3 ED visits related to asthma during the past year • AND the patient is not a White patient • AND the patient was prescribed ≤2 distinct asthma medications during the past year • AND the patient is single • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	25	14 (56)	14 (1.18)
3	• The patient's most recent outpatient visit related to asthma was from ≤104 days ago • AND the patient had 2 or 3 ED visits related to asthma during the past year • AND the patient was prescribed ≥1 unit of medications during the past year • AND the patient had no public insurance on the last day of the past year • AND the patient had between 1 and 13 outpatient visits during the past year • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	32	16 (50)	16 (1.35)

[a]ED: emergency department.

XSL•FO
RenderX

Table 7. The top three association rules that our explanation ranking method produced for the seventh selected patient (patient 7). This patient satisfied conditions 1 and 2.

Rank	Association rule	Confidence of the rule		Commonality of the rule (n=1184), n (%)
		Total, n	Value, n (%)	
1	• The patient had ≥7 ED[a] visits related to asthma during the past year • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	51	39 (76.47)	39 (3.29)
2	• The patient had between 17 and 27 asthma diagnoses during the past year • AND the patient had no outpatient visit during the past year • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	48	28 (58.33)	28 (2.36)
3	• The patient's mean length of stay of an ED visit during the past year was between 0.025 and 0.205 day • AND the patient had ≥3 ED visits during the past year • AND the patient was prescribed ≥3 asthma relievers that are neither short-acting β-2 agonists nor systemic corticosteroids during the past year • AND the patient was prescribed ≥4 systemic corticosteroids during the past year • AND the patient is single • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	116	58 (50)	58 (4.9)

[a]ED: emergency department.

Table 8. The top three association rules that our explanation ranking method produced for the eighth selected patient (patient 8). This patient satisfied condition 5.

Rank	Association rule	Confidence of the rule		Commonality of the rule (n=1184), n (%)
		Total, n	Value, n (%)	
1	• The patient had between 9 and 17 primary or principal asthma diagnoses during the past year • AND the patient was prescribed asthma medications ≥16 times during the past year • AND the patient had no outpatient visit to the primary care provider during the past year • AND the patient is not a White patient • → The patient will likely have ≥1 inpatient stay or ED[a] visit for asthma in the succeeding 12 months.	87	45 (51.72)	45 (3.8)
2	• For the patient's most recent visit, the time from making the request to the actual visit was ≤0.6 day • AND the patient was prescribed asthma medications ≥16 times during the past year • AND the patient is a Black or an African American patient • AND the patient's first encounter related to asthma was from ≥1 year ago • AND the patient's lowest SpO_2[b] level during the past year was between 94.5% and 95.5% • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	19	12 (63.16)	12 (1.01)
3	• The patient was prescribed ≥12 distinct asthma medications during the past year • AND the patient had ≥12 encounters during the past year • AND the patient's most recent outpatient visit related to asthma was from ≤104 days ago • AND the patient had ≤82 laboratory tests during the past year • AND the patient is not a White patient • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	19	12 (63.16)	12 (1.01)

[a]ED: emergency department.

[b]SpO_2: peripheral capillary oxygen saturation.

XSL•FO
RenderX

Table 9. The interventions linked to each of the top three association rules that our explanation ranking method produced for patient 7.

Rank	Association rule	Linked interventions
1	• The patient had ≥7 ED[a] visits related to asthma during the past year • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	• An intervention linked to the item "the patient had ≥7 ED visits related to asthma during the past year" is to use control strategies to prevent needing emergency care.
2	• The patient had between 17 and 27 asthma diagnoses during the past year • AND the patient had no outpatient visit during the past year • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	• An intervention linked to the item "the patient had between 17 and 27 asthma diagnoses during the past year" is to give the patient suggestions on how to improve asthma control. • An intervention linked to the item "the patient had no outpatient visit during the past year" is to make sure that the patient has a primary care provider and to suggest the patient to regularly visit the provider.
3	• The patient's mean length of stay of an ED visit during the past year was between 0.025 and 0.205 day • AND the patient had ≥3 ED visits during the past year • AND the patient was prescribed ≥3 asthma relievers that are neither short-acting β-2 agonists nor systemic corticosteroids during the past year • AND the patient was prescribed ≥4 systemic corticosteroids during the past year • AND the patient is single • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.	• An intervention linked to the items "the patient's mean length of stay of an ED visit during the past year was between 0.025 and 0.205 day" and "the patient had ≥3 ED visits during the past year" is to use control strategies to prevent needing emergency care. • An intervention linked to the items "the patient was prescribed ≥3 asthma relievers that are neither short-acting β-2 agonists nor systemic corticosteroids during the past year" and "the patient was prescribed ≥4 systemic corticosteroids during the past year" is to tailor the prescribed asthma medications, to help the patient adhere to asthma controllers, and to improve avoidance of triggers.

[a]ED: emergency department.

As illustrated by the cases shown in Tables 1-9, the top few explanations that our explanation ranking method produces for a patient offer five benefits for clinical decision support. We describe these five benefits sequentially in the following sections.

Benefit 1: The Top Few Explanations Provide Succinct Summaries on a Wide Range of Aspects of the Patient's Situation

To make good clinical decisions for a patient, the clinician needs to understand the patient's situation well. For each of the eight selected patients, the top three rule-based explanations produced by our explanation ranking method provide succinct summaries on a wide range of aspects of the patient's situation, such as demographics, encounters, vital signs, laboratory tests, and medications. From these summaries, the user of the automated explanation function can quickly gain a comprehensive understanding of the patient's situation related to the prediction target. This saves the user a significant amount of time and effort. In comparison, to gain this understanding in a clinical setting, even if a clinician knows all of the features needed for this purpose, the clinician currently often needs to spend a significant amount of time laboriously checking many pages of information scattered in various places in the EHR system and performing manual calculations. For example, patient 1 had a total of >1000 encounters recorded in the EHR system at the UWM over time. In 2018, this patient had 164 encounters, only two of which were related to asthma, and both were ED visits. As Table 1 shows, the statistics of two ED visits related to asthma are reflected by the first item on the left-hand side of the first association rule produced for this patient. As another example, in 2018, patient 2 had 740 medication prescriptions, 153 of which were asthma medication prescriptions covering a total of 72 short-acting β-2 agonists. As Table 2 shows, the statistic of 72 short-acting β-2 agonists is reflected by the first item on the left-hand side of the first rule produced for this patient. The statistics of 153 asthma medication prescriptions are reflected by the first item on the left-hand side of the second rule produced for this patient. The cases with the other items on the left-hand sides of the top three rules produced for these two patients were similar.

To gain a comprehensive understanding of a patient's situation quickly, a clinician could ask the patient to describe his or her situation. However, the patient often cannot perform this well. For example, patients 1, 3, and 7 had severe mental disorders, which affected their memory and ability to describe their situation. This was a common scenario. Over 29.99% (4393/14,644) of patients with asthma at the UWM have mental disorders. Moreover, when making clinical decisions, the clinician does not always have direct access to the patient. For instance, when identifying candidate patients for care management, care managers are sitting in a back office and cannot talk to patients. In either of these two cases, the summaries provided by the top few rule-based explanations can help the clinician gain an understanding of the patient.

Benefit 2: Showing the Top Few Explanations Can Save the User of the Automated Explanation Function From Having to Manually Think of Many Features Summarizing the Patient's Situation and Computing Their Values

Often, many features must be used to adequately summarize a patient's situation related to the prediction target. In a busy

XSL•FO
RenderX

clinical environment, a clinician cannot be expected to enumerate all of these features in a short amount of time. The top few rule-based explanations that our explanation ranking method produces for a patient cover the values of various features summarizing the patient's situation related to the prediction target. This saves the user of the automated explanation function from having to manually think of these features and to compute their values.

Benefit 3: The Top Few Explanations Can Provide Information Not Easily Obtainable From Using the Existing Search and Browsing Functions of the EHR System to Check the Patient's Data

The EHR system provides some browsing and basic search functions. However, for certain important features summarizing a patient's situation related to the prediction target, we cannot easily obtain their values by using these functions to check the patient's EHR data. The top few rule-based explanations that our explanation ranking method produces for a patient cover the values of several such features. This saves the user of the automated explanation function a significant amount of work. For example, many different asthma medications exist. In 2018, patient 2 had 740 medication prescriptions. It is difficult and time-consuming to manually compute the number of asthma medication prescriptions and the total number of short-acting β-2 agonists prescribed for this patient in 2018. In comparison, as mentioned before, these two statistics are directly reflected by the first and second rules produced for this patient. As a second example, in 2018, patient 7 had 14 ED visits, eight of which were related to asthma. For two of these eight ED visits, asthma was not the primary diagnosis. To compute the patient's number of ED visits related to asthma in 2018, a clinician needs to find all of the patient's ED visits in 2018 and check each of them to see whether it has an asthma diagnosis code. This requires a nontrivial amount of time. In comparison, as Table 7 shows, the statistics of eight ED visits related to asthma are directly reflected by the first item on the left-hand side of the first rule produced for this patient. As a third example, in 2018, patient 8 had 12 outpatient visits, none of which was to the patient's primary care provider. To compute the patient's number of outpatient visits to the primary care provider, a clinician needs to find all of the patient's outpatient visits in 2018 and manually check each of them to see whether it involved the patient's primary care provider. This requires a nontrivial amount of time. In comparison, as Table 8 shows, the third item on the left-hand side of the first rule produced for this patient directly shows that the patient had 0 outpatient visits to the primary care provider in 2018.

Benefit 4: The Top Few Explanations Can Help the User of the Automated Explanation Function Avoid Overlooking Certain Important Information of the Patient and Discover Errors in the Data Recorded on the Patient in the EHR System

A patient with asthma often has several other diseases, which could distract the clinicians and cause them to pay insufficient attention to the patient's asthma and record incorrect data on the patient in the EHR system. For example, in 2018, asthmatic patient 3 also had major depression disorder, anxiety, posttraumatic stress disorder, visual disturbance, chronic pain, and knee osteoarthritis. In the patient's problem list, these diseases were recorded as major problems, whereas asthma was recorded as a minor problem. However, the patient had 15 primary asthma diagnoses, some of which were severe persistent asthma and indicated that asthma was a major problem for the patient at that time. In 2020, asthma was first recorded as two major problems in the patient's problem list: one on asthma exacerbation and another on persistent asthma with status asthmaticus. As shown in Table 3, the first and third rules produced for the patient covered the patient's number of asthma diagnoses and the highest severity of these diagnoses in 2018, reflecting that the patient had severe persistent asthma at that time. This can help the user of the automated explanation function avoid overlooking this aspect and discover that asthma should be recorded as a major problem in the patient's problem list in 2018.

Benefit 5: The Top Few Explanations Can Help the User of the Automated Explanation Function Identify Certain Problems of the Patient Not Easily Findable in the EHR System

This can help the user of the automated explanation function identify suitable interventions for the patient. For example, as shown in Table 6, the first and second rules produced for patient 6 showed that this patient had quite a few ED visits related to asthma; however, very few asthma medications were prescribed for this patient in 2018. This patient did not adhere to albuterol prescriptions due to personal preference. Realizing this, the user could consider adopting the intervention of replacing albuterol with some other asthma medications that the patient is willing to take. As another example, as shown in Tables 4 and 7, for patients 4 and 7, the top three rules produced for each patient revealed that the patient had many ED visits related to asthma but no outpatient visit in 2018. These two patients were found to be homeless. With this information, the user could consider providing social resources to reduce the socioeconomic burden of homelessness, which leads to ineffective access to health care.

Description of the 5 Example Patient Cases, One Case Per Each of Conditions 1-5

In this section, for each of conditions 1-5, we choose one example patient satisfying it and show how this patient was an informative case.

As an example case for condition 1, patient 1 had 164 encounters and 644 medication prescriptions in 2018. As shown in Table 1, the top three explanations produced for this patient effectively capture and summarize various aspects of the patient's key information related to future asthma hospital encounters.

As an example case for condition 2, patient 7 had eight asthma-related encounters in 2018, all of which were ED visits. As shown in Table 7, the top three explanations produced for this patient revealed that the patient had many asthma diagnoses, had no outpatient visit, and was prescribed ≥4 systemic corticosteroids during 2018, reflecting poor asthma control. As shown in Table 9, the interventions linked to the top three

XSL•FO
RenderX

explanations address various aspects related to poor asthma control.

Patient 6 provides an example for condition 3. As shown in Table 6, for each of the top three association rules produced for this patient, the rule's confidence value is close to the minimum confidence threshold of 50%, and the rule's commonality value is close to the minimum commonality threshold of 1%. These three rules cover a wide range of aspects of the patient's situation, including demographics, encounters, diagnoses, vital signs, and medications.

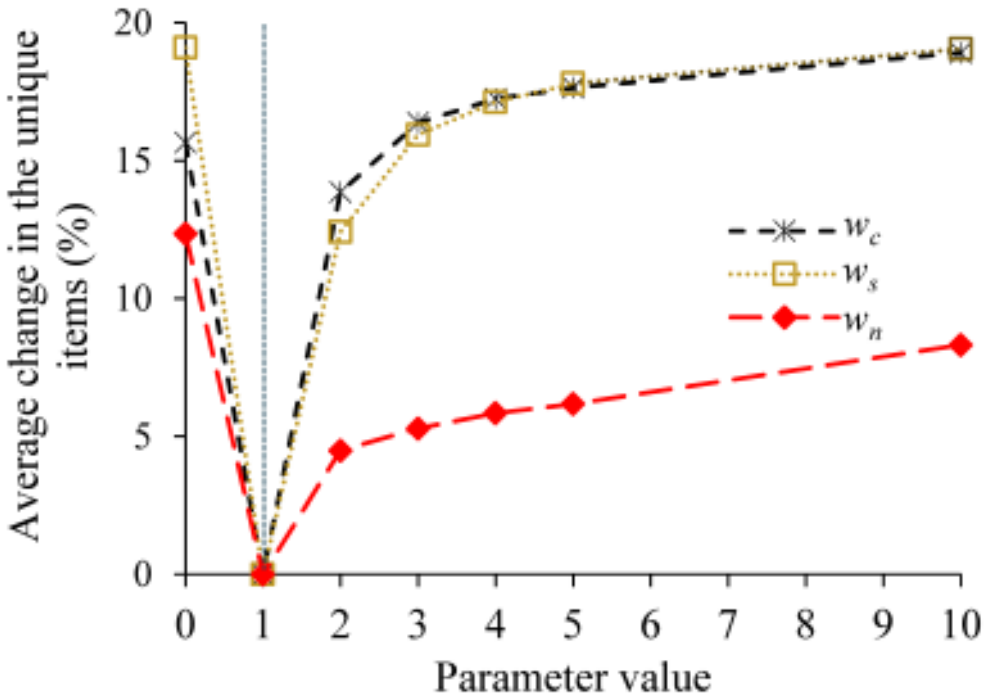
As an example case for condition 4, patient 6 had only three encounters and one medication order, and subsequently, a small amount of information was recorded for this patient in the EHR system in 2018. As shown in Table 6, the top three explanations produced for this patient share three common feature-value pair items on their left-hand sides. Despite having moderate information redundancy, these explanations still cover a wide range of aspects of the patient's situation, including demographics, encounters, diagnoses, vital signs, and medications.

As an example case for condition 5, patient 8 had no hospital encounters related to asthma in 2018. As shown in Table 8, the top three explanations produced for this patient capture several feature values of the patient correlated with future asthma hospital encounters, such as the patient having between 9 and 17 primary or principal asthma diagnoses during the past year, the patient having ≥16 asthma medication prescriptions during the past year, the patient having no outpatient visit to the primary care provider during the past year, and the patient having ≥12 encounters during the past year.

Sensitivity Analysis Results of the Parameters Used in the Rule Scoring Function

We performed 5 sensitivity analysis experiments, 1 for each of the 5 parameters w_c, w_s, w_n, w_d, and d used in the rule scoring function. In each experiment, we changed the corresponding parameter's value and kept the other parameters at their default values. In comparison with the case where all 5 parameters took their default values and for each of these 5 parameters, Figures 3-5 show the average percentage change in the unique feature-value pair items contained in the top min(3, q) association rules for a patient versus the parameter's value. In each figure, the vertical dotted line represents the default value of the corresponding parameter. For each parameter value tested, the average percentage change in the unique items was relatively small (<20%). The only exception is the case of either w_d=0 or d=0, where the average percentage change in the unique items was 43.57% (453.18/1040). In both cases, our explanation ranking method ignores the need for the top-ranked rules to provide diversified information (factor 4).

Figure 3. In comparison with the case where all five parameters took their default values and for each of the three parameters w_c, w_s, and w_n, the average percentage change in the unique feature-value pair items contained in the top min (3, q) association rules for a patient versus the parameter's value. The vertical dotted line represents the default value of w_c, w_s, and w_n.



XSL•FO
RenderX

Figure 4. In comparison with the case where all five parameters took their default values, the average percentage change in the unique feature-value pair items contained in the top min (3, q) association rules for a patient versus the value of the parameter w_d. The vertical dotted line represents the default value of w_d.

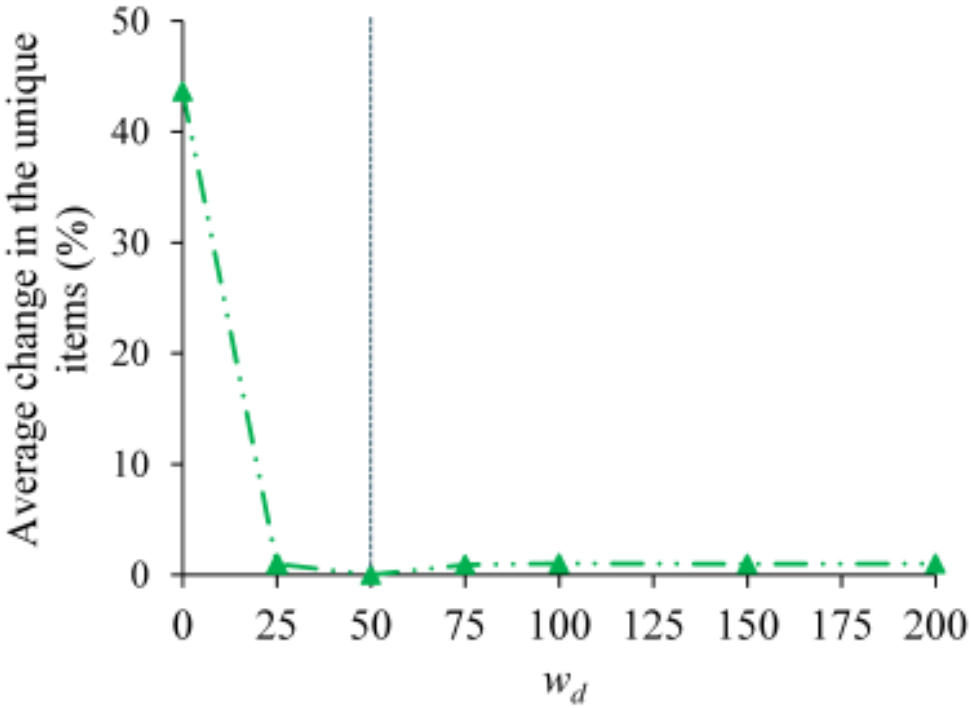
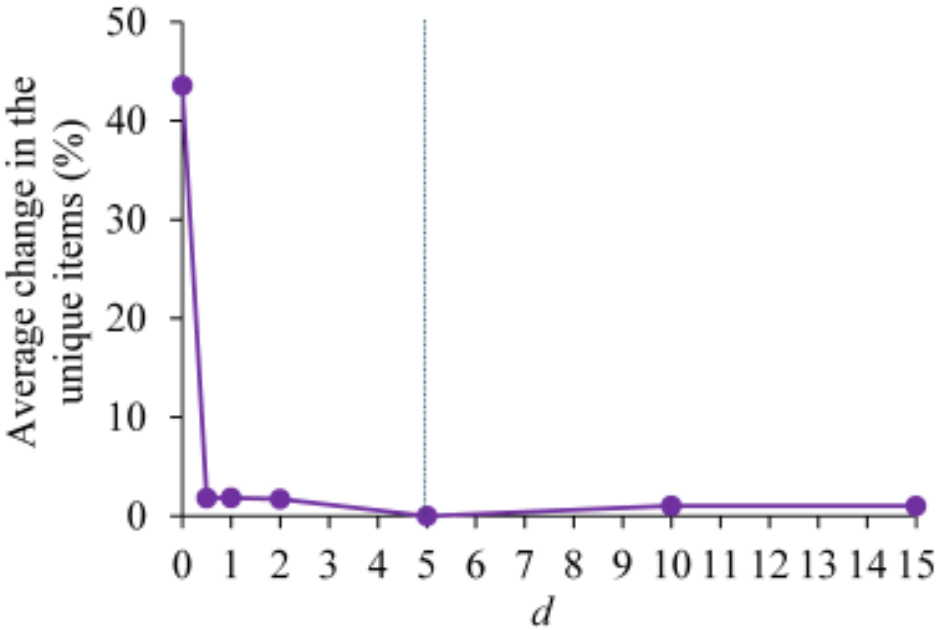


Figure 5. In comparison with the case where all five parameters took their default values, the average percentage change in the unique feature-value pair items contained in the top min (3, q) association rules for a patient versus the value of the parameter d. The vertical dotted line represents the default value of d.



Discussion

Principal Findings

In a busy clinical environment, the explanation ranking module is essential for our automated explanation function for machine learning predictions to provide high-quality real-time decision support. For an average patient with asthma correctly predicted by our UWM model to have future asthma hospital encounters, our automated explanation method generated over 5000 rule-based explanations, if any. Within a negligible amount of time, our explanation ranking method can appropriately rank them and return the few highest-ranked explanations. These few explanations typically have high quality and low redundancy. From these few explanations, the user of the automated explanation function can gain useful insights on various aspects of the patient's situation. Many of these insights cannot be easily obtained by viewing the patient's data in the

XSL•FO
RenderX

current EHR system. With further improvements in model accuracy, our UWM model coupled with our automated explanation method and our explanation ranking method could be deployed to better guide the use of asthma care management to save costs and improve patient outcomes.

Similar to our automated explanation method, our explanation ranking method is general purpose and does not rely on any specific property of a particular prediction target, disease, patient cohort, or health care system. Our automated explanation method coupled with our explanation ranking method can be used for any predictive modeling problem on any tabular data set. This provides a unique solution to the interpretability issue that deters the widespread adoption of machine learning predictive models in clinical practice.

In our sensitivity analysis, when we changed any parameter used in our explanation ranking method from its default value, the resulting average percentage change in the unique feature-value pair items contained in the top min(3, q) association rules for a patient was typically <20%. This is not a large change, as most (>80%) of the distinct feature-value pair items contained in these rules and, subsequently, most of the information seen by the user of the automated explanation function remain the same. For instance, if the top min(3, q) association rules contain 15 unique feature-value pair items, at most three of these feature-value pair items would vary due to the change in the parameter value, whereas the other 12 or more remain the same as before. Thus, each parameter used in our explanation ranking method has a reasonably large stable range, within which the top few explanations produced by our method do not vary greatly as the parameter value changes. The default value of the parameter was within this stable range. According to our test results, the stable ranges are 0 to 10 for w_c, 0 to 10 for w_s, 0 to 10 for w_n, 25 to 200 for w_d, and 0.5 to 15 for d.

Adjusting Certain Parameters Used in the Rule Scoring and the Item Scoring Functions

Both the rule scoring and item scoring functions have several parameters. On the basis of the preferences of the users of the automated explanation function and the specific needs of the particular health care application, the developer of the automated explanation function could change some of these parameters from their default values. In the UWM test case used in this study, all association rules used by our automated explanation method were actionable. For some other predictive modeling problems, certain rules used by our automated explanation method are nonactionable [36]. In this case, if we want to allow some nonactionable rules to rank higher than some non-top-scored actionable rules on any patient, we need to reduce the weight w_a. Similarly, if we want to allow some nonactionable items to rank higher than some actionable items in any non-top-scored rule that our automated explanation method finds for any patient, we need to reduce the weight w_b.

Considerations on the Threshold That Is Used to Determine the Top Rules That Will Be Displayed by Default

Different patients have different distributions of the ranking scores for the association rules found for the patients. No single threshold on the ranking score works for all patients. Thus, we use a threshold on the number of rules rather than a threshold on the ranking score to determine the top rules that will be displayed by default. This is similar to the case with a web search engine such as Google. Google does not use any ranking score threshold to determine the search results that will be displayed on each search result page. Instead, by default, Google displays 10 search results on each search result page. The user can request to see more search results by clicking the *next* button.

Considerations Regarding Potential Clinical Use

Understanding how a predictive model works requires a global interpretation. Understanding a single prediction of a model requires only local interpretation [29,30]. Our automated explanation method provides local interpretations. For clinical applications, the user of the automated explanation function is frequently a clinician who has little or no background in machine learning, can see only the prediction results but not the internal of the machine learning predictive model, cares about understanding the prediction on an individual patient but not much about how the predictive model works internally, and possibly does not even know which predictive model is used because the model is often embedded in the clinical software. In this case, it does not matter whether the explanations provided by the automated explanation function match how the predictive model works internally, as long as the explanations can help the user understand the prediction for a specific patient. For a patient predicted to have a poor outcome, our automated explanation method will give the same set of explanations regardless of which machine learning model is used to make the prediction. In the case where a deep learning model built on longitudinal data is used to make predictions, we can use the method proposed in our paper [45] to extract temporal features from the deep learning model and longitudinal data, use these temporal features to convert longitudinal data to tabular data, and then apply our automated explanation method to a predictive model built on the tabular data.

To use our automated explanation method in clinical practice, we could implement our automated explanation method together with our explanation ranking method as a software library with an application programming interface. For any clinical decision support software that uses a machine learning predictive model, we could use the application programming interface to add the automated explanation function into the software to explain the model's predictions.

Related Work

As surveyed in the book written by Molnar [29] and the previous papers written by several research groups [30,46-48], other researchers have proposed many automated methods to explain machine learning predictions. Some of these methods are used for traditional machine learning algorithms, whereas others are specifically designed for deep learning algorithms [48]. The explanations given by most of these methods are not in a rule form. Many of these methods can handle only a specific machine learning algorithm or degrade the performance measures of the predictive model. None of these methods can automatically suggest tailored interventions. Ribeiro et al [49] and Rudin and

XSL•FO
RenderX

Shaposhnik [50] used rules to explain any machine learning model's predictions automatically. However, automatically recommending tailored interventions is still beyond the reach of the methods proposed by Ribeiro et al [49] and Rudin and Shaposhnik [50], as the rules are not generated until the prediction time. In comparison, our automated explanation method mines the association rules before the prediction time, provides rule-based explanations, works for any machine learning predictive model built on tabular data, does not degrade model performance, and automatically recommends tailored interventions. Compared with other types of explanations, rule-based explanations can more directly recommend tailored interventions and are easier to understand.

As surveyed in previous studies [39,51,52], association rules have been used in various applications to discover interesting patterns in the data and to make predictions. Various methods have been proposed to rank the rules mined from a data set for these purposes [39,51-55]. In comparison, we mine and rank association rules to automatically explain machine learning predictions and to recommend tailored interventions.

Limitations

This work has three limitations that are excellent areas for future work:

1. This study used data from a single health care system. In the future, it would be beneficial to test our explanation ranking method on data from other health care systems.
2. This study tested our explanation ranking method for predicting one specific target in one disease. In the future, it would be beneficial to test our method on predictive modeling problems that address other prediction targets and diseases.
3. The data set used in this work contains no information on patients' encounters outside the UWM. This forced us to limit the prediction target to asthma hospital encounters at the UWM rather than asthma hospital encounters in any health care system. In addition, the features used in this study were computed solely from the data recorded for the patients' encounters at the UWM. In the future, it would be worth investigating how the top few explanations produced by our explanation ranking method would differ if we have data on the patients' encounters in other health care systems.

Conclusions

In this study, we developed a method to rank the rule-based explanations generated by our automated explanation method for machine learning predictions. Within a negligible amount of time, our explanation ranking method ranks the explanations and returns the few highest-ranked explanations. These few explanations typically have high quality and low redundancy. Many of them provide useful insights on the various aspects of the patient's situation, which cannot be easily obtained by viewing the patient's data in the current EHR system. Both our automated explanation method and our explanation ranking method are designed based on general computer science principles and rely on no special property of any specific disease, prediction target, patient cohort, or health care system. Although only tested in the case of predicting asthma hospital encounters in patients with asthma, our explanation ranking method is general and can be used for any predictive modeling problem on any tabular data set. The explanation ranking module is an essential component of the automated explanation function, which addresses the interpretability issue that deters the widespread adoption of machine learning predictive models in clinical practice. In the next few years, we plan to test our explanation ranking method on predictive modeling problems addressing other diseases as well as on data from other health care systems.

Acknowledgments

The authors thank Brian Kelly for useful discussions. GL was partially supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number R01HL142503. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

XZ participated in designing the study, conducting a literature review, writing the paper's first draft, performing the computer coding implementation, and conducting experiments. GL conceptualized and designed the study, conducted a literature review, and rewrote the entire paper. Both authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
A summary of the demographic and clinical characteristics of patients with asthma at the University of Washington Medicine.
[PDF File (Adobe PDF File), 94 KB - medinform_v9i8e28287_app1.pdf]

References

1. Most recent National Asthma Data. Centers for Disease Control and Prevention. 2020. URL: https://www.cdc.gov/asthma/most_recent_national_asthma_data.htm [accessed 2021-01-29]

XSL•FO
RenderX

2. Chronic respiratory diseases: asthma. World Health Organization. 2020. URL: https://www.who.int/news-room/q-a-detail/chronic-respiratory-diseases-asthma [accessed 2021-01-31]
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. Mays GP, Claxton G, White J. Managed care rebound? Recent changes in health plans' cost containment strategies. Health Aff (Millwood) 2004;Suppl Web Exclusives:427-436 [FREE Full text] [doi: 10.1377/hlthaff.w4.427] [Medline: 15451964]
5. 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]
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 [FREE Full text]
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: KDD'16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA, USA 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 RJ, Debray TP, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Schermer TR, et al. Exacerbations in adults with asthma: a systematic review and external validation of prediction models. J Allergy Clin Immunol Pract 2018;6(6):1942-1952. [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]

XSL•FO
RenderX

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. 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]
28. Luo G, Johnson MD, Nkoy FL, He S, Stone BL. Automatically explaining machine learning prediction results on asthma hospital visits in asthmatic patients: secondary analysis. JMIR Med Inform 2020 Dec 31;8(12):e21965 [FREE Full text] [doi: 10.2196/21965] [Medline: 33382379]
29. Molnar C. Interpretable Machine Learning. Morrisville, NC: lulu.com; 2020.
30. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv 2019 Jan 23;51(5):93. [doi: 10.1145/3236009]
31. Desai JR, Wu P, Nichols GA, Lieu TA, O'Connor PJ. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. Med Care 2012 Jul;50 Suppl:30-35. [doi: 10.1097/MLR.0b013e318259c011] [Medline: 22692256]
32. Wakefield DB, Cloutier MM. Modifications to HEDIS and CSTE algorithms improve case recognition of pediatric asthma. Pediatr Pulmonol 2006 Oct;41(10):962-971. [doi: 10.1002/ppul.20476] [Medline: 16871628]
33. 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]
34. Luo G, Nau CL, Crawford WW, Schatz M, Zeiger RS, Koebnick C. Generalizability of an automatic explanation method for machine learning prediction results on asthma-related hospital visits in patients with asthma: quantitative analysis. J Med Internet Res 2021 Apr 15;23(4):e24153 [FREE Full text] [doi: 10.2196/24153] [Medline: 33856359]
35. 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]
36. Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. Health Inf Sci Syst 2016;4:2 [FREE Full text] [doi: 10.1186/s13755-016-0015-4] [Medline: 26958341]
37. Alaa AM, van der Schaar M. Prognostication and risk factors for cystic fibrosis via automated machine learning. Sci Rep 2018 Jul 26;8(1):11242 [FREE Full text] [doi: 10.1038/s41598-018-29523-2] [Medline: 30050169]
38. Alaa AM, van der Schaar M. AutoPrognosis: automated clinical prognostic modeling via Bayesian optimization with structured kernel learning. In: Proceedings of 35th International Conference on Machine Learning. 2018 Presented at: ICML'18: 35th International Conference on Machine Learning; July 10-15, 2018; Stockholm, Sweden p. 139-148.
39. Thabtah FA. A review of associative classification mining. The Knowledge Engineering Review 2007 Mar 01;22(1):37-65. [doi: 10.1017/s0269888907001026]
40. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. In: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. 1998 Presented at: KDD'98: 4th International Conference on Knowledge Discovery and Data Mining; August 27-31, 1998; New York City, NY p. 80-86.
41. Fayyad UM, Irani KB. Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence. 1993 Presented at: IJCAI'93: 13th International Joint Conference on Artificial Intelligence; August 28-September 3, 1993; Chambéry, France p. 1022-1029.
42. Luo G, Thomas SB, Tang C. Automatic home medical product recommendation. J Med Syst 2012 Apr;36(2):383-398. [doi: 10.1007/s10916-010-9483-2] [Medline: 20703712]
43. Luo G, Tang C, Yang H, Wei X. MedSearch: a specialized search engine for medical information retrieval. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. 2008 Presented at: CIKM'08: Conference on Information and Knowledge Management; October 26-30, 2008; Napa Valley, CA, USA p. 143-152. [doi: 10.1145/1458082.1458104]
44. Santos RL, Macdonald C, Ounis I. Search result diversification. Foundations and Trends in Information Retrieval 2015;9(1):1-90. [doi: 10.1561/1500000040]
45. 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]
46. Du M, Liu N, Hu X. Techniques for interpretable machine learning. Commun ACM 2020;63(1):68-77. [doi: 10.1145/3359786]
47. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. In: Proceedings of the 5th IEEE International Conference on Data Science and Advanced Analytics.

XSL•FO
RenderX

2018 Presented at: DSAA'18: IEEE 5th International Conference on Data Science and Advanced Analytics; October 1-3, 2018; Turin, Italy p. 80-89. [doi: 10.1109/DSAA.2018.00018]
48. Samek W, Montavon G, Lapuschkin S, Anders CJ, Muller K. Explaining deep neural networks and beyond: a review of methods and applications. Proc IEEE 2021 Mar;109(3):247-278. [doi: 10.1109/jproc.2021.3060483]
49. Ribeiro MT, Singh S, Guestrin C. Anchors: high-precision model-agnostic explanations. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018 Presented at: AAAI'18: 32nd AAAI Conference on Artificial Intelligence; February 2-7, 2018; New Orleans, LA p. 1527-1535.
50. Rudin C, Shaposhnik Y. Globally-consistent rule-based summary-explanations for machine learning models: application to credit-risk evaluation. In: Proceedings of INFORMS 11th Conference on Information Systems and Technology. 2019 Presented at: CIST'19: 11th Conference on Information Systems and Technology; October 19-20, 2019; Seattle, WA p. 1-19. [doi: 10.2139/ssrn.3395422]
51. Altaf W, Shahbaz M, Guergachi A. Applications of association rule mining in health informatics: a survey. Artif Intell Rev 2017;47(3):313-340. [doi: 10.1007/s10462-016-9483-9]
52. Pazhanikumar K, Arumugaperumal S. Association rule mining and medical application: a detailed survey. Int J Comput Appl 2013 Oct 18;80(17):10-19. [doi: 10.5120/13967-1698]
53. Yang G, Shimada K, Mabu S, Hirasawa K. A personalized association rule ranking method based on semantic similarity and evolutionary computation. In: Proceedings of the IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence). 2008 Presented at: CEC'08: IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence); June 1-6, 2008; Hong Kong, China p. 487-494. [doi: 10.1109/CEC.2008.4630842]
54. Bouker S, Saidi R, Yahia SB, Nguifo EM. Ranking and selecting association rules based on dominance relationship. In: Proceedings of the IEEE 24th International Conference on Tools with Artificial Intelligence. 2012 Presented at: ICTAI'12: IEEE 24th International Conference on Tools with Artificial Intelligence; November 7-9, 2012; Athens, Greece p. 658-665. [doi: 10.1109/ICTAI.2012.94]
55. Chen MC. Ranking discovered rules from data mining with multiple criteria by data envelopment analysis. Expert Syst Appl 2007 Nov;33(4):1110-1116. [doi: 10.1016/j.eswa.2006.08.007]

Abbreviations

ED: emergency department
EHR: electronic health record
ICD: International Classification of Diseases
UWM: University of Washington Medicine
XGBoost: extreme gradient boosting

Edited by C Lovis; submitted 06.03.21; peer-reviewed by P Elkin, A Rovetta; comments to author 17.05.21; revised version received 19.05.21; accepted 06.06.21; published 11.08.21.

Please cite as:
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;9(8):e28287
URL: https://medinform.jmir.org/2021/8/e28287
doi:10.2196/28287
PMID:34383673

©Xiaoyi Zhang, Gang Luo. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 11.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Current-Visit and Next-Visit Prediction for Fatty Liver Disease With a Large-Scale Dataset: Model Development and Performance Comparison

Cheng-Tse Wu[1], MS; Ta-Wei Chu[2,3], MD, PhD; Jyh-Shing Roger Jang[1], PhD

[1]Department of Computer Science & Information Engineering, National Taiwan University, Taipei, Taiwan

[2]Department of Obstetrics and Gynecology, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[3]MJ Health Screening Center, Taipei, Taiwan

Corresponding Author:
Ta-Wei Chu, MD, PhD
Department of Obstetrics and Gynecology
Tri-Service General Hospital, National Defense Medical Center
No. 325, Sec. 2, Chenggong Rd., Neihu Dist.,
Taipei, 114
Taiwan
Phone: 886 287923311 ext 88083
Email: taweichu@gmail.com

Abstract

Background: Fatty liver disease (FLD) arises from the accumulation of fat in the liver and may cause liver inflammation, which, if not well controlled, may develop into liver fibrosis, cirrhosis, or even hepatocellular carcinoma.

Objective: We describe the construction of machine-learning models for current-visit prediction (CVP), which can help physicians obtain more information for accurate diagnosis, and next-visit prediction (NVP), which can help physicians provide potential high-risk patients with advice to effectively prevent FLD.

Methods: The large-scale and high-dimensional dataset used in this study comes from Taipei MJ Health Research Foundation in Taiwan. We used one-pass ranking and sequential forward selection (SFS) for feature selection in FLD prediction. For CVP, we explored multiple models, including k-nearest-neighbor classifier (KNNC), Adaboost, support vector machine (SVM), logistic regression (LR), random forest (RF), Gaussian naïve Bayes (GNB), decision trees C4.5 (C4.5), and classification and regression trees (CART). For NVP, we used long short-term memory (LSTM) and several of its variants as sequence classifiers that use various input sets for prediction. Model performance was evaluated based on two criteria: the accuracy of the test set and the intersection over union/coverage between the features selected by one-pass ranking/SFS and by domain experts. The accuracy, precision, recall, F-measure, and area under the receiver operating characteristic curve were calculated for both CVP and NVP for males and females, respectively.

Results: After data cleaning, the dataset included 34,856 and 31,394 unique visits respectively for males and females for the period 2009-2016. The test accuracy of CVP using KNNC, Adaboost, SVM, LR, RF, GNB, C4.5, and CART was respectively 84.28%, 83.84%, 82.22%, 82.21%, 76.03%, 75.78%, and 75.53%. The test accuracy of NVP using LSTM, bidirectional LSTM (biLSTM), Stack-LSTM, Stack-biLSTM, and Attention-LSTM was respectively 76.54%, 76.66%, 77.23%, 76.84%, and 77.31% for fixed-interval features, and was 79.29%, 79.12%, 79.32%, 79.29%, and 78.36%, respectively, for variable-interval features.

Conclusions: This study explored a large-scale FLD dataset with high dimensionality. We developed FLD prediction models for CVP and NVP. We also implemented efficient feature selection schemes for current- and next-visit prediction to compare the automatically selected features with expert-selected features. In particular, NVP emerged as more valuable from the viewpoint of preventive medicine. For NVP, we propose use of feature set 2 (with variable intervals), which is more compact and flexible. We have also tested several variants of LSTM in combination with two feature sets to identify the best match for male and female FLD prediction. More specifically, the best model for males was Stack-LSTM using feature set 2 (with 79.32% accuracy), whereas the best model for females was LSTM using feature set 1 (with 81.90% accuracy).

(JMIR Med Inform 2021;9(8):e26398) doi:10.2196/26398

XSL•FO
RenderX

KEYWORDS

machine learning; sequence forward selection; one-pass ranking; fatty liver diseases; alcohol fatty liver disease; nonalcoholic fatty liver disease; long short-term memory; current-visit prediction; next-visit prediction

Introduction

Background

Prior research on the use of machine learning for early disease prediction has focused on diabetes, fatty liver disease (FLD), hypotension, and other metabolic syndromes [1]. This study focused on the prediction of FLD, which is widespread in Taiwan, and could lead to liver cirrhosis, fibrosis, and liver cell death. If left untreated for up to 3 years, FLD has a 25% chance of developing into nonalcoholic steatohepatitis and a 10%-15% chance of developing into liver cirrhosis [2,3]. Moreover, FLD increases the prevalence of diabetes, metabolic syndrome, and obesity, creating enormous medical and economic burdens for society. This situation raises an urgent need for early and precise prediction, followed by personalized treatment and lifestyle management. Typically, FLD has been classified into two types according to its cause: alcohol-related fatty liver disease (AFLD) and nonalcoholic fatty liver disease (NAFLD). AFLD is commonly caused by excessive alcohol consumption, whereas NAFLD is due to other more complex factors. Although most prior research has focused on NAFLD prediction rather than AFLD prediction [4-8], there is no inherent reason to conduct separate prediction processes. The previous focus on NAFLD is partly due to the datasets used being insufficiently large to predict both types of FLD. Previous studies have relied on leave-one-out (LOO) cross-validation to avoid overfitting [4-10] on these small datasets. Some prior studies have performed feature selection through human intervention rather than automatic selection [7,11-14], although this is not a common practice in machine learning.

Recently, machine learning has been used extensively in medicine and health care. Dealing with large datasets with many features requires efficient methods to reduce the computing time. We adopted one-pass ranking (OPR) for automatic feature selection, with accuracy similar to the features selected by sequential forward selection (SFS). OPR enables finding good features for current-visit prediction (CVP) and next-visit prediction (NVP). The contributions of this paper can be summarized as follows. First, we compared the performance of OPR and SFS for automatic feature selection, demonstrating that OPR offers great efficiency with decent accuracy when dealing with a large-dimensional dataset. Second, in addition to CVP, we propose the task of NVP, which is much more important for practicing preventive medicine. To our knowledge, this is the first attempt to perform NVP on FLD. Third, we modeled NVP as a sequence classification problem and proposed two feature sets with fixed or variable intervals for the long short-term memory (LSTM) classifier and some of its variants. Before describing the study, we first provide a review of some important prior work on FLD prediction along with a brief overview of automatic feature selection in machine learning.

Related Work

Literature Survey

Table 1 summarizes the differences between this study and prior research. The dataset used in this study is much larger and covers a much longer period. All of the prior research [4-8,11] summarized in Table 1 used smaller datasets, with sample sizes ranging from less than 100 to 11,000 individuals, covering periods ranging from less than 1 year to 2 years at most. Furthermore, most of these studies only used male data for analysis, such as Jamali et al [5], Yip et al [8], and Wu et al [7], with data sizes below 600 individuals. Although Birjandi et al [4], Islam et al [11], and Ma et al [6] used both male and female data for analysis, their data sizes were at most 11,000 individuals, which is still much smaller than the dataset used in this study. The dataset used in this study is far larger than other datasets reported in the literature, and is thus suitable for separate construction of male and female models, which are much more robust and reliable.

Table 1. Comparison of prior research and this study for fatty liver disease (FLD) prediction.

Reference	Sample size	Years of study	Feature selection	FLD type	Gender	Next-visit prediction	Data source
Birjandi et al [4]	<1700	2012	Yes	NAFLD[a]	Male/Female	No	Health screening centers
Jamali et al [5]	<100	2012-2014	No	NAFLD	Male	No	Hospital
Yip et al [8]	<1000	2015	Yes	NAFLD	Male	No	Hospital
Islam et al [11]	<1000	2012-2013	Yes	NAFLD/AFLD[b]	Male/Female	No	Hospital
Ma et al [6]	<11,000	2010	Yes	NAFLD	Male/Female	No	Hospital
Wu et al [7]	<600	2009	No	NAFLD/AFLD	Male	No	Hospital
This study	>150,000	2009-2016	Yes	NAFLD/AFLD	Male/Female	Yes	Health screening dataset

[a]NAFLD: nonalcoholic fatty liver disease.

[b]AFLD: alcoholic fatty liver disease.

In various application domains, LSTM has proven to be the state-of-the-art sequence classifier that can achieve better results than classical methods. For instance, Kim et al [15] developed an epidemic disease spread and economic situation model based

XSL•FO
RenderX

on LSTM to predict the economic impact of future COVID-19 spread. Pal et al [16] proposed an LSTM framework to predict a country-based COVID-19 risk category at a given time with a dataset from 180 countries. Zhang et al [17] used LSTM to reproduce soil stress-strain behavior, demonstrating better accuracy than other models. For stock price prediction, Sunny et al [18] proposed an LSTM-based framework to forecast stock trends with high accuracy. In surface-guided radiation therapy, Wang et al [19] created a framework to predict internal liver motion signals and external respiratory motion signals, finding that LSTM can achieve better results. Moreover, Qiao et al [20] proposed a high-precision LSTM model to monitor mooring line responses by using the vessel motion as input. The superior performance of LSTM in previous studies motivated us to use this approach for NVP in the context of FLD prediction.

Automatic Feature Selection

Automatic feature selection is an important step in machine learning, since it can identify a feature subset to construct a better model while requiring less computing time for training and testing. Automatic feature selection methods can be divided into three categories: wrappers, filters, and embedded methods. Wrapper methods use a classifier to score the feature subsets, which produces accurate results but is time-consuming. Filter methods use a proxy measure instead of accuracy to score a feature subset, which is efficient but does not always produce a good model since the proxy measure does not always relate to classification accuracy [21]. Embedded methods perform feature selection as part of the model construction process, which tends to lie between wrappers and filters in terms of accuracy and computational complexity [22,23]. This study used more accurate wrapper methods for feature selection, including OPR and SFS [24].

Not all approaches covered in the literature use the wrapper methods for feature selection. For example, as shown in in Table 1, Wu et al [7] manually selected only 10 predictor variables, including age, gender, systolic blood pressure, diastolic blood pressure, abdominal girth, glucose AC, triglyceride, high-density lipoprotein cholesterol, serum glutamic-oxaloacetic transaminase-aspartate aminotransferase, and serum glutamic-pyruvic transaminase-alanine aminotransferase, and then derived their weights by information gain without further verifying their ranking by classification accuracy.

Common Classifiers Used in This Study

This study used different conventional classifiers for CVP, including Adaboost [25], support vector machine (SVM) [26], logistic regression (LR) [27], random forest (RF) [28,29], Gaussian naïve Bayes (GNB) [30], decision tree C4.5 [31], and classification and regression trees (CART) [32]. For NVP, since the input is a variable-length sequence, we used LSTM [33], bidirectional LSTM (biLSTM) [34], Stack-LSTM [35], Stack-biLSTM [36], and Attention-LSTM [37].

Methods

Study Design and Process

Flowchart

This study explored feature selection schemes for CVP and NVP, and proposes two feature sets for NVP using LSTM. Figure 1 shows the flowchart for FLD prediction. First, we needed to perform data preprocessing and cleaning, which is covered in further detail in the Dataset subsection below. We then used different feature selection methods and different classifiers for the two prediction types (CVP and NVP). As shown in Figure 2, we used automatic feature selection (such as OPR or SFS) to select the most critical features from a given classifier, including K-nearest neighbor classification (KNNC), and then adopted a procedure for performance evaluation (such as k-fold cross-validation). Following feature selection, we constructed other more complicated models for prediction and evaluation.

XSL•FO
RenderX

Figure 1. Flowchart of current-visit prediction and next-visit prediction for fatty liver disease (FLD).

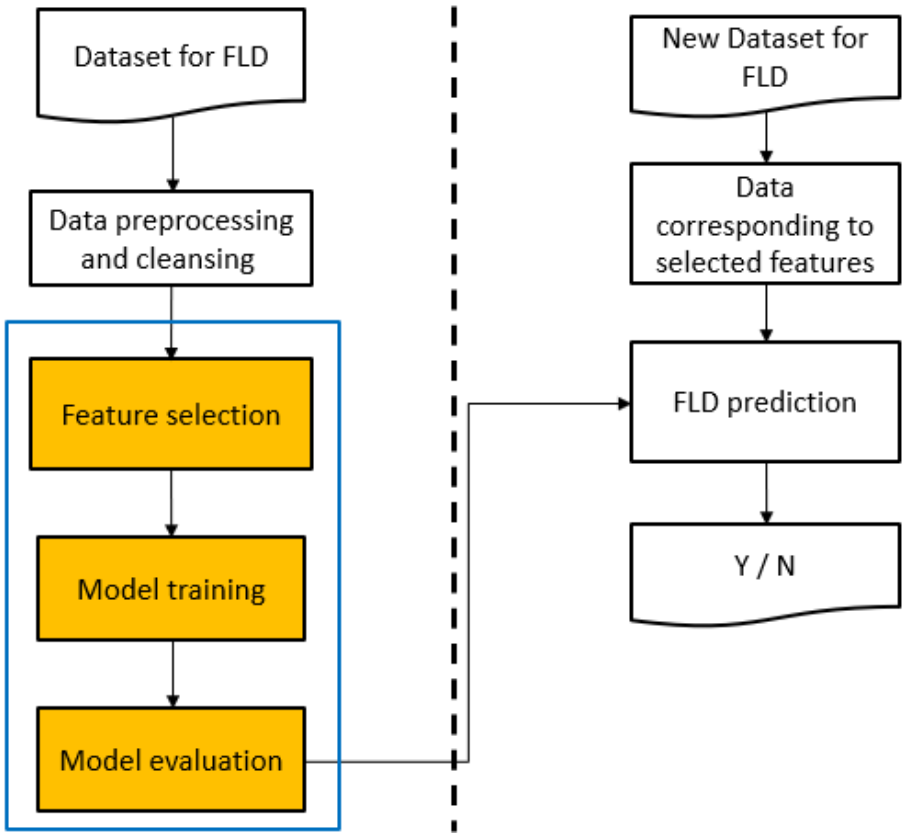
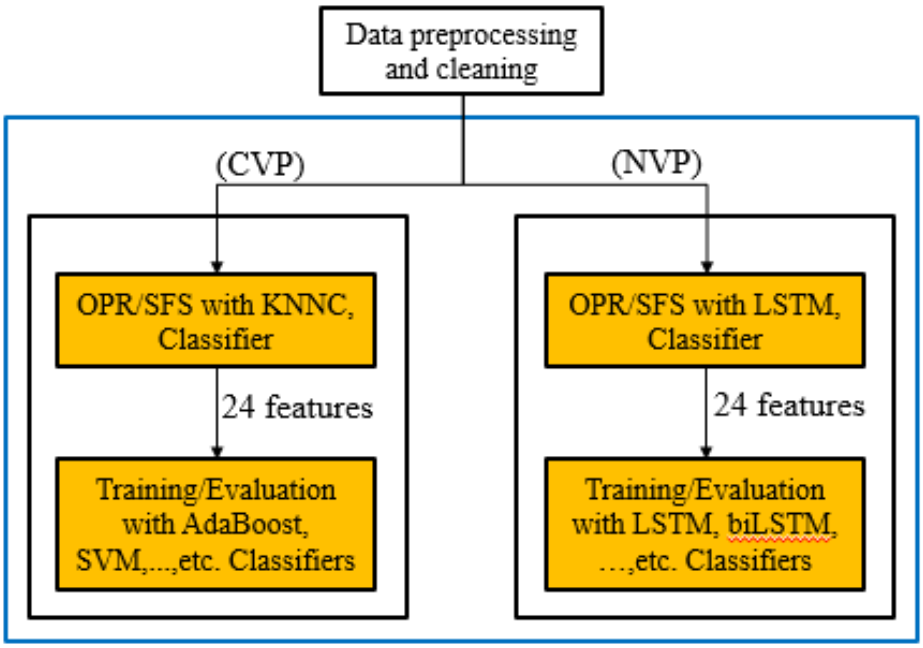


Figure 2. Flowchart of current-visit prediction (CVP) and next-visit prediction (NVP) for fatty liver disease (FLD) with different classifiers. OPR: one-pass ranking; SFS: sequential forward selection; KNNC: k-nearest neighbor classifier; SVM: support vector machine; LSTM: long short-term memory; biLSTM: bidirectional long short-term memory.



CVP Model

Although fatty liver has no special symptoms, there is a certain chance that fatty hepatitis will develop in the long term, and it may progress to serious liver diseases such as cirrhosis, liver failure, and even liver cancer [38,39]. Through the CVP model, the risk of FLD can be predicted directly. For those with a low FLD risk, there is no need to spend time and money in arranging abdominal ultrasound examinations. However, groups with a high risk of FLD are recommended to receive an abdominal ultrasound for early detection and prevention of significant liver diseases. Therefore, CVP can achieve the goal of rapid screening with timely and appropriate intervention, if necessary.

XSL•FO
RenderX

For this task, CVP uses a classifier with all important information (including lab and questionnaire results) at the current visit as inputs to predict whether or not the patient currently has FLD. Correct execution of CVP with selected features can help the doctor better understand what features are more likely to contribute to FLD. Sufficiently high CVP accuracy allows patients with a low FLD risk to forego a time-consuming and costly abdominal ultrasound. That is, CVP can be used for rapid screening at medical clinics that do not have the equipment or specialists needed to manually diagnose FLD. This can effectively reduce staff and equipment requirements at clinics and hospitals, which is of particularly importance in the era of the COVID-19 pandemic.

For CVP feature selection, we used two wrapper-based methods, OPR and SFS, with a simple classifier of KNNC and LOO cross-validation for performance evaluation. Following this rapid feature selection, we used the selected features for model training and evaluation with other advanced classifiers, including Adaboost, SVM, LR, RF, GNB, decision trees C4.5, and CART.

NVP Model

Early prediction also plays an essential role in disease prevention, especially for chronic diseases. With NVP, our system can even predict the next visit result, allowing physicians to arrange abdominal ultrasound examinations or other appropriate interventions for patients with a high future risk of FLD. For this task, we used a sequence classifier with all historical information (up to the current visit) as inputs to predict whether or not the patient will be diagnosed with FLD at the next visit. NVP is more important than CVP from the perspective of preventive medicine. If the patient is predicted to have a high probability of FLD risk at the next visit, the physician can suggest lifestyle changes (eg, diet, smoking, alcohol consumption) to effectively modify the key features that contribute to FLD in NVP, along with other appropriate interventions, including abdominal ultrasound at the next health check.

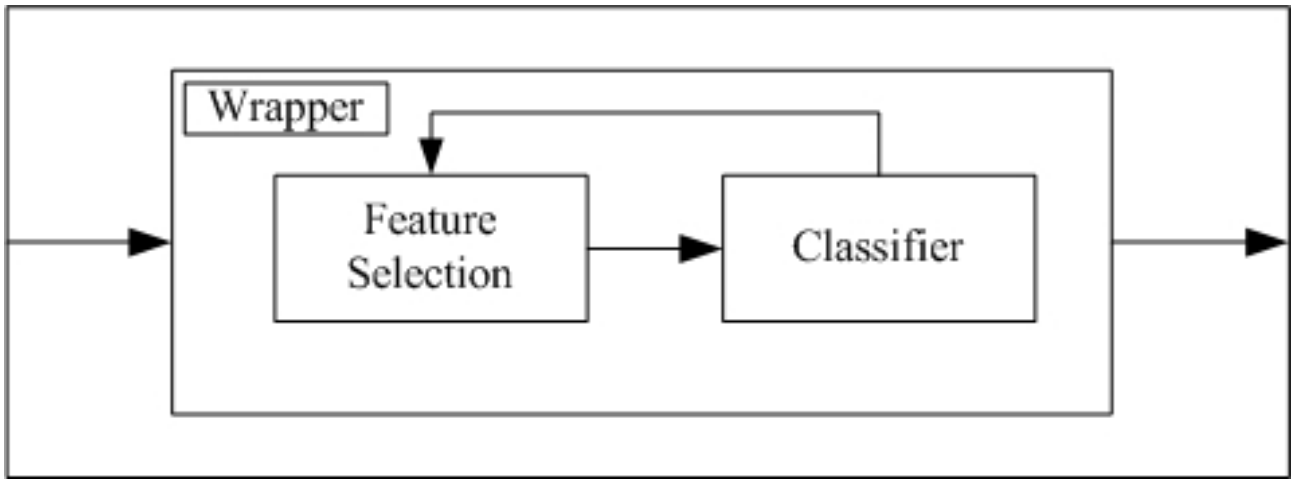
For feature selection in NVP, we used OPR with the LSTM classifier and a hold-out test (ie, training and testing) for performance evaluation. Note that we could not use SFS for feature selection since it is too time-consuming for LSTM. If we want to create equal-spaced features for each month between two visits for LSTM, we need to perform linear interpolation between these two visits for each subject. For lab test features (with continuous numerical values), this is achieved by spline interpolation with the piecewise cubic method. For questionnaire features (with categorical values of integers), this is achieved by linear interpolation with rounding off to the nearest labels, as shown in Figure 3.

Figure 3. Interpolation for the questionnaire features between any two medical checkups.

Feature Selection

As mentioned above, there are three categories of feature selection methods: wrappers, filters, and embedded methods [40]. In general, classification accuracy is strongly dependent on wrapper-selected features; however, this is a time-consuming approach. To strike a balance between efficiency and effectiveness, we compared two wrappers, OPR and SFS, for rapid feature selection based on our large dataset and a given classifier, as shown in Figure 4.

Figure 4. Conceptual diagram of wrappers that interact with a given classifier to select critical features.



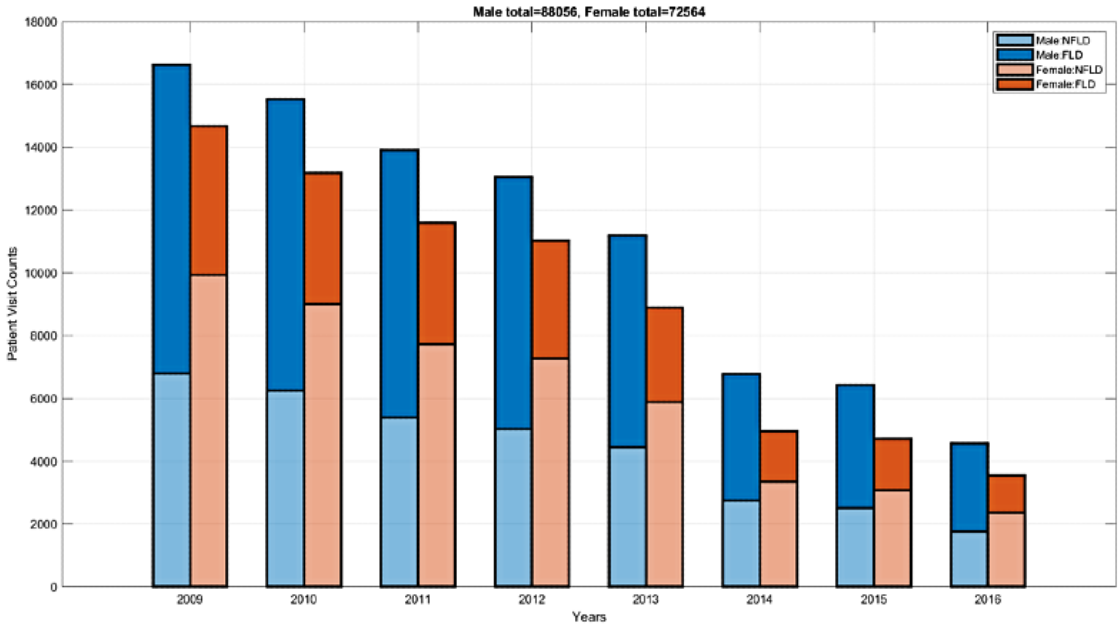
XSL•FO
RenderX

Dataset

General Characteristics of the Dataset

This study is primarily related to the MJ-FLD dataset [41], which was collected from a medical checkup clinic in Taipei from 2009 to 2016. This large dataset consists of 160,620 unique (people) visits (88,056 males and 72,546 females) with 446 features (also known as biodata) in total, including 289 from questionnaires and 157 from lab tests. Figure 5 shows the annual visit counts of males and females per year. Our goal is to predict whether a given person has FLD or not at the current and next visits. The following subsections explore the dataset in various ways. The sample sizes indicated refer to the total number of visits for all patients.

Figure 5. Visit counts for males (blue) and females (red) per year in the MJ-FLD dataset and statistics of no fatty liver disease (NFLD) and fatty liver disease (FLD) per year. The drop from 2013 to 2014 is likely due to the implementation of Taiwan's Personal Data Protection Act.



Data Size Over 8 Years

Figure 5 shows the annual visit counts of males and females per year of the dataset. The large disparity between 2013 and 2014 is likely due to enforcement of Taiwan's Personal Data Protection Act that set opt-in as the default for participation in medical research.

Therefore, between 2013 and 2014, the male count falls from 11,184 to 6770 (60.53% decrease), and the female count falls from 8896 to 4958 (55.73% decrease). Furthermore, over this 8-year period, the class size ratio of no fatty liver disease (NFLD) vs FLD was 0.66 (34,885 vs 53,171) for males and 2.02 (48,574 vs 23,990) for females. For each year from 2009 to 2016, the class size ratios of NFLD vs FLD were 0.69, 0.67, 0.63, 0.63, 0.66, 0.68, 0.64, and 0.63 for males, and 2.09, 2.16, 2.0, 1.93, 1.94, 2.1, 1.89, and 1.96 for females, respectively (Figure 5). These statistics indicate that the overall dataset is not highly imbalanced, and the class size ratios broken down by gender and year do not vary excessively.

Dataset Properties

Another characteristic of the dataset is its high ratio of missing values, as shown in Figure 6, which plots the percentage of missing values for all features and the top 20 features. Since the features with missing value ratios of 90% or higher are hard to impute, these 17 features were eliminated, leaving 252 features for further processing. The histograms of important features for males and females are shown in Figure 7. Some features such as waist-hip ratio displayed very different gender-dependent histograms.

XSL•FO
RenderX

Figure 6. The ratio of missing values for all features and for the top 20 features in the MJ-FLD dataset.

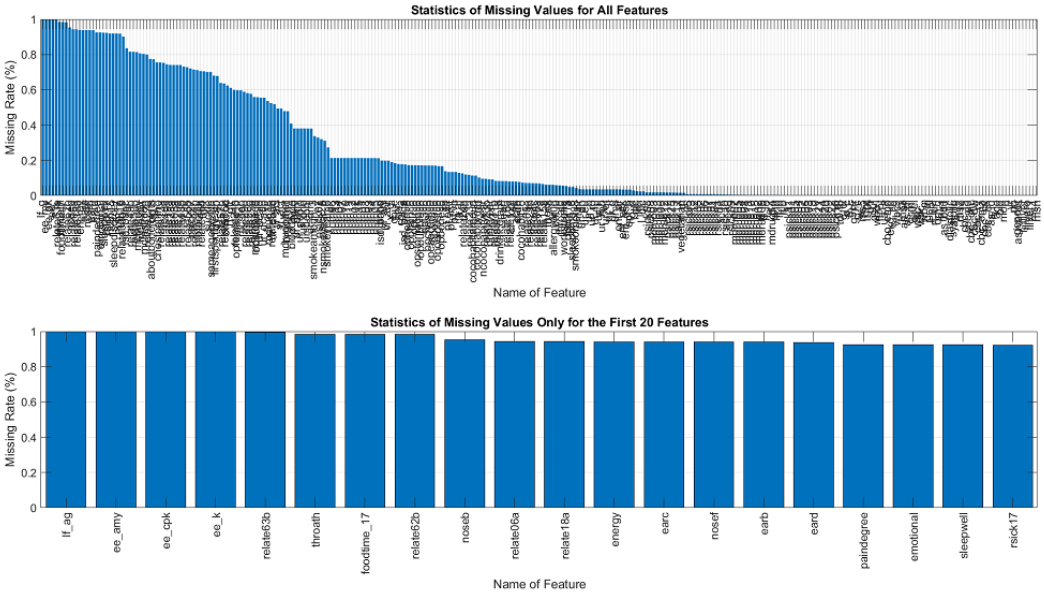
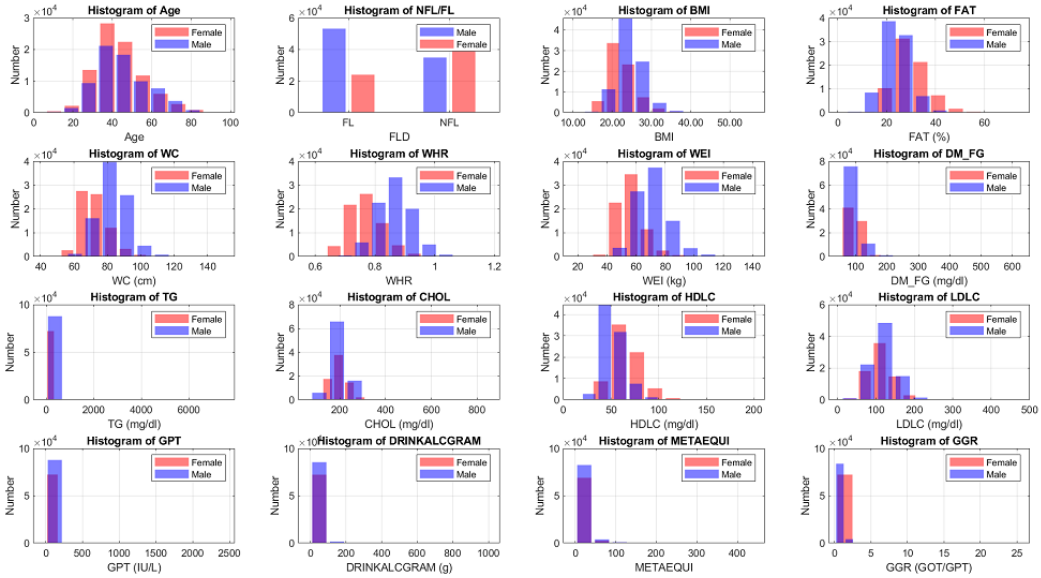


Figure 7. Histograms of important features of the MJ-FLD dataset for males (blue) and females (red). NFL: no fatty liver; FL: fatty liver; FAT: body fat; WC: waist circumference; WHR: waist-to-hip ratio; WEI: weight; DM_FG: diabetes for fasting glucose; TG: triglyceride; CHOL: total cholesterol; HDLC: high-density lipoprotein cholesterol; LDLC: low-density lipoprotein cholesterol; GPT: serum glutamic-pyruvic transaminase; DRINKALCGRAM: alcohol per gram; METAEQUI: metabolic equivalent for exercise per week; GGR: serum glutamic-oxaloacetic transaminase to glutamic-pyruvic transaminase ratio.

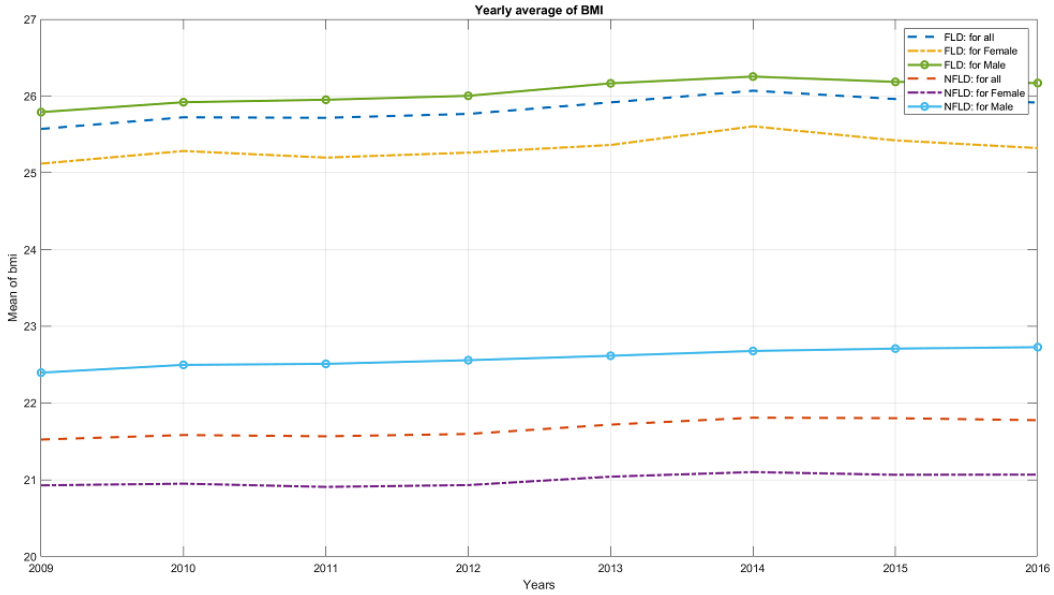


BMI Progression Over 8 Years

Some features such as BMI are strong indicators of FLD. Figure 8 plots the yearly average BMI for FLD and NFLD, broken down by males, females, and overall. Six curves are clearly divided into two groups of FLD and NFLD, with BMI for FLD consistently higher than that of NFLD. Within the same class (FLD or NFLD), males usually have a higher BMI than females. Moreover, the three curves for FLD show higher variance than the other three curves for NFLD, indicating that FLD patients might have a more dramatic BMI progression.

XSL•FO
RenderX

Figure 8. Progression of yearly average BMI over 8 years, broken down by [FLD, NFLD] x [male, female, overall] into 6 curves. FLD: fatty liver disease; NFLD: no fatty liver disease.



Data Preprocessing

Our dataset is based on health screening results from individuals, some of whom underwent multiple screenings at different intervals with different sets of screening items. As a result, there are several missing values in the dataset that needed to be imputed before further processing. Moreover, the questionnaires also changed over these 8 years when the dataset was compiled; therefore, we needed to consolidate the answers to different questionnaires of the same type.

To perform missing values imputation in our dataset, we used the mean for numerical features and the mode for questionnaire features. This is a quick and dirty method, especially for such a large dataset. Missing value imputation could be accomplished using other more complicated methods such as MICE (Multivariate Imputation by Chained Equations) [42], which imputes each missing value sequentially by another machine learning method. The process iterates until all of the imputed values converge, which usually takes a long time and is thus not feasible for a large dataset with many missing values.

To consolidate the answers to different questionnaires of the same type in the dataset, we needed to use some heuristics to derive consistent numerical values as features for machine learning. For instance, “grams of alcohol” represents the average weekly alcohol intake in grams [43,44], which was derived by combining some questionnaire items related to drinking from the MJ-FLD dataset. Similarly, to derive “weekly exercise metabolic equivalent,” we needed to combine some questionnaire items related to exercise.

In summary, the steps involved in data preprocessing were performed as follows:

1. Deletion of useless features: Our first step in data preprocessing was to drop features that are apparently not related to FLD, such as “cervical cancer,” “prostate cancer,” “other forms of cancer,” “other hereditary diseases,” “Chinese medicine,” and “has your mother or sister had breast cancer, ovarian cancer, or endometrial cancer?”
2. Missing value handling: Missing values in the dataset were replaced by the average for numerical features and by the mode for categorical features.
3. Feature conversion: To create consistent features from questionnaires, we consolidated highly related questionnaires and expressed the corresponding responses in numeric terms. For example, the feature “grams of alcohol consumption” was derived from responses to the questionnaire items “type of drink,” “amount of drink,” “drink or not,” and “alcohol density.” Similarly, the feature “weekly exercise metabolic equivalent” was derived from responses to the questionnaire items “type of sport,” “frequency of sport,” and “time for sport.”
4. Deletion of redundant features: Some highly redundant features were deleted from the dataset, such as “BMI,” “systolic/diastolic blood pressure while lying down left arm,” and “systolic/diastolic blood pressure while lying down right arm.”
5. Feature-wise normalization: This was achieved by z-score normalization to have a zero mean and unit variance for each feature:

where is the sample mean of feature x and S is the sample standard deviation of feature x.

Environment and Specification

All experiments were performed on a 64-bit Windows-10 server, with an Intel Xeon Silver 4116 CPU at 2.10 GHz, two NVIDIA Quadro GV100 GPUs, 256 GB RAM, 1-TB hard disk, and Matlab R2020b (9.8.0.1538559), and python 3.8.2, scikit-learn 0.24.1, TensorFlow-GPU 2.4.1.

XSL•FO
RenderX

All of the models in this study were constructed based on the MJ-FLD dataset [41]. Each of our experiments was designed with the goal of finding something meaningful in the dataset; therefore, we may use different ways to partition the dataset into subsets for training, validation, and testing for different experiments. We also performed necessary dataset preprocessing before using the data for modeling, including missing value imputation, feature consolidation, and feature-wise *Z*-score normalization, as explained above.

Results

Feature Selection With Various Methods

To investigate the effectiveness of different feature selection methods, we compared the computer-selected features with expert-suggested FLD features. All of the expert-suggested features are listed in Table 2, with a brief explanation for each. For instance, the well-known high-risk factors (or features) suggested by domain experts included BMI, body fat, and waist circumference. The critical factors related to AFLD are also listed, including “drinkalcgram” (average alcohol consumption in grams) and “drinkyear” (how many years the patient has been drinking alcohol).

XSL•FO
RenderX

Table 2. Features of fatty liver disease, including those suggested by domain experts or selected by one-pass ranking (OPR) and sequential forward selection (SFS) for current-visit prediction and next-visit prediction.

Features	Explanation	Suggested by experts	OPR		SFS		OPR (Feature set 1)		OPR (Feature set 2)	
			Selected by OPR	Match[a]	Selected by SFS	Match	Selected by OPR	Match	Selected by OPR	Match
age	Age				✓		✓		✓	
blood type	Blood type				✓					
bmd	Bone mineral density				✓		✓		✓	
bmi	Body mass index	✓	✓	✓	✓	✓	✓	✓	✓	✓
cc (cm)	Chest circumference		✓				✓			
cci (cm)	Chest circumference during inspiration		✓				✓		✓	
cea (ng/ml)	Carcinoembryonic antigen				✓					
ch	The ratio of chol/hdlc	✓	✓	✓			✓	✓	✓	✓
chol (mg/dl)	Total cholesterol	✓			✓	✓				
diastolic	Diastolic blood pressure						✓			
drinkalcgram (g)	Alcohol per gram	✓								
drinkyear	How many years have you been drinking?	✓								
e (%)	Eosinophils								✓	
ery (10^6/μl)	Red blood cells				✓					
fat (g)	Body fat	✓	✓	✓	✓	✓	✓	✓	✓	✓
fg (mg/dl)	Diabetes mellitus fasting glucose	✓	✓	✓	✓	✓	✓	✓	✓	✓
food18	How many servings of bread do you eat?	✓								
food19	Do you add jam or honey to your food?	✓								
food20	Do you add sugar to your coffee, tea, cola/soda, fruit juices, or other beverages?	✓								
food21	How many servings of your food intake are fried in oil?	✓			✓	✓				
ggr	The ratio of got/gpt	✓	✓	✓	✓	✓	✓	✓	✓	✓
ggt (IU/L)	Gamma-glutamyl transferase		✓		✓		✓		✓	
got (IU/L)	Serum glutamic-oxaloacetic transaminase (sGOT)		✓		✓		✓		✓	
gpt (IU/L)	Serum glutamic-pyruvic transaminase (sGPT)	✓	✓	✓			✓	✓	✓	✓
hc (cm)	Hip circumference		✓		✓		✓		✓	
hdlc (mg/dl)	High-density lipoprotein cholesterol	✓	✓	✓	✓	✓	✓	✓	✓	✓
hei (cm)	Height				✓				✓	
hema (%)	Hematocrit								✓	
Ldlc (mg/dl)	Low-density lipoprotein cholesterol	✓					✓	✓		
leu (10^3/ml)	White blood cells						✓		✓	
mcv (fl)	Mean corpuscular volume				✓					

XSL•FO
RenderX

Features	Explanation	Suggested by experts	OPR		SFS		OPR (Feature set 1)		OPR (Feature set 2)	
			Selected by OPR	Match[a]	Selected by SFS	Match	Selected by OPR	Match	Selected by OPR	Match
mdrug10	Steroids	✓			✓	✓				
mdrug8	Medicine for asthma	✓			✓	✓				
metaequi	Metabolic equivalent for exercise per week	✓							✓	✓
n (%)	Neutrophils								✓	
p (mg/dl)	Phosphorus		✓							
pul (beat/mint)	Pulse rate						✓		✓	
relate33b	In the last 3 months, have you lost weight by more than 4 kg?	✓								
relate17a	Have your defecation habits changed?		✓							
sdephi (/HPF)	Sediment epithelial cells high		✓							
sdrhi (/HPF)	Sediment red blood cells high		✓							
sdwhi (/HPF)	Sediment white blood cells high		✓							
sg	Specific gravity		✓							
smokeornot	Have you ever smoked?	✓								
systolic	Systolic blood pressure						✓			
tb (mg/dl)	Total bilirubin								✓	
tg (mg/dl)	Triglyceride	✓	✓	✓	✓	✓	✓	✓	✓	✓
tp (g/dl)	Total protein						✓			
tsh (μIU/ml)	Thyroid stimulating hormone		✓		✓					
ua (mg/dl)	Uric acid						✓		✓	
vanl	Visual acuity (naked left eye)				✓					
wc (cm)	Waist circumference	✓	✓	✓	✓	✓	✓	✓	✓	✓
wei (kg)	Weight	✓	✓	✓	✓	✓	✓	✓		
Whr	Waist-to-hip ratio	✓	✓	✓			✓	✓		
Workstreng	What is your level of activity at work?		✓							

[a]Indicates a match with the features selected by domain experts based on the literature.

Intersection Over Union and Coverage

To evaluate the similarity between the feature sets manually selected by human experts (set S1) and automatically selected by OPR/SFS (set S2), we used two similarity indices, intersection over union (IoU) and coverage, defined as follows:

$$\mathrm{IoU}(S1, S2)=|S1\cap S2|/|S1\cup S2|$$

$$\mathrm{Coverage}(S1, S2)=|S1\cap S2|/|S1|$$

Both similarity indices range from 0 to 1, and a higher value indicates higher similarity.

Experiment 1: CVP With Optimum Years of Training Data and Feature Selection

Given the size of the dataset, we can explore it in different directions. First, we needed to confirm the modeling accuracy of CVP across years, which was achieved using the previous year data for training and the current year data for testing. The test accuracy for each year is shown in Figure 9.

Next, we wanted to further explore the optimum duration in years considered for modeling in feature selection. In general, using a long period of historical data for modeling may result in mismatching with the test data since the optimum model may change over time. However, a short period of historical data may not be sufficient for stable model construction. As a result,

XSL•FO
RenderX

we needed to identify the optimum duration in years where the training data are obtained for predicting the data in 2016. More specifically, we defined seven subtasks for training data in intervals (2015, 2014-2015, 2013-2015, 2012-2015, 2011-2015, 2010-2015, 2009-2015), and the test data were from 2016. This arrangement is illustrated in Figure 10. Moreover, we performed feature selection for each subtask to select the best features. The modeling specifications are as follows: dataset, male part of the MJ-FLD dataset; classifier, KNNC; feature selection, OPR with LOO cross-validation for the performance index to select the most important 24 features (this number was used to match the number of features suggested by the domain experts.)

The result is shown in Figure 11, where the best interval was 2012-2015, achieving the best test accuracy of 80.00%. The corresponding OPR-selected features are shown in Figure 12. For comparison, if we used the same training/test pair to evaluate SFS-selected and expert-suggested 24 features, the accuracies were 78.37% and 79.78%, respectively. Using the same evaluation steps on female data produced the same result; that is, the best interval was 2012-2015.

Figure 9. Test accuracy for each year using the previous year data for training and the current year data for testing for both males and females.

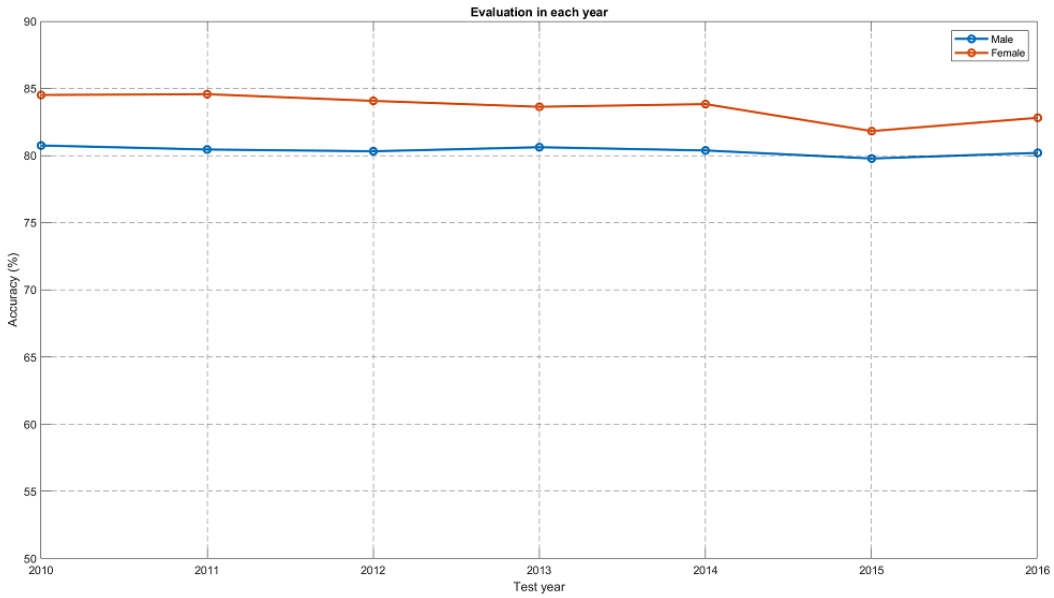
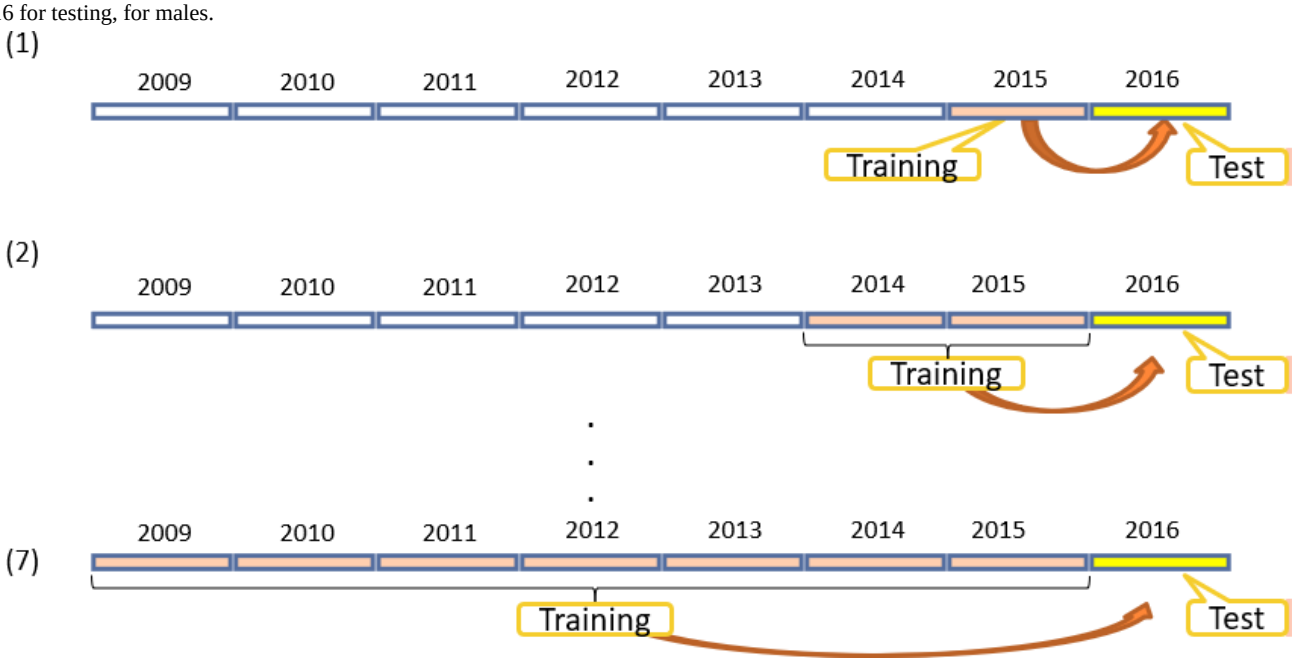


Figure 10. The models of seven subtasks for training in intervals (2015, 2014-2015, 2013-2015, 2012-2015, 2011-2015, 2010-2015, 2009-2015) and 2016 for testing, for males.



XSL•FO
RenderX

Figure 11. The best year interval for the model of male fatty liver disease prediction is 2012-2015.

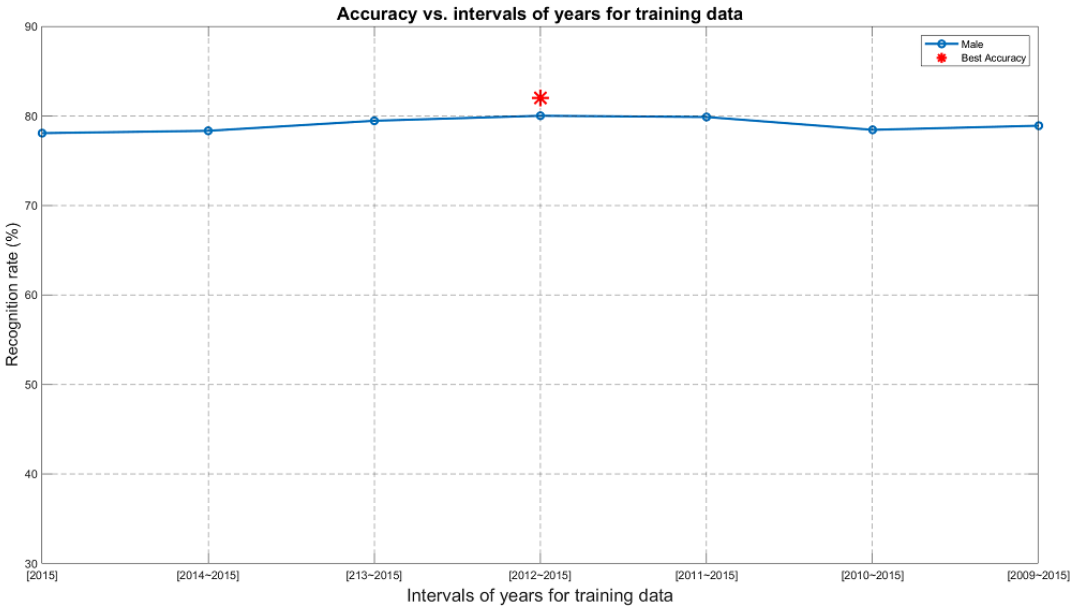
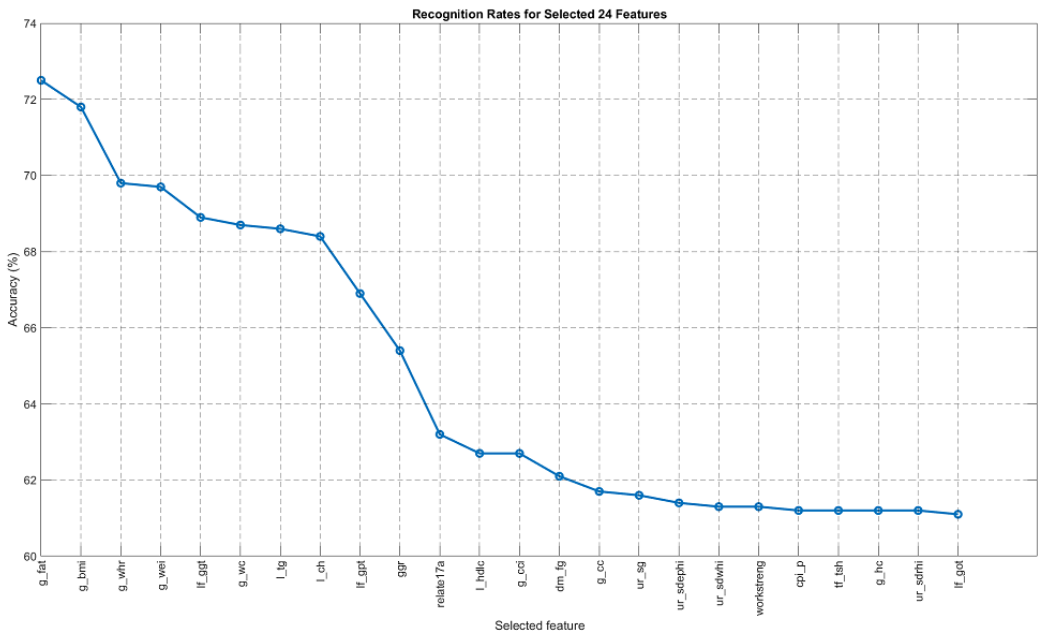


Figure 12. Features selected by one-pass ranking based on the standard set, in descending order of recognition rate.



For easy reference, we refer to the training set of the interval 2012-2015 and the test set from 2016 as the "standard set." Based on the standard set, we applied OPR and SFS, as shown in Table 3. The result indicated that SFS is slightly better than OPR in terms of classification accuracy (80.92% vs 80.32%). In terms of the selected features, SFS was also slightly better than OPR, with 50.00% vs 45.83% for coverage rate and 33.33% vs 29.73% for IoU. However, SFS achieved these marginal improvements at the cost of computing time, which was approximately three times slower than that of OPR. The features selected by OPR, SFS, and domain experts are listed in Table 2, including the most common features for FLD with a simple explanation. In the table, any matched features selected by OPR or SFS are indicated with a check mark in the "Match" column.

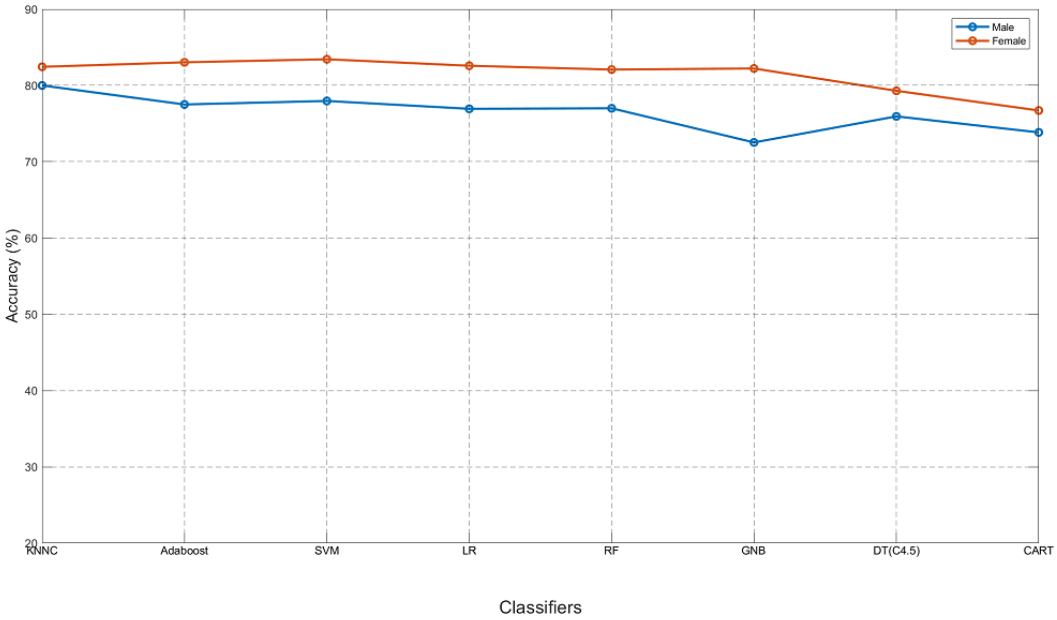
Finally, we tested other classifiers on the standard set, including KNNC, Adaboost, SVM, LR, RF, GNB, decision trees C4.5, and CART, as shown in Figure 13. The classifiers of Adaboost and SVM showed higher accuracy than the others. We also noticed that for all classifiers, the accuracy for the females outperformed that for the males, which will be discussed in the next subsection. The area under the receiver operating characteristic curve (AUROC), precision, recall, and F1 scores for these 7 classifiers are shown in Table 4. In particular, the AUROC values for these classifiers for CVP were all higher for females than for males.

XSL•FO
RenderX

Table 3. Comparison of one-pass ranking (OPR) and sequential forward selection (SFS) in terms of feature selection and classification.

Metric	OPR	SFS
Feature selection		
Intersection over union	29.73% (11/37)	33.33% (12/36)
Coverage	45.83% (11/24)	50.00% (12/24)
Classification accuracy	80.32%	80.92%

Figure 13. Performance of various classifiers on the standard set. KNNC: k-nearest neighbor classifier; SVM: support vector machine; LR: logistic regression; RF: random forest; GNB: Gaussian naive Bayes; CART: classification and regression trees.



XSL•FO
RenderX

Table 4. Performance metrics for eight different classifiers.

Classifier	AUROC[a]	Precision	Recall	F1 score	Accuracy
KNNC[b]					
Males	0.80	0.77	0.82	0.79	80.00%
Females	0.87	0.77	0.68	0.72	82.45%
Adaboost					
Males	0.85	0.80	0.85	0.82	77.51%
Females	0.90	0.77	0.70	0.74	83.07%
SVM[c]					
Males	0.85	0.80	0.86	0.83	77.97%
Females	0.90	0.80	0.69	0.74	83.44%
LR[d]					
Males	0.85	0.83	0.78	0.81	76.94%
Females	0.90	0.71	0.82	0.76	82.59%
RF[e]					
Males	0.85	0.83	0.79	0.81	77.01%
Females	0.90	0.72	0.80	0.76	82.90%
GNB[f]					
Males	0.79	0.83	0.70	0.76	72.53%
Females	0.88	0.77	0.67	0.72	82.23%
DT[g] (C4.5)					
Males	0.83	0.83	0.76	0.80	75.95%
Females	0.87	0.67	0.78	0.72	79.30%
CART[h]					
Males	0.73	0.79	0.78	0.79	73.85%
Females	0.76	0.63	0.75	0.68	76.72%

[a]AUROC: area under the receiver operating characteristic curve.
[b]KNNC: k-nearest-neighbor classifier.
[c]SVM: support vector machine.
[d]LR: logistic regression.
[e]RF: random forest.
[f]GNB: Gaussian naïve Bayes.
[g]DT: decision tree.
[h]CART: classification and regression trees.

Experiment 2: Hormonal Influence in CVP

As shown in Figure 13, the accuracy for females was consistently higher than that for males. This may be due to data imbalance, which is further addressed in the Discussion section. Moreover, we can also explore the influence of hormones for both males and females in CVP. To this end, we assumed that menopause/andropause occurs at a certain age and then performed modeling/evaluation before and after the age to determine the difference in prediction accuracy. More specifically, we split the whole dataset (2009-2016) into two subsets, “before” and “after,” according to the assumed age of menopause. Within each subset, the period of 2009-2015 was used for training and 2016 was used for testing with the naïve Bayes classifier. The results are shown in Figure 14, in which we assumed that menopause/andropause occurs at ages 53, 54, 55, 56, and 57, and derived the accuracy before and after menopause/andropause for both males and females. We observed that the “before” accuracy is consistently higher than that of “after” for females. Moreover, the accuracy differences between “before” and “after” were much higher for females than for males. This is because female hormones can maintain the basal metabolic rate at a certain level before menopause such that the accumulation of fat in the internal organs is less likely to occur, thus improving the FLD prediction accuracy. After menopause, women do not have normal hormone

XSL•FO
RenderX

secretion, leading to a less balanced body status and more challenging FLD prediction. For fatty liver, lifestyle intervention is usually recommended for treatment. Chalasani et al [45] reviewed several population-based studies and pointed out that because body fat, sex hormone metabolism, and lifestyle have gender differences, the occurrence of FLD will vary by gender [46]. Therefore, we believe that the accuracy of CVP will also differ due to these indicators.

Figure 14. Investigation of hormonal influence, assuming menopause/andropause occurs at ages 53, 54, 55, 56, and 57, respectively. The upper plot is for males and the lower plot is for females. Each yellow-purple bar pair indicates the accuracy before and after menopause at a specific age. The dataset used for this analysis corresponds to the years 2009-2016.

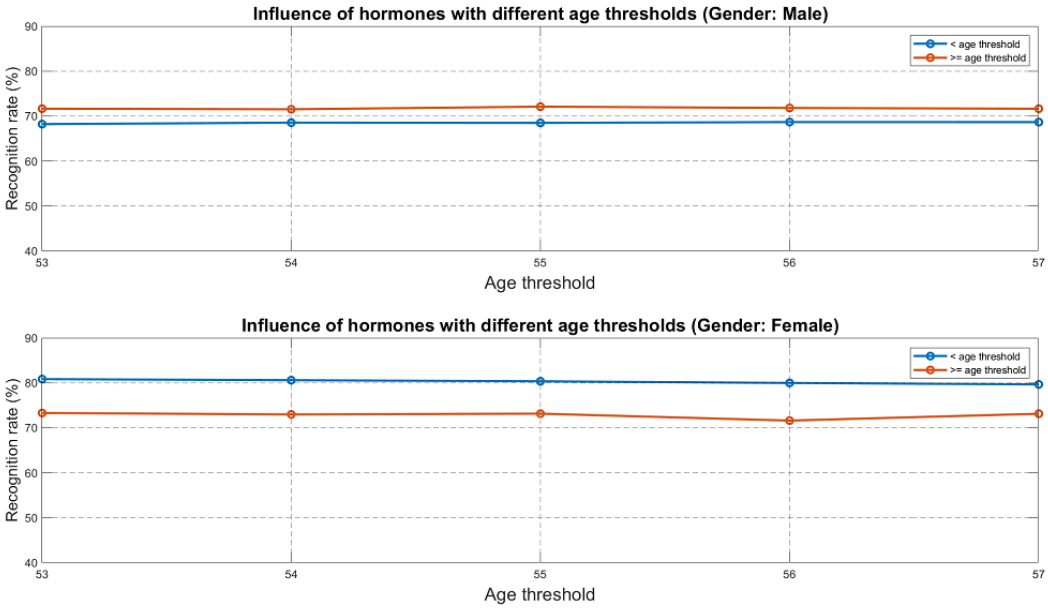


Figure 14 shows that the difference in recognition rate for males does not change obviously between the "before" and "after" age threshold, but it does for females within each subset. This means that sex hormones play an important role in FLD prediction for females. In other words, the greater the effect of sex hormones will result in a higher recognition rate for prediction.

For females, sex hormones will be affected not only by the lifestyle habits an individual engages in to maintain a good figure but also by factors such as dieting and drugs. To achieve a slim figure, many women try various types of diets that have several side effects, which may affect specific biochemical tests related to FLD. In addition, some women may resort to the ingestion of nutritional supplements or other forms of "diet pills" to lose weight. However, many of these drugs contain unknown ingredients or illegal substances that could significantly affect the results of tests associated with FLD.

Experiment 3: LSTM for NVP

In this experiment, we used LSTM with various setups for NVP. LSTM is a well-known sequence classifier that can use information from historical visits, with no length limit, to predict the possibility of FLD at the patient's next clinic visit. As explained earlier, from the perspective of preventive medicine, NVP is much more important than CVP. The specifications for feature selection of NVP are as follows: dataset, male subjects in the MJ-FLD dataset; classifier, LSTM; feature selection, OPR with 3-fold cross-validation to select the most important 24 features.

In general, clinic visits do not always occur at regular intervals. For a given visit pattern of length N, we can extract N – 1 input-output pairs for NVP modeling using LSTM, as shown in Figure 15 where N=5. To deal with this situation of nonregular intervals, we designed two types of LSTM that have two types of feature sets. In feature set 1 with fixed intervals, interpolation was performed to obtain a fixed-interval input sequence to our sequence classifier. For instance, the input can have a fixed interval of 1 month and the output can be 12 months into the future, as shown in Figure 16. If the next visit is less than or equal to 12 months away from the current visit, then we can easily perform interpolation for the input. However, if the next visit is more than 12 months away from the current visit, then we simply duplicate the data at the current visit to the subsequent months until we have enough data to perform NVP. In feature set 2 with variable intervals, we used the visit pattern directly with extra inputs to preserve the interval information and target time for prediction. For instance, if we have d features for a visit, then the number of inputs should be d+2, with the additional first feature indicating the time span from the previous visit and the additional second feature indicating how far in the future the prediction should be made, as shown in Figure 17.

For feature set 1 with fixed-interval data, the dataset included the number of input/output pairs for males (13,315) and for females (10,998). The mean input sequence length for males and females was 42.03 (SD 21.25, range 5-96) and 41.44 (SD 20.84, range 4-96), respectively. For feature set 2 with variable-interval data, there were 16,081 input/output pairs for males with a mean input sequence length of 3.32 (SD 1.46,

XSL•FO
RenderX

range 2-13), and 13,364 input/output pairs for females with a mean input sequence length of 3.15 (SD 1.35, range 2-15).

Feature set 2 with input data from variable intervals showed three major advantages: (1) the unfolded LSTM network has considerably fewer stages, resulting in much shorter training and prediction times; (2) the dataset is used directly with no need to perform extra interpolation in advance, thus reducing time requirements and increasing precision; and (3) it can perform any prediction at any time in the future directly.

Figure 15. A typical visit pattern and the extracted input/output pairs for training long short-term memory (LSTM). If the visit pattern is denoted by [v1, v2, v3, v4, v5], then we can extract 4 input/output pairs for training LSTM: {v1⇒v2}, {v1,v2⇒v3}, {v1, v2, v3⇒v4}, {v1,v2,v3,v4⇒v5}. Note that patients with only a single visit are discarded in this next-visit prediction task.

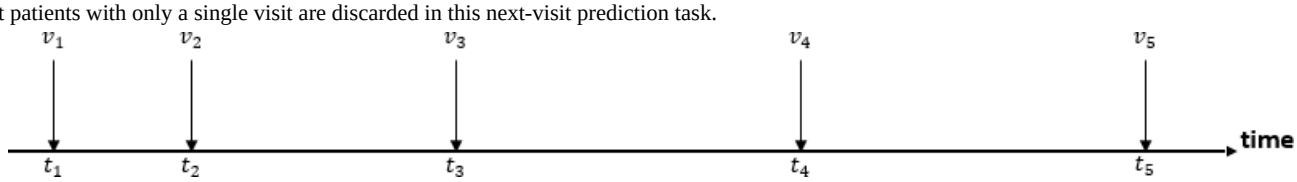


Figure 16. To create fixed-interval data for feature set 1, we need to perform interpolation on the input/output parts. For this case, the input part is interpolated to have a fixed interval of 1 month and the output part is interpolated to have a time distance of 12 months from the nearest time of the input.

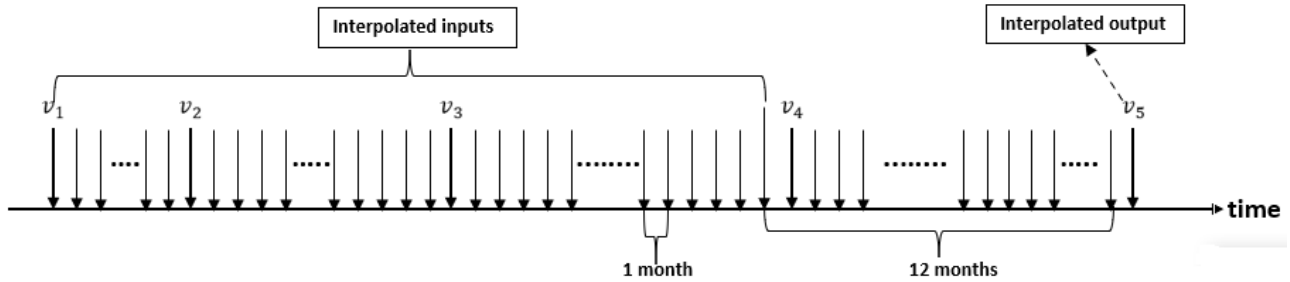
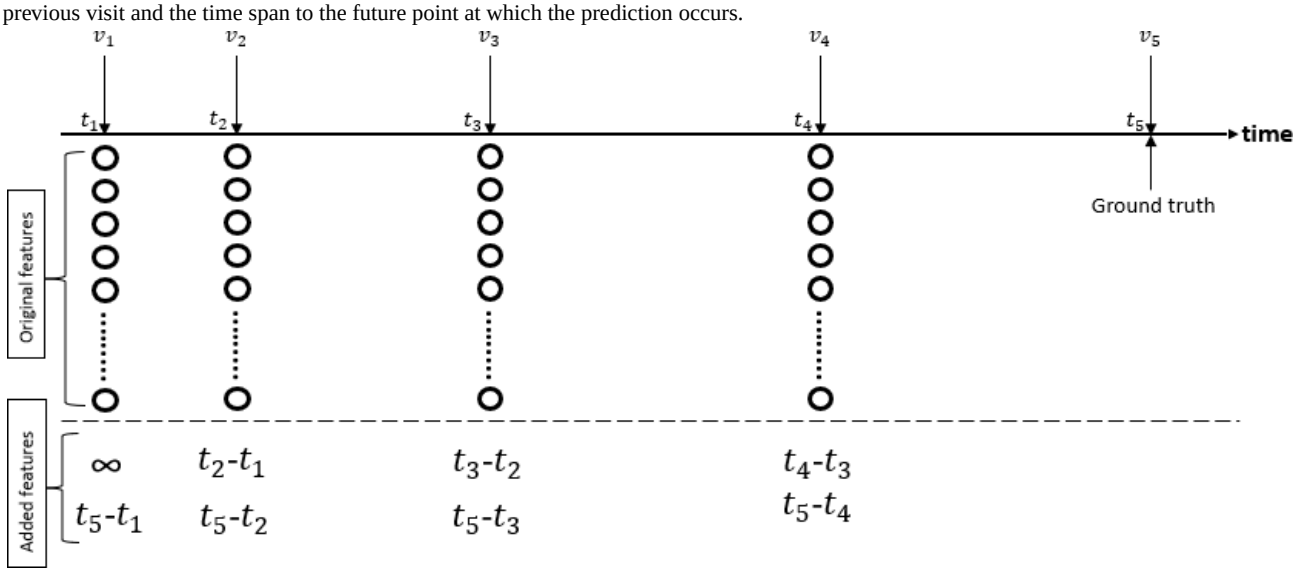


Figure 17. To create variable-interval data for feature set 2, we need to add two extra inputs to long short-term memory, including the time span from the previous visit and the time span to the future point at which the prediction occurs.



First, the input/output pairs used to train feature set 1 (with 24 features for males in the dataset) were prepared as follows. All patients with only a single visit were removed from the dataset, reducing the total number of males from 34,856 to 22,972. From the historical data for each patient, we interpolated data between any two consecutive visits to the monthly values. For a specific visit (excluding the last one), the first 12 months of the interpolated data right before the visit were used as the feature set 1 input, while the interpolated output at 12 months right after the visit was used as the output. The input-output data pairs were then collected using moving windows with a stride of 1 month.

The final count of input-output data pairs for trained feature set 1 with 24 features was 469,159. These data pairs were divided into 70% used for training (10% of which was used for validation) and 30% used for testing, all with stratified partitioning. All training options and parameters for LSTM are listed in Multimedia Appendix 1. Figure 18 shows the training and validation accuracy/loss vs epochs during the training process. As usual, the best model was selected at the epoch where the validation loss reached its minimum or the validation accuracy reached its maximum. In this case, the best model was selected at epoch 93 where the validation accuracy reached its maximum of 81.72%.

XSL•FO
RenderX

Based on the above process, we then performed OPR on top of feature set 1 to derive 24 features. As shown in Figure 19, when compared with the expert-selected features, the OPR-selected features achieved an IoU of 29.73% and a coverage of 45.83%, which is satisfactory based on the opinions of the domain experts we consulted. By contrast, the OPR on top of feature set 2 achieved an IoU of 23.08% and the coverage was 37.50%. All results for feature sets 1 and 2 are shown in Table 5. The AUROC, precision, recall, and F1 scores are shown in Table 6.

Figure 18. The accuracy (upper plot) and loss (lower plot) for training and validation during the training of feature set 1 for male subjects of the MJ-FLD dataset. The best model was selected at epoch 93 where the validation accuracy reached its maximum of 81.72%.

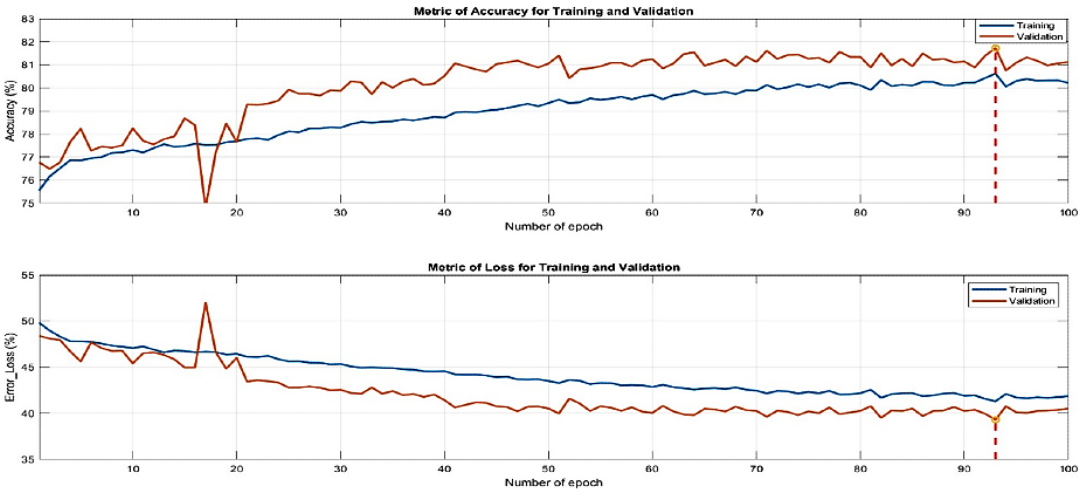


Figure 19. Features selected by one-pass ranking based on feature set 1, ranked by accuracy.

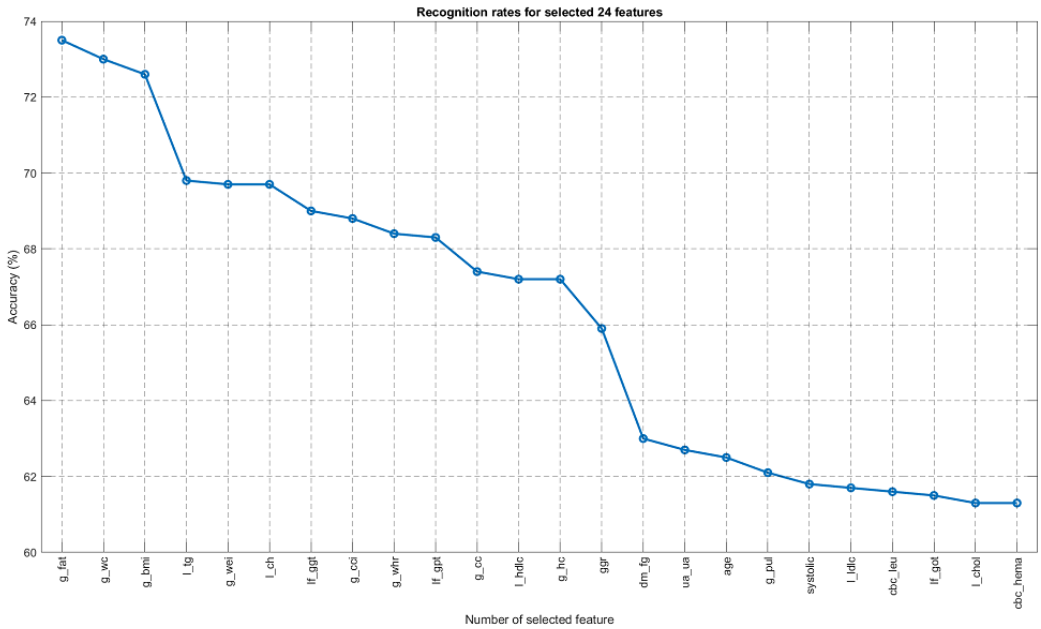


Table 5. Comparison of intersection over union (IoU), coverage, and accuracy of the features selected by one-pass ranking (OPR) and domain experts in the two feature sets.

Metric	OPR		Experts	
	Feature set 1	Feature set 2	Feature set 1	Feature set 2
IoU	29.73% (11/37)	23.08% (9/39)	N/A[a]	N/A
Coverage	45.83% (11/24)	37.50% (9/24)	N/A	N/A
Accuracy	75.91%	77.32%	75.40%	74.95%
Computing time (seconds)	5875	1452	N/A	N/A

[a]N/A: not applicable.

XSL•FO
RenderX

Table 6. Comparison of performance, computing time, and error reduction rate with five long short-term memory (LSTM)-based classifiers.

Classifier	AUROC[a]	Precision	Recall	F1 score	Accuracy	Computing time (s)	Error reduction rate
LSTM							
FSI[b]							
Males	0.83	0.75	0.74	0.75	76.54%	2713	5.33%
Females	0.88	0.80	0.78	0.79	81.90%	2485	30.86%
FS2[c]							
Males	0.86	0.78	0.77	0.78	79.29%	1466	16.42%
Females	0.87	0.79	0.77	0.78	80.81%	1469	26.70%
biLSTM[d]							
FS1							
Males	0.83	0.75	0.74	0.75	76.66%	3380	5.81%
Females	0.88	0.81	0.78	0.79	81.70%	3155	30.10%
FS2							
Males	0.87	0.78	0.77	0.78	79.12%	1789	15.74%
Females	0.88	0.79	0.77	0.78	80.79%	1800	26.63%
Stack-LSTM							
FS1							
Males	0.84	0.76	0.75	0.75	77.23%	3764	8.11%
Females	0.88	0.80	0.78	0.79	81.87%	3524	30.75%
FS2							
Males	0.87	0.78	0.77	0.78	79.32%	1952	16.55%
Females	0.87	0.79	0.77	0.78	80.51%	2016	25.55%
Stack-biLSTM							
FS1							
Males	0.84	0.76	0.75	0.75	76.84%	6085	6.54%
Females	0.88	0.80	0.78	0.79	81.77%	5429	30.37%
FS2							
Males	0.87	0.78	0.77	0.78	79.29%	2714	16.42%
Females	0.88	0.79	0.77	0.78	80.78%	2802	26.59%
Attention-LSTM							
FS1							
Males	0.84	0.83	0.80	0.81	77.31%	N/A[e]	8.43%
Females	0.89	0.69	0.79	0.74	80.81%	N/A	26.70%
FS2							
Males	0.87	0.87	0.77	0.82	78.36%	N/A	12.67%
Females	0.89	0.70	0.81	0.75	81.46%	N/A	29.18%

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

[b]FS1: feature set 1.

[c]FS2: feature set 2.

[d]biLSTM: bidirectional long short-term memory.

[e]N/A: not applicable.

We next compared the performances of feature sets 1 and 2 to two baseline models, as shown in Figure 20. The predictor for baseline 1 always outputs the class with a larger percentage in the ground truth. In the case of the MJ-FLD dataset, the output

XSL•FO
RenderX

is always NFLD. Baseline 2 is a simple inference model that always outputs the class of the previous visit. In other words, the prediction is based on the ground truth of the previous visit.

The test accuracy of NVP using feature set 1 (with fixed intervals) and feature set 2 (with variable intervals) for males was 77.31% with Attention-LSTM (8.43% error reduction) and 79.32% with Stack-LSTM (16.55% error reduction), respectively. The error reduction rates were compared with a baseline model of simple inference. For females, the corresponding values were 81.90% with LSTM (30.86% error reduction) and 81.46% with Attention-LSTM (29.18% error reduction). The error reduction rates of four classifiers for males and females are listed in Table 6.

Figure 20. Accuracy for two baseline models and 10 long short-term memory (LSTM) models for males and females. biLSTM: bidirectional LSTM.

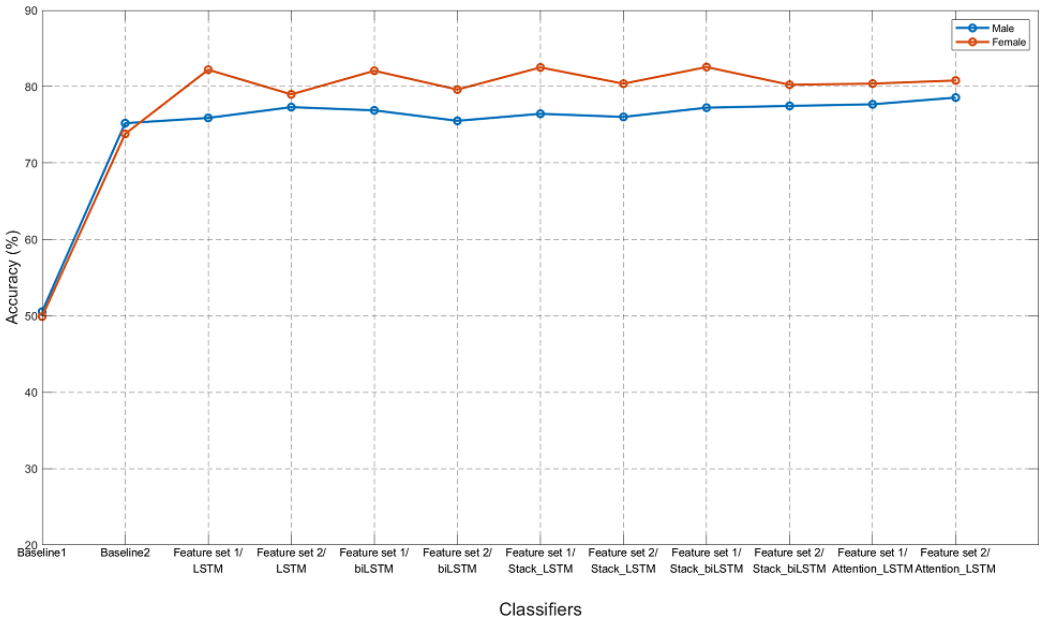


Table 5 shows the IoU and coverage rates of OPR-selected features based on feature sets 1 and 2. The accuracy of feature set 2 was comparable with that of feature set 1 for both males and females. However, the training times were 5875 and 1452 seconds, respectively, indicating that the proposed feature set 2 provides much better efficiency. Note that it is almost impossible to perform SFS in this case due to its lengthy computation. Moreover, for both feature sets 1 and 2, the accuracy results of OPR-selected features (78.20% and 76.79%) were higher than those of the expert-selected features (75.40% and 74.95%), indicating the feasibility of OPR for feature selection of a large dataset with a complex model of LSTM.

For feature set 2, we discarded patients with a single visit to obtain 76,172 input-output pairs; therefore, the number of male patient visits dropped from 34,856 to 22,972. The results of OPR-selected features are listed in Table 2 for comparison. Note that the table does not include feature set 2–based SFS, simply because the computational time for SFS with feature set 2 takes more than 7 days.

Discussion

Principal Findings

The computing time of OPR was much lower than that of SFS; however, it can achieve comparable performance (in terms of the overlap between the automatically selected features and the manually selected features) as SFS, especially when dealing with a large-scale dataset with high-dimensional features. The best model for CVP was KNNC for males (80.00%) and SVM for females (83.44%). The best model for NVP was Stack-LSTM using feature set 1 (79.32%) for males and LSTM using feature set 2 (81.90%) for females.

For NVP, the proposed feature set 2 is highly flexible and can achieve comparable results to those obtained with feature set 1; however, the computing time is much shorter, and the prediction can be derived at any time in the future. Both feature sets 1 and 2 outperformed a simple inference model (baseline 2), achieving an error reduction of 16.53% (Stack-LSTM) for males and 30.86% (LSTM) for females.

As shown in Table 4, by comparing two rows of SVM/male and KNNC/male, we can observe that SVM outperformed KNNC in all metrics except for accuracy. As a result, for males, SVM can be used to replace KNNC if accuracy is not the only concern. According to Figure 9 and Figure 13, the CVP for females was consistently better than that for males. This is simply due to the fact that the female dataset is more imbalanced than the male dataset. To demonstrate this, we computed the imbalance factors (data size of the bigger class divided by that of the smaller class) across 8 years: (1.45, 1.49, 1.58, 1.60, 1.52, 1.47, 1.57, 1.60) for males and (2.09, 2.16, 2.0, 1.93, 1.94, 2.1, 1.89, 1.96) for females. Therefore, the imbalance factors for females are consistently higher than those for males, leading to better accuracy for the female dataset.

For CVP, the influence of hormones for females was more intense than that for males, leading to difficulty in FLD prediction for females after menopause, as shown in Figure 14, where the difference in accuracy before and after menopause

XSL•FO
RenderX

age is more dramatic for females than for males. In other words, hormones play an important role for FLD prediction in females. However, after menopause, women lose protection from sex hormones, which can increase the risk of chronic and/or metabolic diseases. This would make FLD prediction harder due to women's imbalanced postmenopausal physiology.

For males in Figure 14, the accuracy of the "bigger-age group" is higher than that of the "smaller-age group." This difference is not related to hormones since men do not exhibit obvious menopause. It is more likely due to the data imbalance, as demonstrated by the imbalance factors of the "smaller-age group" at (1.54, 1.56, 1.57, 1.58, 1.59) and "bigger-age group" at (1.78, 1.70, 1.71, 1.67, 1.64). Note that a higher imbalance factor usually leads to higher accuracy.

In Table 6 for NVP, the best classifiers are Stack-LSTM (using feature set 2) for males and LSTM (using feature set 1) for females. This indicates that there is no single model and no single feature set that are best for both males and females.

It should be noted that by using Attention-LSTM with feature set 2, the accuracy only dropped by 0.96% for female FLD prediction and by 0.44% for male FLD prediction. The advantages in using feature set 2 include better efficiency in training/evaluation and more flexible prediction at any future time. Thus, if efficiency and flexibility are major concerns, we can sacrifice accuracy to a certain degree to achieve high efficiency and flexibility.

Conclusions and Future Work

This study explored the use of a large health checkup dataset for FLD prediction in terms of current-visit and next-visit predictions. We used OPR and SFS for feature selection in CVP and then compared the results against expert-selected features. In our experiment with CVP, OPR was more efficient and provided comparable results with those obtained using SFS in terms of classification accuracy and the similarity between the automatically selected features and the expert-selected features.

For NVP, we propose two feature sets (feature sets 1 and 2) for various LSTM models. For females, the best accuracy of 81.90% was obtained when using feature set 1 for LSTM. For males, the best accuracy of 79.32% was obtained when using feature set 2 for LSTM. This indicates that the best models and best features are gender-dependent. However, it should be noted that feature set 2 is a much more compact representation; thus, it requires less time for training/evaluation, and there is no need for prior feature interpolation. Moreover, the model trained by feature set 2 is more flexible and it allows for FLD prediction at any time in the future.

In practice, NVP is much more valuable from the perspective of preventive medicine since whenever a positive prediction occurs, the physician can suggest lifestyle changes to prevent FLD at the next visit. To our knowledge, this is the first use of machine learning for NVP using a large-scale dataset.

Our immediate future work will focus on extending our LSTM-based NVP system to develop a comprehensive recommendation system, in which precise and personal recommendations will be given to prevent the potential future development of FLD, such as reduction in alcohol consumption, weight loss, and increased exercise. Such precise, personalized recommendations can be made based on patient clustering according to influential features. In general, such a system for preventive treatment can also be extended to other chronic or metabolic syndrome diseases, as long as we have a large dataset that covers many years for longitudinal studies.

Acknowledgments

All data used in this study were authorized by and received from MJ Health Research Foundation (authorization code MJHRF2019014C). Any interpretations or conclusions described in this paper are those of the authors and do not represent the views of MJ Health Research Foundation. The work presented herein was partly supported by the Ministry of Science and Technology, Taiwan (grant MOST 110-2634- F-002-032).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Parameters of training options for the several variants of long short-term memory (LSTM).
[DOCX File , 18 KB - medinform_v9i8e26398_app1.docx]

References

1. Rajabi Shishvan O, Zois D, Soyata T. Machine intelligence in healthcare and medical cyber physical systems: a survey. IEEE Access 2018;6:46419-46494. [doi: 10.1109/access.2018.2866049]
2. Fan J, Kim S, Wong VW. New trends on obesity and NAFLD in Asia. J Hepatol 2017 Oct;67(4):862-873. [doi: 10.1016/j.jhep.2017.06.003] [Medline: 28642059]
3. Hsu C, Kao J. Non-alcoholic fatty liver disease: an emerging liver disease in Taiwan. J Formos Med Assoc 2012 Oct;111(10):527-535 [FREE Full text] [doi: 10.1016/j.jfma.2012.07.002] [Medline: 23089687]

XSL•FO
RenderX

4. Birjandi M, Ayatollahi SMT, Pourahmad S, Safarpour AR. Prediction and diagnosis of non-alcoholic fatty liver disease (NAFLD) and identification of its associated factors using the classification tree method. Iran Red Crescent Med J 2016 Nov 09;18(11):e32858 [FREE Full text] [doi: 10.5812/ircmj.32858] [Medline: 28191344]
5. Jamali R, Arj A, Razavizade M, Aarabi M. Prediction of nonalcoholic fatty liver disease via a novel panel of serum adipokines. Medicine (Baltimore) 2016 Feb;95(5):e2630. [doi: 10.1097/MD.0000000000002630] [Medline: 26844476]
6. Ma H, Xu C, Shen Z, Yu C, Li Y. Application of machine learning techniques for clinical predictive modeling: a cross-sectional study on nonalcoholic fatty liver disease in China. Biomed Res Int 2018 Oct 03;2018:4304376. [doi: 10.1155/2018/4304376] [Medline: 30402478]
7. Wu C, Yeh W, Hsu W, Islam MM, Nguyen PA, Poly TN, et al. Prediction of fatty liver disease using machine learning algorithms. Comput Methods Programs Biomed 2019 Mar;170:23-29. [doi: 10.1016/j.cmpb.2018.12.032] [Medline: 30712601]
8. Yip TC, Ma AJ, Wong VW, Tse Y, Chan HL, Yuen P, 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 06;46(4):447-456. [doi: 10.1111/apt.14172] [Medline: 28585725]
9. Gu S, Cheng R, Jin Y. Feature selection for high-dimensional classification using a competitive swarm optimizer. Soft Comput 2016 Oct 7;22(3):811-822. [doi: 10.1007/s00500-016-2385-6]
10. Juha R. Overfitting in making comparisons between variable selection methods. J Machine Learn Res 2003;3:1371-1382 [FREE Full text]
11. Islam MM, Wu CC, Poly TN, Yang HC, Li YCJ. Applications of machine learning in fatty liver disease prediction. In: Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth. 2006 Apr Presented at: 40th Medical Informatics in Europe Conference, MIE 2018; April 26, 2018; Gothenburg, Sweden p. 166-170 URL: https://tmu.pure.elsevier.com/en/publications/applications-of-machine-learning-in-fatty-live-disease-prediction [doi: 10.3233/978-1-61499-852-5-166]
12. Paul D, Su R, Romain M, Sébastien V, Pierre V, Isabelle G. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. Comput Med Imaging Graph 2017 Sep;60:42-49. [doi: 10.1016/j.compmedimag.2016.12.002] [Medline: 28087102]
13. Tian Y, Zhang X, Wang C, Jin Y. An evolutionary algorithm for large-scale sparse multiobjective optimization problems. IEEE Trans Evol Computat 2020 Apr;24(2):380-393. [doi: 10.1109/tevc.2019.2918140]
14. Wei X, Jiang F, Wei F, Zhang J, Liao W, Cheng S. An ensemble model for diabetes diagnosis in large-scale and imbalanced dataset. 2017 Presented at: CF'17: Computing Frontiers Conference; May 15-17, 2017; Siena, Italy p. 71-78. [doi: 10.1145/3075564.3075576]
15. Kim MH, Kim JH, Lee K, Gim G. The prediction of COVID-19 using LSTM algorithms. Int J Netw Distrib Comput 2021;9(1):19. [doi: 10.2991/ijndc.k.201218.003]
16. Pal R, Sekh AA, Kar S, Prasad DK. Neural network based country wise risk prediction of COVID-19. Appl Sci 2020 Sep 16;10(18):6448. [doi: 10.3390/app10186448]
17. Zhang N, Shen S, Zhou A, Jin Y. Application of LSTM approach for modelling stress–strain behaviour of soil. Appl Soft Comput 2021 Mar;100:106959. [doi: 10.1016/j.asoc.2020.106959]
18. Sunny M, Maswood M, Alharbi A. Deep learning-based stock price prediction using LSTM and bi-directional LSTM model. 2020 Presented at: 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES); October 24-26, 2020; Giza, Egypt p. 87-92. [doi: 10.1109/niles50944.2020.9257950]
19. Wang G, Li Z, Li G, Dai G, Xiao Q, Bai L, et al. Real-time liver tracking algorithm based on LSTM and SVR networks for use in surface-guided radiation therapy. Radiat Oncol 2021 Jan 14;16(1):13 [FREE Full text] [doi: 10.1186/s13014-020-01729-7] [Medline: 33446245]
20. Qiao D, Li P, Ma G, Qi X, Yan J, Ning D, et al. Realtime prediction of dynamic mooring lines responses with LSTM neural network model. Ocean Eng 2021 Jan;219:108368. [doi: 10.1016/j.oceaneng.2020.108368]
21. Davagdorj K, Yu S, Kim S, Huy P, Park J, Ryu K. Prediction of 6 months smoking cessation program among women in Korea. Int J Machine Learn Comput 2019 Jul;9(1):83-90 [FREE Full text] [doi: 10.18178/ijmlc.2019.9.1.769]
22. Park H, Batbaatar E, Li D, Ryu K. Risk factors rule mining in hypertension: Korean National Health and Nutrient Examinations Survey 2007-2014. 2016 Presented at: 2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB); January 14-15, 2016; Noida, India p. 1-4. [doi: 10.1109/cibcb.2016.7758128]
23. Park KH, Ishag MIM, Rya KS, Li M, Ryu KH. Efficient ensemble methods for classification on clear cell renal cell carcinoma clinical dataset. In: Nguyen N, Hoang D, Hong TP, Pham H, Trawiński B, editors. Intelligent Information and Database Systems. ACIIDS 2018. Lecture Notes in Computer Science, vol 10752. Cham: Springer; Feb 14, 2018:235-242.
24. Whitney AW. A direct method of nonparametric measurement selection. IEEE Trans Comput 1971 Sep;C-20(9):1100-1103. [doi: 10.1109/T-C.1971.223410]
25. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 1997 Aug;55(1):119-139. [doi: 10.1006/jcss.1997.1504]
26. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/BF00994018]

XSL•FO
RenderX

27. Wright RE. Logistic regression. In: Grimm LG, Yarnold PR, editors. Reading and understanding multivariate statistics. Washington, DC: American Psychological Association; 1995:217-244.
28. Breiman L. Random Forests. Machine Learn 2001 Oct;45:5-32. [doi: 10.1023/A:1010933404324]
29. Dereli O, Oğuz C, Gönen M. A multitask multiple kernel learning algorithm for survival analysis with application to cancer biology. 2019 Presented at: 36th International Conference on Machine Learning; 2019; Long Beach, CA p. 1576-1585 URL: http://proceedings.mlr.press/v97/dereli19a.html
30. Bishop CM. Pattern Recognition and Machine Learning. New York, NY: Springer; Jan 2006.
31. Salzberg SL. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Mach Learn 1994 Sep;16(3):235-240 [FREE Full text] [doi: 10.1007/BF00993309]
32. Breiman L, Friedman JH, Stone CJ, Olshen RA. Classification and regression trees. United Kingdom: Hall/CRC Press; 1984.
33. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
34. Schuster M, Paliwal K. Bidirectional recurrent neural networks. IEEE Trans. Signal Process 1997;45(11):2673-2681. [doi: 10.1109/78.650093] [Medline: 16112549]
35. Dyer C, Ballesteros M, Ling W, Matthews A, Noah A. Transition-based dependency parsing with stack long short-term memory, arXiv, in proceedings of ACL. arXiv. 2015. URL: https://arxiv.org/abs/1505.08075 [accessed 2015-05-29]
36. Cai L, Zhou S, Yan X, Yuan R. A stacked BiLSTM neural network based on coattention mechanism for question answering. Comput Intell Neurosci 2019 Aug 21;2019:9543490 [FREE Full text] [doi: 10.1155/2019/9543490] [Medline: 31531011]
37. Song H, Rajan D, Thiagarajan J, Spanias A. Attend and diagnose: clinical time series analysis using attention models. 2018 Presented at: Thirty-Second AAAI Conference on Artificial Intelligence; February 2-7, 2018; New Orleans, LA URL: https://ojs.aaai.org/index.php/AAAI/article/view/11635
38. Chalasani N, Younossi Z, Lavine JE, Diehl AM, Brunt EM, Cusi K, et al. The diagnosis and management of non-alcoholic fatty liver disease: practice Guideline by the American Association for the Study of Liver Diseases, American College of Gastroenterology, and the American Gastroenterological Association. Hepatology 2012 Jun;55(6):2005-2023 [FREE Full text] [doi: 10.1002/hep.25762] [Medline: 22488764]
39. Marengo A, Rosso C, Bugianesi E. Liver cancer: connections with obesity, fatty liver, and cirrhosis. Annu Rev Med 2016;67:103-117. [doi: 10.1146/annurev-med-090514-013832] [Medline: 26473416]
40. Hancer E. Differential evolution for feature selection: a fuzzy wrapper–filter approach. Soft Comput 2018 Oct 6;23(13):5233-5248 [FREE Full text] [doi: 10.1007/s00500-018-3545-7]
41. MJ Health Database. URL: http://www.mjhrf.org/main/page/resource/en/#resource03 [accessed 2020-02-21]
42. Buuren SV, Groothuis-Oudshoorn K. MICE: multivariate imputation by chained equations in R. J Stat Soft 2011;45(3):1-68. [doi: 10.18637/jss.v045.i03]
43. Ainsworth BE, Haskell WL, Herrmann SD, Meckes N, Bassett DR, Tudor-Locke C, et al. 2011 Compendium of Physical Activities: a second update of codes and MET values. Med Sci Sports Exerc 2011 Aug;43(8):1575-1581. [doi: 10.1249/MSS.0b013e31821ece12] [Medline: 21681120]
44. The Compendium of Physical Activities Tracking Guide. 2002. URL: http://prevention.sph.sc.edu/tools/docs/documents_compendium.pdf [accessed 2021-08-04]
45. Chalasani N, Younossi Z, Lavine JE, Charlton M, Cusi K, Rinella M, et al. The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases. Hepatology 2018 Jan;67(1):328-357. [doi: 10.1002/hep.29367] [Medline: 28714183]
46. Pan JJ, Fallon MB. Gender and racial differences in nonalcoholic fatty liver disease. World J Hepatol 2014 May 27;6(5):274-283 [FREE Full text] [doi: 10.4254/wjh.v6.i5.274] [Medline: 24868321]

Abbreviations

AFLD: alcohol-related fatty liver disease
AUROC: area under the receiver operating characteristic curve
biLSTM: bidirectional long short-term memory
CART: classification and regression trees
CVP: current-visit prediction
FLD: fatty liver disease
GNB: Gaussian naive Bayes
IoU: intersection over union
KNNC: k-nearest neighbor classification
LOO: leave one out
LR: logistic regression
LSTM: long short-term memory
NAFLD: nonalcoholic fatty liver disease

XSL•FO
RenderX

NFLD: no fatty liver disease
NVP: next-visit prediction
OPR: one-pass ranking
RF: random forest
SFS: sequential forward selection
SVM: support vector machine

Edited by G Eysenbach; submitted 10.12.20; peer-reviewed by X Zhang, Y Xiang, E Mastriani; comments to author 05.01.21; revised version received 27.04.21; accepted 03.06.21; published 12.08.21.

Please cite as:
Wu CT, Chu TW, Jang JSR
Current-Visit and Next-Visit Prediction for Fatty Liver Disease With a Large-Scale Dataset: Model Development and Performance Comparison
JMIR Med Inform 2021;9(8):e26398
URL: https://medinform.jmir.org/2021/8/e26398
doi:10.2196/26398
PMID:34387552

©Cheng-Tse Wu, Ta-Wei Chu, Jyh-Shing Roger Jang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 12.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development and Validation of an Arterial Pressure-Based Cardiac Output Algorithm Using a Convolutional Neural Network: Retrospective Study Based on Prospective Registry Data

Hyun-Lim Yang[1,2], PhD; Chul-Woo Jung[1,3], MD, PhD; Seong Mi Yang[1,3], MD, PhD; Min-Soo Kim[4], PhD; Sungho Shim[5], BSc; Kook Hyun Lee[1,3], MD, PhD; Hyung-Chul Lee[1,3], MD, PhD

[1]Department of Anesthesiology and Pain Medicine, Seoul National University Hospital, Seoul, Republic of Korea

[2]Biomedical Research Institute, Seoul National University Hospital, Seoul, Republic of Korea

[3]Department of Anesthesiology and Pain Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea

[4]School of Computing, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

[5]Department of Information and Communication Engineering, Daegu Gyeongbuk Institute of Science & Technology (DGIST), Daegu, Republic of Korea

Corresponding Author:
Hyung-Chul Lee, MD, PhD
Department of Anesthesiology and Pain Medicine
Seoul National University College of Medicine
101 Daehak-ro
Jongno-gu
Seoul, 03080
Republic of Korea
Phone: 82 2 2072 0640
Fax: 82 27478363
Email: vital@snu.ac.kr

Abstract

Background: Arterial pressure-based cardiac output (APCO) is a less invasive method for estimating cardiac output without concerns about complications from the pulmonary artery catheter (PAC). However, inaccuracies of currently available APCO devices have been reported. Improvements to the algorithm by researchers are impossible, as only a subset of the algorithm has been released.

Objective: In this study, an open-source algorithm was developed and validated using a convolutional neural network and a transfer learning technique.

Methods: A retrospective study was performed using data from a prospective cohort registry of intraoperative bio-signal data from a university hospital. The convolutional neural network model was trained using the arterial pressure waveform as input and the stroke volume (SV) value as the output. The model parameters were pretrained using the SV values from a commercial APCO device (Vigileo or EV1000 with the FloTrac algorithm) and adjusted with a transfer learning technique using SV values from the PAC. The performance of the model was evaluated using absolute error for the PAC on the testing dataset from separate periods. Finally, we compared the performance of the deep learning model and the FloTrac with the SV values from the PAC.

Results: A total of 2057 surgical cases (1958 training and 99 testing cases) were used in the registry. In the deep learning model, the absolute errors of SV were 14.5 (SD 13.4) mL (10.2 [SD 8.4] mL in cardiac surgery and 17.4 [SD 15.3] mL in liver transplantation). Compared with FloTrac, the absolute errors of the deep learning model were significantly smaller (16.5 [SD 15.4] and 18.3 [SD 15.1], $P<.001$).

Conclusions: The deep learning–based APCO algorithm showed better performance than the commercial APCO device. Further improvement of the algorithm developed in this study may be helpful for estimating cardiac output accurately in clinical practice and optimizing high-risk patient care.

(JMIR Med Inform 2021;9(8):e24762) doi:10.2196/24762

XSL•FO
RenderX

KEYWORDS

cardiac output; deep learning; arterial pressure

Introduction

Cardiac output (CO; L/min), the amount of blood pumped from the left ventricle per minute, is the main determinant of oxygen delivery to the body, including to the brain and vital organs, and is an important monitoring parameter during hemodynamic optimization. It is sometimes referred to as the stroke volume (SV; mL/beat), calculated by dividing the CO by the heart rate (HR; beats per minute). Particularly, in the perioperative phase, hemodynamic optimization is directly related to postoperative complications, which are the third leading cause of death worldwide [1]. Patients' outcomes can potentially be improved by applying immediate treatment to maintain CO within 4-8 L/min or maintain SV within 60-100 mL/beat during major surgery [2]. Optimization of CO is also essential for high-risk patients [3]. Early interventions for hemodynamic control can significantly reduce mortality by more than 20% in high-risk patients [4].

The thermodilution method using a pulmonary artery catheter (PAC) has been regarded as a gold standard for measuring CO in clinical practice [5]. However, owing to its invasiveness, the risks associated with placement limit its use to only cardiac surgery, liver transplantations, and some critically ill patients. Instead, arterial pressure-based cardiac output (APCO) methods have been proposed as a less invasive method for estimating CO from the arterial pressure waveform without the risk of complications associated with a PAC [6,7]. These methods estimate systemic vascular resistance from arterial pressure waveform and general patient characteristics and predict the SV. As the arterial line is less invasive and usually inserted for continuous blood pressure monitoring, APCO devices such as the FloTrac (Edwards Lifesciences, Irvine, CA, United States) or LiDCO Rapid (LiDCO Ltd, London, UK) are widely used in perioperative CO management. However, inaccuracies of the commercially available APCO devices have been reported, especially, in sepsis or liver transplantation patients [8,9]. Improvements of the algorithm by researchers are also not possible, because only a subset of the algorithm has been openly released.

Recent advances in machine learning techniques have led to many new approaches to solving clinical problems [10]. Deep learning techniques, such as convolutional neural networks (CNNs), have performed well in bio-signal analysis [11]. In contrast, the shortage of clinical bio-signal data makes it difficult to train deep learning models properly [12,13]. Publicly available bio-signal datasets are still limited compared with medical imaging or structured datasets [14-16]. Furthermore, data with reduced clinical use, such as PAC-based CO, worsen this tendency. In this case, after training a model with a relatively common dataset, a transfer learning technique can be used to refine the model parameters with relatively rare data. Previous studies also reported performance gains from transfer learning with bio-signal data [17-20].

In this study, a novel APCO algorithm was built using a transfer learning technique. The algorithm learned the commercial APCO algorithm and then was trained with less-common PAC data. In addition, several preprocessing techniques were proposed to analyze the arterial pressure waveforms for predicting CO. Finally, the deep learning model was validated using real-world bio-signal data, which was collected during a separate period than the training data and includes cardiac surgery and liver transplantation patients. This study hypothesized that a model developed using real-world clinical data, deep learning techniques, and transfer learning techniques can be more accurate than a commercial APCO device for estimating CO.

Methods

Study Approval

All data used in this study were obtained from the prospective registry of the vital signs for surgical patients at Seoul National University Hospital. The registry was approved by the Institutional Review Board of Seoul National University Hospital (H-1408-101-605) and registered at the clinical trial registration site (ClinicalTrials.gov, NCT02914444). This retrospective study was also approved by the Institutional Review Board (H-2007-015-1138). However, the need for written informed consent was waived because of the anonymity of the data.

Data Collection

The registry collected synchronous vital signs and bio-signal data from various medical devices using the Vital Recorder Program [21]. Among the cases in the registry, those from between August 2016 and September 2019 were used in this study. The cases collected in the last 8 months of the study period (February 2019 to September 2019) were used for the testing dataset. The remaining cases were used for training the model.

During data collection, CO monitors were used according to the discretion of the anesthesiologist. CO values were collected at 2-second intervals from the serial port of APCO devices, such as the EV1000 clinical platform or the Vigileo system (Edwards Lifesciences, Irvine, CA, United States) with a fourth-generation FloTrac algorithm or a PAC-based device such as the Vigilance II (Edwards Lifesciences, Irvine, CA, United States). The arterial pressure waveform was recorded at 500 Hz from the analog output port of the TRAM module (GE Healthcare, Chicago, IL, United States), and the heart rate was recorded at 2-second intervals from the serial port of the Solar 8000 patient monitor (GE Healthcare, Chicago, IL, United States). General patient characteristics (ie, age, sex, weight, and height) were collected from electronic medical records.

The deep learning model requires massive amounts of data for good performance. However, with PAC-based CO monitoring data, such as the Vigilance II, it is difficult to retain massive amounts of data, which hinders the ability to develop of a good

XSL•FO
RenderX

deep learning model using real-world databases. Hence, the model was pretrained using APCO data from the EV1000 or Vigileo, from which data are relatively easy to obtain. After that, we tuned the model using PAC data from the Vigilance II, from which data are hard to obtain. In total, 1572 cases of surgery were recorded with APCO monitoring devices for pretraining, 290 cases were recorded with PAC-based CO monitoring devices for tuning or testing, and 195 cases were recorded with both APCO and PAC-based CO monitoring devices for tuning or testing. Among the 2057 cases, 1958 cases (95.19%) that were operated on from August 2016 to January 2019 were used for pretraining or tuning, and the remaining 99 cases (4.81%) that were operated on since February 2019 were used for testing.

Data Preprocessing

A dataset of arterial pressure waveforms was preprocessed, including corresponding SV values. The arterial pressure waveforms were resampled from 500 Hz to 100 Hz and sliced into 20-second segments. Each pair of a 20-second segment and an SV was referred to as a "sample." For preprocessing the samples, the following 4 steps were performed: (1) converting the output of PAC-based CO data to SV, (2) smoothing the APCO data, (3) delaying the PAC data, and (4) removing unsuitable samples.

The APCO monitoring device provides the SV value, because it estimates the amount of blood moving from arterial waves when a stroke occurs. In contrast, a PAC-based CO monitoring device emits the CO value, because it measures the temperature change by blood flow from the pulmonary artery catheter. Owing to the physiological differences between the 2 methods, CO values needed to be converted from the Vigilance II monitor to SV, using HR values (SV = CO/HR), to synchronize with APCO data.

In addition, because there were larger fluctuations in the APCO data than clinically expected CO changes, the APCO data were smoothed using a locally weighted scatterplot smoother (LOWESS) algorithm [22]. We used the hyperparameter, λ=0.03, of the LOWESS algorithm. If APCO data were not recorded for more than 200 seconds in a single case due to recording errors, the LOWESS algorithm was applied separately.

In addition, the PAC-based SV value was delayed, to synchronize the time differences between arterial pressure waveforms and PAC data. There are several minutes of delay in the CO values using the Vigilance II in its "Trend" mode [23,24]. To determine the time lag of the CO values in the Trend modes, we used the device with the "STAT" mode in several cases, allowing both the CO values and the "CO stat" values to be transferred. Then, the CO values were compared with the CO stat values, and we obtained the minimum mean absolute difference with the delay of 2 minutes. Thus, PAC-based SV values were shifted to the time 2 minutes earlier than the recorded time. More detailed descriptions and examples of delaying PAC values are provided in Multimedia Appendix 1.

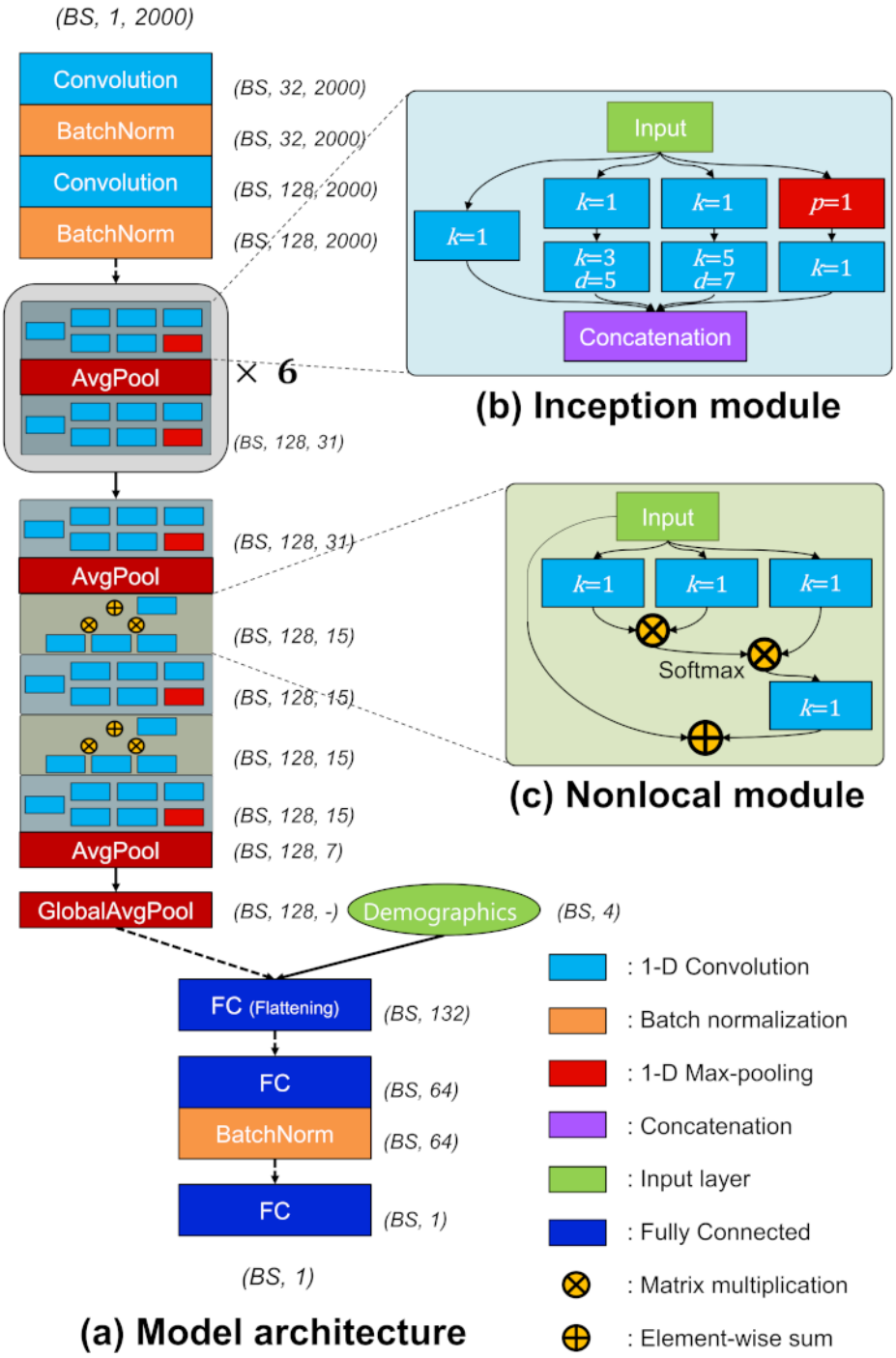
Finally, unsuitable samples were removed for robust deep learning model training. Samples with a blood pressure <25 mmHg or >250 mmHg and an SV <20 mL or >200 mL were removed from the dataset. After applying a beat-detection algorithm, we eliminated samples with an HR <30 beats/min or >180 beats/min, a pulse pressure <20 mm Hg, or with frequent (>50%) ventricular premature beats [25].

Model Building

A CNN model was designed to learn the appropriate feature extraction from the 20-second segments of arterial pressure waveforms. The goal of the model was to estimate PAC-based CO values using the arterial pressure waveform and the patient's demographics (ie, age, sex, weight, and height) (Figure 1). The model consists of 2 parts: feature extraction and regression. The feature extraction part of the model was composed of 2 successive pairs of convolution and batch-normalization layers, 15 inception modules, 2 nonlocal modules with pooling, and dropout layers [26-29]. The regression part of the model, which was composed of 3 fully connected, batch-normalization, and dropout layers, takes a concatenation of the extracted features from the feature extraction part and the patient's demographic information as input and returns a predicted SV value. For all layers, a rectified linear unit was used as the activation function.

XSL•FO
RenderX

Figure 1. Proposed convolutional neural network model for estimating stroke volume from arterial pressure waveform. (A) Overall model architecture. (B) Details of the inception module. (C) Details of the nonlocal module. The variable *k* indicates kernel size of the convolution layer, *d* means the dilation rate of the convolution layer, and *p* represents pooling rate. GlobalAvgPool indicates the global average pooling layer, which computes the mean value for each feature map and supplies abstracted feature maps to the flattening layer. Dotted arrows represent dropout of 0.5, while solid arrows mean full connections. BS: batch size; FC: fully connected.



Input samples were abstracted by 2 consecutive pairs of convolution and batch-normalization layers (dropout rate of 0.5), before feeding them into the inception modules. The inception modules consisted of 4 paths with 6 convolution layers and 1 pooling layer. The filter size of each path in the inception module was 32, and the outputs of the 4 paths were concatenated to 128. The detailed configurations of the inception module are illustrated in Figure 1B. Note that *p* represents the pooling rate of the average pooling layer, *k* indicates the kernel size of the convolution layer, and *d* is the dilation rate of the dilated convolution layer. After an odd-numbered inception module, an average pooling layer intervened to reduce the size of feature

XSL•FO
RenderX

maps from the previous inception module. The shaded inception module block in Figure 1A was repeated 6 times. Nonlocal modules were added just before the last 2 inception modules to consider the global covariance in each segment. The first fully connected layer was a flattening layer (size of 132) that took the concatenation of the average of feature maps by global average pooling (size of 128) and demographic information (size of 4). The number of neurons in the last 2 fully connected layers was 64 and 1, respectively. Immediately after the second fully connected layers, there was a batch-normalization layer to achieve robustness for cases in which a batch is biased to a specific SV range. The last fully connected layer consisted of a single neuron, which represents a float value of the SV. For all fully connected layers, a dropout of 0.5 was used. The output dimensions of each module or layer are described in Figure 1. Note that dimensions are represented as *BS*, channel, and length, where *BS* indicates batch size. The dotted arrows indicate a dropout of 0.5.

Model Training

Our model was trained using a transfer learning method with 2 steps: pretraining and tuning. During pretraining, SV values were used from EV1000 or Vigileo to find rough parameters of the model for analyzing arterial pressure waveforms. The input of pretraining was 20-second segments of arterial pressure waveform and patient demographic information, and the output was the predicted SV. After that, the parameters were tuned using PAC data. The input and output of tuning were the same as in pretraining; however, the parameters of the regressor part were initialized with the Xavier algorithm to be retrained with PAC data [30]. Gradient-descent optimizers, RAdam, and Lookahead were used to update parameters [31,32]. The batch size was 512, and the loss function was the root mean squared error. Note that the loss function was calculated based on the equation ($\hat{Y}$ = predicted value; Y = ground truth; N = batch size):

$$\text{Root mean squared error} = (\sum^{N}(\hat{Y}-Y)^2/N)^{1/2} \quad \textbf{(1)}$$

The model was tested every 200 steps using 30% of the training datasets to calculate the validation errors. The training was stopped when the validation errors no longer decreased after 50 times and then was restarted with a smaller learning rate (decay rate of 0.5). The training was performed using our own program (code available at [33]), written in Python using PyTorch 1.1.0 on a graphics processing unit server with two 10-core Intel Xeon central processing units and eight Nvidia GTX 1080Ti graphics processing units.

Statistical Analysis

The statistics of patient demographics were described in the training and testing groups, comparing the heterogeneity of the variables. Note that training groups contained pretraining and tuning datasets, and the testing group contained the testing dataset. For continuous variables (ie, age, height, and weight), a Mann-Whitney *U* test was performed for comparisons after testing for normality using a Shapiro-Wilk test. For categorical variables (ie, sex), a Pearson chi-square test was conducted for comparisons between groups.

The performance of the deep learning model was validated using error, absolute error, percentage error, and absolute percentage error, using the testing dataset. Each metric was calculated based on the equation ($\hat{Y}$ = predicted value; Y = PAC value; N = number of samples):

$$\text{Error (mL)} = \sum^{N}(\hat{Y}-Y)/N \quad \textbf{(2)}$$

$$\text{Absolute error (mL)} = \sum^{N}|\hat{Y}-Y|/N \quad \textbf{(3)}$$

$$\text{Percentage error (\%)} = [\sum^{N}\{(\hat{Y}-Y)/Y\}/N] \times 100 \quad \textbf{(4)}$$

$$\text{Absolute percentage error (\%)} = \{\sum^{N}|(\hat{Y}-Y)/Y|/N\} \times 100 \quad \textbf{(5)}$$

Efforts were made to validate the generalizability and substitutability of our model by comparing its performance with the FloTrac in two radically different patient groups: the patients who underwent cardiac surgery and the patients who underwent liver transplantation surgery. Among the test dataset, both the FloTrac and Vigilance II devices were used simultaneously in 16 cases of cardiac surgery and 40 cases of liver transplantation surgery. With these 56 cases, direct comparisons were performed between the deep learning model and FloTrac using a paired *t* test for overall cases and each subgroup.

Spearman correlation coefficients were calculated between the SV of the deep learning model and PAC, and between the SV of the EV1000 or Vigileo and PAC. Bland-Altman analysis was used to test the agreement of either the pair of the deep learning model and PAC-based SV or the pair of the FloTrac and PAC-based SV [34]. Bias was defined as the mean difference between SVs, and the upper and lower limits of agreement were defined as ±1.96 SDs of the bias. The trending ability of the deep learning model was examined using a 4-quadrant plot analysis [35]. The concordance rate of the association for percentage changes in SV was calculated between our model or the FloTrac and PAC, with the exclusion of 10% of the changes [36].

All data are expressed as the mean (SD), median (interquartile range), or absolute numbers (%). *P* values <.05 were considered statistically significant. Statistical analyses were performed using Python Scipy 1.4.1.

Results

Data from 2057 surgical cases (1232 general, 59.89%; 636 thoracic, 30.92%; 159 urologic, 7.73%; 23 gynecologic, 1.12%; 6 otolaryngologic, 0.29%; and 1 plastic surgery, 0.05%) in the registry were extracted and preprocessed. Of these 2057 surgical cases, we used the data from 1958 cases (95.19%) for training and 99 cases (4.81%; 59 cardiac surgery, 60%; 40 liver transplantation, 40%) for testing. For transfer learning, of 2057 surgical cases, we used the data from 1572 cases (76.42%) for pretraining and the data from 386 cases (18.77%; 245 cardiac surgeries, 63.5%; 141 liver transplantations, 36.5%) for tuning. Demographic information of the patients was not different between the training and testing datasets, except, that the patients in the testing dataset were slightly older (Table 1).

XSL•FO
RenderX

Table 1. Demographics of patients for the training and testing dataset.

Characteristic	Training dataset (n=1958)	Testing dataset (n=99)	Statistical test	
			Method used	*P* value
Age, median (interquartile range), years	61.2 (51.2-69.5)	63.8 (57.4-71.9)	Mann-Whitney *U* test	.02
Sex, number male (%)	1195 (61.03%)	66 (67%)	Pearson chi-square test	.309
Height, median (interquartile range), years	164.0 (157.3-170.0)	163.4 (156.7-168.0)	Mann-Whitney *U* test	.231
Weight, median (interquartile range), kg	63.1 (55.2-72.4)	64.0 (57.5-72.1)	Mann-Whitney *U* test	.494

The absolute error of the deep learning model for the testing dataset was 14.5 (SD 13.4) mL (Table 2). In the subgroup analysis, the absolute errors of the deep learning model were 10.2 (SD 8.4) mL for cardiac surgery and 17.4 (SD 15.3) for liver transplantation.

Table 2. Stroke volume estimation of the deep learning model.

Measure	Overall (n=99), mean (SD)	Cardiac surgery (n=59), mean (SD)	Liver transplantation (n=40), mean (SD)
Error (mL)	–4.4 (19.2)	2.3 (13.0)	–9.0 (21.3)
Absolute error (mL)	14.5 (13.4)	10.2 (8.4)	17.4 (15.3)
Percentage error (%)	0.4 (27.4)	9.0 (26.8)	–5.5 (26.2)
Absolute percentage error (%)	20.5 (18.2)	20.5 (19.4)	20.4 (17.4)

In the testing dataset, the data from 56 cases with both PAC and FloTrac (16 cardiac surgery, 29%; 40 liver transplantation, 71%) data were used to compare the performance of the deep learning model with that of the FloTrac. The absolute error of the deep learning model was significantly lower than that of the FloTrac (16.5 mL vs 18.3 mL, *P*<.001; Table 3). In the subgroup analysis, the absolute errors of the deep learning model were lower than those of the FloTrac, in both cardiac surgery (11.1 mL vs 14.3 mL, *P*<.001) and liver transplantation (17.4 mL vs 19.0 mL, *P*<.001; Table 3). The individual plots of the time course of the measured and predicted SVs in all 99 testing cases are available in Multimedia Appendix 2.

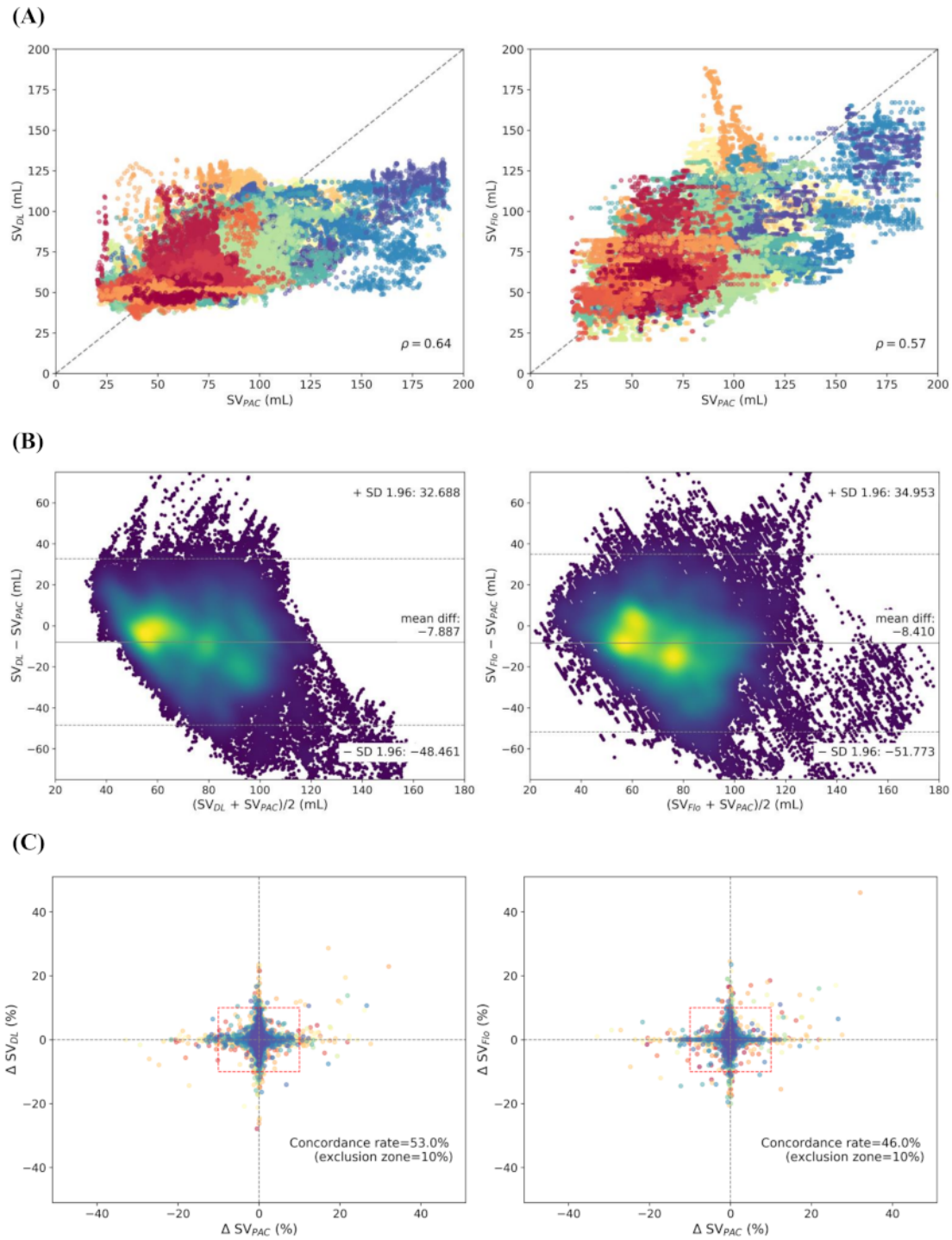
Table 3. Comparison of performance in stroke volume estimation between the deep learning model and FloTrac algorithm.

Measure	Deep learning model, mean (SD)	FloTrac, mean (SD)	Statistical test	
			Count, n	*P* value
Error (mL)				
Overall (n=56)	–7.9 (20.7)	–8.4 (22.1)	158725	<.001
Cardiac surgery (n=16)	–1.7 (15.2)	2.7 (18.3)	65260	<.001
Liver transplantation (n=40)	–9.0 (21.3)	–10.3 (22.2)	93465	<.001
Absolute error (mL)				
Overall (n=56)	16.5 (14.8)	18.3 (15.1)	158725	<.001
Cardiac surgery (n=16)	11.1 (10.5)	14.3 (11.8)	65260	<.001
Liver transplantation (n=40)	17.4 (15.3)	19.0 (15.4)	93465	<.001
Percentage error (%)				
Overall (n=56)	–4.4 (26.9)	–5.6 (28.6)	158725	<.001
Cardiac surgery (n=16)	1.8 (29.9)	9.5 (34.9)	65260	<.001
Liver transplantation (n=40)	–5.5 (26.2)	–8.3 (26.5)	93465	<.001
Absolute percentage error (%)				
Overall (n=56)	20.3 (18.3)	22.5 (18.5)	158725	<.001
Cardiac surgery (n=16)	19.3 (22.9)	25.4 (25.7)	65260	<.001
Liver transplantation (n=40)	20.4 (17.4)	22.0 (16.9)	93465	<.001

XSL•FO
RenderX

The Spearman rho value of our deep learning model was 0.64 (P<.001), whereas the rho value of the FloTrac was 0.57 (P<.001; Figure 2). Bland-Altman analysis demonstrated that the lower and upper limits of agreements, respectively, were –48.5 (95% CI –48.7 to –48.3) mL and 32.7 (95% CI 32.5-33.0) mL for the deep learning model and –51.8 (95% CI –52.0 to –51.5) mL and 35.0 (95% CI 34.8-35.3) mL for the FloTrac. The 4-quadrant plot showed concordance rates of 53% for the deep learning model and 46% for FloTrac (Figure 2). Mean differences were smaller in our deep learning model than in the commercial APCO device.

Figure 2. Scatter plot. (A) Bland-Altman plot with density highlight. (B) Four-quadrant plots. (C) Plot between target stroke volume and predicted stroke volume. DL: deep learning; Flo: FloTrac; PAC: pulmonary artery catheter; SV: stroke volume.



XSL•FO
RenderX

Discussion

Principal Results

In this study, we built and evaluated an open-source deep learning–based APCO algorithm using a large set of prospectively collected registry data. The performance of the deep learning model was better than that of the FloTrac, the commercially available APCO algorithm.

A mathematical analysis of the association between arterial blood flow and pressure waveform has a long history of over 100 years [37]. A key factor in this flow-pressure association is the estimation of systemic vascular resistance (SVR), because the flow is determined by the pressure gradient and vascular resistance [38]. However, since SVR changes with the patient's condition, the coefficient also needs to be updated in real-time [39-41]. An uncalibrated APCO algorithm, which automatically updates the coefficient using the patient's general characteristics and arterial pressure waveforms, can be a convenient solution in clinical situations. However, currently available commercial uncalibrated APCO devices have been reported to work poorly in patients with vasodilatory states such as sepsis or liver transplantation [8,9,38,42,43]. Our results showed that our deep learning–based APCO algorithm outperformed the FloTrac algorithm in both cardiac surgery and liver transplantation patients. However, both the deep learning–based and FloTrac algorithms showed a positive bias in cardiac surgery patients, who usually have low SVs, and a negative bias in liver transplantation patients, who usually have high SVs. This tendency to return to the average is also shown in Figure 2C and may be a fundamental limitation of the APCO algorithm, in which SVR should be estimated only from the arterial pressure waveform and patient's demographics. Otherwise, this tendency may occur because most of the data obtained from routine clinical practice are within the normal SV range.

Clinical, Academic, and Technical Implications

The measurement of CO is essential for clinical hemodynamic optimization. As the deep learning model is more accurate than the commercial APCO device, it can help enhance patient management and improve final outcomes. For example, goal-directed fluid therapy or SV optimization can be performed using our model. However, further validation and implementation are required for clinical applications. Disclosing our dataset and model, researchers can improve the model and validate our algorithm or their own algorithm. We believe that this approach can facilitate developing more accurate APCO algorithms and help in its clinical application. In the domain of detecting arrhythmia in the electrocardiogram, numerous studies and technologies have been proposed using publicly available data, such as the MIT-BIH dataset [44-47]. Likewise, our open dataset can be an academic reference for the APCO domain. Finally, a transfer learning method was proposed based on 2 datasets with different characteristics in bio-signal fields, and its scalability was confirmed. These techniques will provide good technical strategies for developing machine learning algorithms in the medical field with scanty data, such as PAC-based CO.

Comparison With Prior Work

In a previous study, Moon et al [48] built a deep learning–based APCO algorithm using the data of 31 liver transplantation patients. However, their model only included the patients who underwent liver transplantation and has not been validated in the other types of surgery. In this study, a larger dataset was used that included both cardiac surgery and liver transplantation cases, with balanced ratios. In addition, a transfer learning technique was used, in which the parameters were pretrained with a large amount of APCO data and then tuned with PAC data. This may explain why our model worked better than the FloTrac for both cardiac and liver transplantation cases.

Limitations

This study has some limitations. First, the data used in this study were from a single-center registry of a surgical cohort, which may contain a limited range of CO. This problem can be overcome by adding more data. However, the clinical use of PAC is gradually decreasing; other modalities such as a Doppler flowmeter or echocardiogram are required. Second, continuous CO measurement methods used as ground-truth values in this study can be less accurate in certain situations, such as in rapid fluid administration, compared with the gold-standard intermittent thermodilution technique [49]. In addition, the delay time for processing in the Vigilance II monitor was not fully revealed [24]. Third, there was a statistical difference in age between the training and testing sets. This was an inevitable problem that occurred because of the use of real-world clinical data based on a prospective registry. However, elderly patients may have isolated systolic hypertension, which may alter the arterial pressure waveforms and affect the results. Fourth, there was no visualization with explainable artificial intelligence algorithms of how the proposed algorithm produces the results [50,51]. A proposal for a method that can display an indication for high or low CO and SVR from a waveform would have great clinical benefit. Finally, although various technological methods have been adopted, our developed model may be a local optimum and not a global optimum. Therefore, the raw data of this study was disclosed, allowing other researchers to improve the model.

Conclusions

In conclusion, an uncalibrated APCO algorithm was developed and validated using a CNN and a transfer learning technique. The performance of our model was better than that of current commercial, uncalibrated APCO devices. Further improvement of the open-source algorithm developed in this study may be helpful for estimating cardiac output accurately in clinical practice and optimizing high-risk patient care.

XSL•FO
RenderX

Acknowledgments

This work was supported by the Seoul National University Hospital (SNUH) research fund (grant 06-2006-2449); the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT), Korea (NRF-2018R1A5A1060031); and Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2018-0-01833), supervised by the Institute for Information & Communications Technology Promotion (IITP).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details of delaying the pulmonary artery catheter cardiac output data.
[DOCX File , 504 KB - medinform_v9i8e24762_app1.docx]

Multimedia Appendix 2
The individual plots of the time course of the measured and predicted stroke volume in all 99 testing cases.
[PDF File (Adobe PDF File), 13581 KB - medinform_v9i8e24762_app2.pdf]

References

1. Nepogodiev D, Martin J, Biccard B, Makupe A, Bhangu A, National Institute for Health Research Global Health Research Unit on Global Surgery. Global burden of postoperative death. Lancet 2019 Feb 02;393(10170):401. [doi: 10.1016/S0140-6736(18)33139-8] [Medline: 30722955]
2. Giglio MT, Marucci M, Testini M, Brienza N. Goal-directed haemodynamic therapy and gastrointestinal complications in major surgery: a meta-analysis of randomized controlled trials. Br J Anaesth 2009 Nov;103(5):637-646 [FREE Full text] [doi: 10.1093/bja/aep279] [Medline: 19837807]
3. Stevenson LW, Tillisch JH. Maintenance of cardiac output with normal filling pressures in patients with dilated heart failure. Circulation 1986 Dec;74(6):1303-1308. [doi: 10.1161/01.cir.74.6.1303] [Medline: 3779915]
4. Kern JW, Shoemaker WC. Meta-analysis of hemodynamic optimization in high-risk patients. Crit Care Med 2002 Aug;30(8):1686-1692. [doi: 10.1097/00003246-200208000-00002] [Medline: 12163777]
5. Drummond KE, Murphy E. Minimally invasive cardiac output monitors. Continuing Education in Anaesthesia Critical Care & Pain 2012 Feb;12(1):5-10. [doi: 10.1093/bjaceaccp/mkr044]
6. Marik PE. Obituary: pulmonary artery catheter 1970 to 2013. Ann Intensive Care 2013;3(1):38. [doi: 10.1186/2110-5820-3-38]
7. Connors AF, Speroff T, Dawson NV, Thomas C, Harrell FE, Wagner D, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. JAMA 1996 Sep 18;276(11):889-897. [doi: 10.1001/jama.276.11.889] [Medline: 8782638]
8. Hattori K, Maeda T, Masubuchi T, Yoshikawa A, Ebuchi K, Morishima K, et al. Accuracy and Trending Ability of the Fourth-Generation FloTrac/Vigileo System in Patients With Low Cardiac Index. Journal of Cardiothoracic and Vascular Anesthesia 2017 Feb;31(1):99-104. [doi: 10.1053/j.jvca.2016.06.016]
9. Sakka SG, Kozieras J, Thuemer O, van Hout N. Measurement of cardiac output: a comparison between transpulmonary thermodilution and uncalibrated pulse contour analysis. Br J Anaesth 2007 Sep;99(3):337-342 [FREE Full text] [doi: 10.1093/bja/aem177] [Medline: 17611251]
10. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun. ACM 2017 May 24;60(6):84-90. [doi: 10.1145/3065386]
11. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019 Jan 7;25(1):65-69. [doi: 10.1038/s41591-018-0268-3]
12. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
13. Lee H, Jung C. Vital Recorder-a free research tool for automatic recording of high-resolution time-synchronised physiological data from multiple anaesthesia devices. Sci Rep 2018 Dec 24;8(1):1527 [FREE Full text] [doi: 10.1038/s41598-018-20062-4] [Medline: 29367620]
14. Beaulieu-Jones B, Finlayson SG, Chivers C, Chen I, McDermott M, Kandola J, et al. Trends and Focus of Machine Learning Applications for Health Research. JAMA Netw Open 2019 Oct 25;2(10):e1914051. [doi: 10.1001/jamanetworkopen.2019.14051]
15. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
16. Faust O, Hagiwara Y, Hong TJ, Lih OS, Acharya UR. Deep learning for healthcare applications based on physiological signals: A review. Computer Methods and Programs in Biomedicine 2018 Jul;161:1-13. [doi: 10.1016/j.cmpb.2018.04.005]

XSL•FO
RenderX

17. Van Steenkiste G, van Loon G, Crevecoeur G. Transfer Learning in ECG Classification from Human to Horse Using a Novel Parallel Neural Network Architecture. Sci Rep 2020 Jan 13;10(1):186. [doi: 10.1038/s41598-019-57025-2]
18. Murugesan B, Ravichandran V, Ram KS. ECGNet: Deep Network for Arrhythmia Classification. In: P. P, Joseph J, Shankaranarayana SM, Sivaprakasam M. ECGNet: Deep Network for Arrhythmia Classification. 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA) Internet Rome: IEEE; 2018 Jun 11 Presented at: 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA); 11-13 June 2018; Rome, Italy. [doi: 10.1109/memea.2018.8438739]
19. Ju C, Gao D, Mane R, Tan B, Liu Y, Guan C. Federated Transfer Learning for EEG Signal Classification. In: Federated Transfer Learning for EEG Signal Classification. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) Internet Montreal, QC. Canada: IEEE; 2020 Presented at: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); 20-24 July 2020; 2020: Montreal, QC, Canada. [doi: 10.1109/embc44109.2020.9175344]
20. Zheng W, Lu B. Personalizing EEG-based affective models with transfer learning. 2016 Presented at: Proceedings of the twenty-fifth international joint conference on artificial intelligence; 9-15 July 2016; New York, NY, USA.
21. Vital DB. URL: https://vitaldb.net [accessed 2021-06-07]
22. Cleveland WS, Devlin SJ. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. Journal of the American Statistical Association 1988 Sep;83(403):596-610. [doi: 10.1080/01621459.1988.10478639]
23. Lazor MA, Pierce ET, Stanley GD, Cass JL, Halpern EF, Bode RH. Evaluation of the accuracy and response time of stat-mode continuous cardiac output. Journal of Cardiothoracic and Vascular Anesthesia 1997 Jun;11(4):432-436. [doi: 10.1016/s1053-0770(97)90050-1]
24. Lakhal K, Ehrmann S, Boulain T. Predictive performance of passive leg raising in patients with atrial fibrillation. Br J Anaesth 2016 Sep;117(3):399 [FREE Full text] [doi: 10.1093/bja/aew233] [Medline: 27543538]
25. Aboy M, McNames J, Thong T, Tsunami D, Ellenby M, Goldstein B. An Automatic Beat Detection Algorithm for Pressure Signals. IEEE Trans. Biomed. Eng 2005 Oct;52(10):1662-1670. [doi: 10.1109/tbme.2005.855725]
26. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. 2015 Presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015; Boston, MA, USA. [doi: 10.1109/cvpr.2015.7298594]
27. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. 2018 Presented at: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2018; Salt Lake City, UT, USA p. 7794. [doi: 10.1109/cvpr.2018.00813]
28. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015 Presented at: Proceedings of the 32nd International Conference on Machine Learning, ICML 2015; 6-11 July 2015; Lille, France p. 448.
29. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res 2014:1929-1958.
30. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. 2010 Presented at: Proceedings of the thirteenth international conference on artificial intelligence and statistics (AISTATS); 2010; Chia Laguna Resort, Sardinia, Italy, p. 249.
31. Liu L, Jiang H, He P, Chen W, Liu X, Gao J, et al. On the Variance of the Adaptive Learning Rate and Beyond. : OpenReview.net; 2020 Presented at: International Conference on Learning Representations (ICLR); April 26-30, 2020; Addis Ababa, Ethiopia URL: https://openreview.net/forum?id=rkgz2aEKDr
32. Zhang M, Lucas J, Ba J, Hinton G. Lookahead Optimizer: k steps forward, 1 step back. 2019 Presented at: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019; December 8-14, 2019; Vancouver, BC, Canada URL: https://proceedings.neurips.cc/paper/2019/file/90fd4f88f588ae64038134f1eeaa023f-Paper.pdf
33. APCONet: AI model for arterial pressure waveform derived cardiac output. GitHub. 2020 Sep 28. URL: https://github.com/hyunlimy/APCONet [accessed 2021-07-27]
34. Bland JM, Altman DG. Agreement Between Methods of Measurement with Multiple Observations Per Individual. Journal of Biopharmaceutical Statistics 2007 Jul 05;17(4):571-582. [doi: 10.1080/10543400701329422]
35. Saugel B, Grothe O, Wagner JY. Tracking Changes in Cardiac Output: Statistical Considerations on the 4-Quadrant Plot and the Polar Plot Methodology. Anesth Analg 2015 Aug;121(2):514-524. [doi: 10.1213/ANE.0000000000000725] [Medline: 26039419]
36. Nordström J, Hällsjö-Sander C, Shore R, Björne H. Stroke volume optimization in elective bowel surgery: a comparison between pulse power wave analysis (LiDCOrapid) and oesophageal Doppler (CardioQ). British Journal of Anaesthesia 2013 Mar;110(3):374-380. [doi: 10.1093/bja/aes399]
37. Frank O. The basic shape of the arterial pulse. First treatise: mathematical analysis. 1899. J Mol Cell Cardiol 1990 Mar;22(3):255-277. [doi: 10.1016/0022-2828(90)91460-o] [Medline: 21438422]
38. Thiele RH, Durieux ME. Arterial Waveform Analysis for the Anesthesiologist. Anesthesia & Analgesia 2011;113(4):766-776. [doi: 10.1213/ane.0b013e31822773ec]

XSL•FO
RenderX

39. Sun JX, Reisner AT, Saeed M, Heldt T, Mark RG. The cardiac output from blood pressure algorithms trial. Critical Care Medicine 2009;37(1):72-80. [doi: 10.1097/ccm.0b013e3181930174]
40. Rhodes A, Sunderland R. Arterial Pulse Power Analysis: The LiDCO plus System. In: Update in Intensive Care and Emergency Medicine. Berlin, Heidelberg: Springer; 2005:183.
41. Gödje O, Höke K, Goetz AE, Felbinger TW, Reuter DA, Reichart B, et al. Reliability of a new algorithm for continuous cardiac output determination by pulse-contour analysis during hemodynamic instability. Crit Care Med 2002 Jan;30(1):52-58. [doi: 10.1097/00003246-200201000-00008] [Medline: 11902287]
42. Hofer CK, Senn A, Weibel L, Zollinger A. Assessment of stroke volume variation for prediction of fluid responsiveness using the modified FloTrac and PiCCOplus system. Crit Care 2008;12(3):R82 [FREE Full text] [doi: 10.1186/cc6933] [Medline: 18570641]
43. Alhashemi JA, Cecconi M, Hofer CK. Cardiac output monitoring: an integrative perspective. Crit Care 2011;15(2):214. [doi: 10.1186/cc9996]
44. Moody G, Mark R. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng. Med. Biol. Mag 2001;20(3):45-50. [doi: 10.1109/51.932724]
45. Mousavi S, Afghah F. Inter- and Intra- Patient {ECG} Heartbeat Classification for Arrhythmia Detection: {A} Sequence to Sequence Deep Learning Approach. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Internet Brighton, United Kingdom: IEEE; 2019 Presented at: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP; May 12-17, 2019; Brighton, United Kingdom. [doi: 10.1109/icassp.2019.8683140]
46. Li R, Zhang X, Dai H, Zhou B, Wang Z. Interpretability Analysis of Heartbeat Classification Based on Heartbeat Activity's Global Sequence Features and BiLSTM-Attention Neural Network. IEEE Access 2019;7:109870-109883. [doi: 10.1109/access.2019.2933473]
47. Kachuee M, Fazeli S, Sarrafzadeh M. ECG Heartbeat Classification: A Deep Transferable Representation. 2018 Presented at: IEEE International Conference on Healthcare Informatics, ICHI; June 4-7, 2018; New York City, NY, USA p. A. [doi: 10.1109/ichi.2018.00092]
48. Moon Y, Moon HS, Kim D, Kim J, Lee J, Shim W, et al. Deep Learning-Based Stroke Volume Estimation Outperforms Conventional Arterial Contour Method in Patients with Hemodynamic Instability. J Clin Med 2019 Sep 09;8(9):1419 [FREE Full text] [doi: 10.3390/jcm8091419] [Medline: 31505848]
49. Slagt C, Malagon I, Groeneveld ABJ. Systematic review of uncalibrated arterial pressure waveform analysis to determine cardiac output and stroke volume variation. Br J Anaesth 2014 Apr;112(4):626-637 [FREE Full text] [doi: 10.1093/bja/aet429] [Medline: 24431387]
50. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis 2019 Oct 11;128(2):336-359. [doi: 10.1007/s11263-019-01228-7]
51. Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. : ACM; 2016 Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA, USA p. 1135. [doi: 10.1145/2939672.2939778]

Abbreviations

APCO: arterial pressure-based cardiac output
CNN: convolutional neural network
CO: cardiac output
HR: heart rate
LOWESS: locally weighted scatterplot smoother
PAC: pulmonary artery catheter
SV: stroke volume
SVR: systemic vascular resistance

Edited by C Lovis; submitted 04.10.20; peer-reviewed by D Pförringer, J Kim, L Rusu; comments to author 17.12.20; revised version received 10.02.21; accepted 17.06.21; published 16.08.21.

Please cite as:
Yang HL, Jung CW, Yang SM, Kim MS, Shim S, Lee KH, Lee HC
Development and Validation of an Arterial Pressure-Based Cardiac Output Algorithm Using a Convolutional Neural Network: Retrospective Study Based on Prospective Registry Data
JMIR Med Inform 2021;9(8):e24762
URL: https://medinform.jmir.org/2021/8/e24762
doi:10.2196/24762
PMID:34398790

XSL•FO
RenderX

©Hyun-Lim Yang, Chul-Woo Jung, Seong Mi Yang, Min-Soo Kim, Sungho Shim, Kook Hyun Lee, Hyung-Chul Lee. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 16.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using a Convolutional Neural Network to Predict Remission of Diabetes After Gastric Bypass Surgery: Machine Learning Study From the Scandinavian Obesity Surgery Register

Yang Cao[1,2], PhD; Ingmar Näslund[3], MD, PhD; Erik Näslund[4], MD, PhD; Johan Ottosson[3], MD, PhD; Scott Montgomery[1,5,6], PhD; Erik Stenberg[3], MD, PhD

[1]Clinical Epidemiology and Biostatistics, School of Medical Sciences, Örebro University, Örebro, Sweden
[2]Unit of Integrative Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
[3]Department of Surgery, Faculty of Medicine and Health, Örebro University, Örebro, Sweden
[4]Division of Surgery, Department of Clinical Sciences, Danderyd Hospital, Karolinska Institutet, Stockholm, Sweden
[5]Clinical Epidemiology Division, Department of Medicine, Karolinska Institutet, Stockholm, Sweden
[6]Department of Epidemiology and Public Health, University College London, London, United Kingdom

Corresponding Author:
Yang Cao, PhD
Clinical Epidemiology and Biostatistics
School of Medical Sciences
Örebro University
X-Huset, Södra Grev Rosengatan 1
Örebro, 70182
Sweden
Phone: 46 196 026 236
Fax: 46 196 026 236
Email: yang.cao@oru.se

Abstract

Background: Prediction of diabetes remission is an important topic in the evaluation of patients with type 2 diabetes (T2D) before bariatric surgery. Several high-quality predictive indices are available, but artificial intelligence algorithms offer the potential for higher predictive capability.

Objective: This study aimed to construct and validate an artificial intelligence prediction model for diabetes remission after Roux-en-Y gastric bypass surgery.

Methods: Patients who underwent surgery from 2007 to 2017 were included in the study, with collection of individual data from the Scandinavian Obesity Surgery Registry (SOReg), the Swedish National Patients Register, the Swedish Prescribed Drugs Register, and Statistics Sweden. A 7-layer convolution neural network (CNN) model was developed using 80% (6446/8057) of patients randomly selected from SOReg and 20% (1611/8057) of patients for external testing. The predictive capability of the CNN model and currently used scores (DiaRem, Ad-DiaRem, DiaBetter, and individualized metabolic surgery) were compared.

Results: In total, 8057 patients with T2D were included in the study. At 2 years after surgery, 77.09% achieved pharmacological remission (n=6211), while 63.07% (4004/6348) achieved complete remission. The CNN model showed high accuracy for cessation of antidiabetic drugs and complete remission of T2D after gastric bypass surgery. The area under the receiver operating characteristic curve (AUC) for the CNN model for pharmacological remission was 0.85 (95% CI 0.83-0.86) during validation and 0.83 for the final test, which was 9%-12% better than the traditional predictive indices. The AUC for complete remission was 0.83 (95% CI 0.81-0.85) during validation and 0.82 for the final test, which was 9%-11% better than the traditional predictive indices.

Conclusions: The CNN method had better predictive capability compared to traditional indices for diabetes remission. However, further validation is needed in other countries to evaluate its external generalizability.

(JMIR Med Inform 2021;9(8):e25612) doi:10.2196/25612

KEYWORDS

forecasting; clinical decision rules; remission induction; type 2 diabetes mellitus; gastric bypass; morbid obesity

XSL•FO
RenderX

Introduction

Bariatric surgery is an efficient and safe treatment for patients with morbid obesity and type 2 diabetes (T2D) [1,2]. In obese patients who also have T2D, more than three-fourths of patients show remission after gastric bypass surgery [3,4]. Although remission rates may differ across different surgical procedures, high remission rates have been reported for Roux-en-Y gastric bypass [1,3]. Despite many patients experiencing remission of diabetes, duration and severity of disease, along with age, have been presented as factors associated with reduced chance of achieving remission [1,5]. Prediction of diabetes remission can be helpful in clinical preoperative consultation and decision-making, and several indices have been constructed for this purpose. Scores like DiaRem [6], Ad-DiaRem [7], DiaBetter [8], and the individualized metabolic surgery (IMS) score [9], as well as the age, body mass index, C-peptide level, and duration of T2D (ABCD) score [10] have been used for predicting diabetes remission after bariatric surgery. Many of the models based on the scores have high predictive capability and may already provide clinical guidance [11]. These tools might be helpful for personalized management of morbidly obese individuals with diabetes when considering bariatric surgery in routine care, ultimately contributing to precision medicine [12]. However, the performance of the scores in various studies is not consistent [7]. Previous prediction models were either limited by small sample sizes or were not validated using external data that were not seen by the models during model construction. Therefore, both the performance and validity of the models or scores need to be further evaluated and improved using a larger bariatric surgery database. In recent years, there have been a number of attempts to use artificial intelligence (AI) algorithms, including support vector machine [13], decision tree [14], random forest [15], and deep learning algorithms, such as artificial neural networks [16,17], to incorporate preoperative predictors in predicting outcomes of bariatric surgery. Compared with the traditional statistical regression models, AI algorithms have shown great promise in the field of bariatric surgery [18,19]. However, to our knowledge, none have thus far reached clinical practice.

The aim of this study was to construct a prediction model for T2D remission using a deep learning AI algorithm (ie, convolutional neural network [CNN)]) and to compare its predictive capability with that of 4 widely used predictive scores.

Methods

Study Participants

The study used the data from the Scandinavian Obesity Surgery Register (SOReg), a validated, national quality register covering virtually all bariatric and metabolic surgical procedures in Sweden [20]. By using the unique Swedish personal identification number, we linked SOReg to the Swedish National Patient Register, the Swedish National Death Register, the Swedish Prescribed Drug Register, and Statistics Sweden to obtain information on inpatient and outpatient hospital visits, mortality, dispensed drugs, and individual socioeconomic data. The inclusion criteria for patients registered in the SOReg were included those operated on with a primary Roux-en-Y gastric bypass procedure between 2007 and 2015 and those diagnosed with T2D preoperatively, as defined by the American Diabetes Association (ie, fasting plasma glucose ≥ 126 mg/L [7.0 mmol/L], hemoglobin A1c [HbA_{1c}] ≥ 48 mmol/mol [6.5%], or pharmacological treatment for diabetes) [21].

Outcome and Predictor Variables

The main outcome measure was complete remission of diabetes 2 years after surgery, defined as being without diabetes medication within a time frame of +/- 6 months; that is, 18-30 months postoperatively with normal HbA_{1c} value <42 mmol/mol (6.0%) in accordance with the definitions of the American Diabetes Association [22]. Due to loss of information of HbA_{1c} at follow-up, analyses of a secondary outcome, complete remission, defined as discontinuance of pharmacological treatment from 18-30 months, was performed.

The predictor variables were patients' demographic and socioeconomic information including age, sex, education level (primary, secondary, higher education <3 years, and high education ≥3 years), and region of residence characteristics (large city, medium city or town, and small town or rural area); preoperative BMI, HbA_{1c}, and treatment information including insulin treatment, metformin use, other noninsulin pharmacological treatment, and number of antidiabetic drugs; and preoperative comorbidities including sleep apnea, hypertension, dyslipidemia, depression, and cardiovascular comorbidity.

Descriptive Analysis

Continuous variables are presented as mean and SD, and ordered and nominal variables are presented as median and interquartile range (IQR) and count and percentage, respectively. For comparison between 2 groups, the *t* test and Mann-Whitney test were used for continuous and ordered variables, respectively, while the Pearson chi-square test was used for categorical variables. A 2-tailed *P* value <.05 was considered to be statistically significant.

Multiple Imputation for Missing Values

Missing values were assumed missing at random and imputed using a random forest algorithm, which has the desirable properties of being able to handle mixed types of missing data, being adaptive to interactions and nonlinearity, and having the potential to scale to big data settings [23]. To allow for the uncertainty of the imputation, 100 imputed data sets were generated in the current study.

Data Normalization

Because the range of values of variables varies widely (such as for age and BMI) in some machine learning (ML) algorithms, objective function will not work properly [24]. Therefore, the continuous and ordered variables were normalized to have a mean of 0 and a standardization of 1, and the multicategory nominal variables (education and residence) were converted into several binary variables before they were entered into the ML models [25].

XSL•FO
RenderX

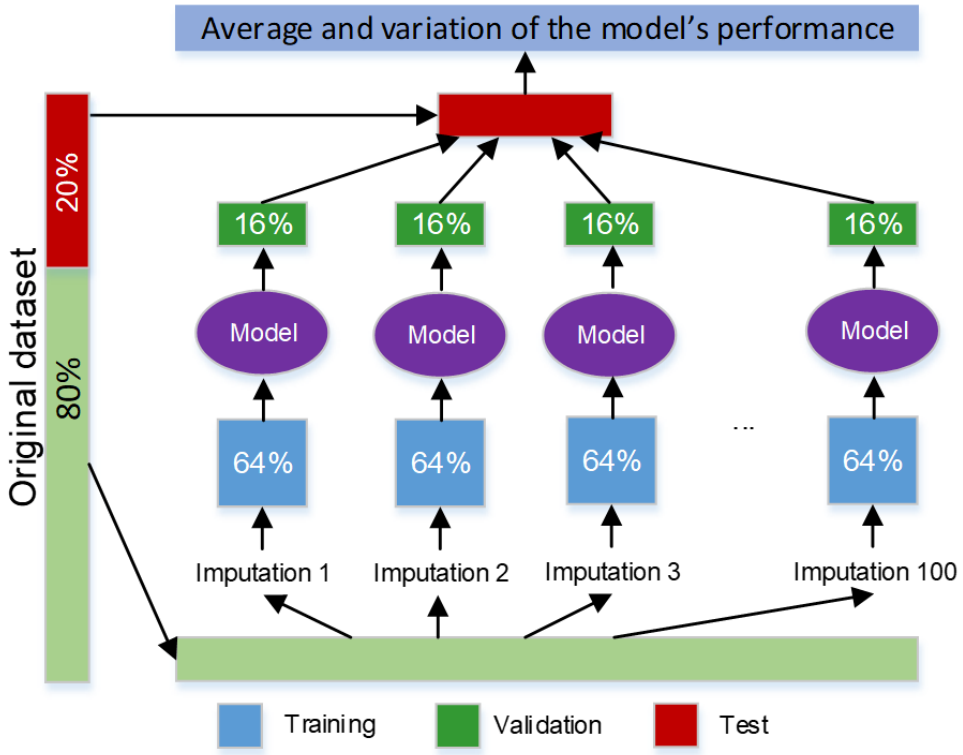
Predictive Model

In the current study, we used a 7-layer CNN model with two 1D convolution layers (with 100 filters for each), two 1D max pooling layers, one flatten layer, and two dense layers (with 1000 computation units) [26,27]. The rectified linear unit activation function was used for the two 1D convolution layers and the first dense layers, and the sigmoid activation function was used for the last dense layer. The binary cross-entropy loss function and the adaptive moment estimation (Adam) optimizer were used when compiling the model [28].

Model Training, Validation, and Test

The whole data set was randomly split into 2 parts: a training data set with 80% (6446/8057) of the patients and a test data set with 20% (1611/8057) of the patients. During the model training stage, the training data set was further divided into 2 data sets: one data set with 64% (5156/8057) of the patients to train the CNN model and another with 16.01% (1290/8057) of the patients to validate the model. Finally, the model was tested using the test data set that was never seen by the CNN model. The CNN model was trained, validated, and tested with the 100 imputations (Figure 1).

Figure 1. Procedure for training, validation, and testing for the convolutional neural network model.



Indices of Predictive Ability

Predictive ability of the CNN model was evaluated using the following indices: area under the receiver operating characteristic (ROC) curve, sensitivity, specificity, and the Youden J [29]. The terminology and derivations of the values have been previously presented in detail [18]. The sensitivity and specificity presented in this study are the values on the ROC curve where the Youden J achieves the maximum value. The acceptable, excellent, and outstanding predictive models were defined as those with an area under the ROC curve (AUC) greater than 0.7, 0.8, and 0.9, respectively [30,31]. The average and the 95% CI of the indices were calculated based on 100 imputations.

Comparison Between the CNN Model and DiaRem, Ad-DiaRem, DiaBetter, and IMS

We also evaluated the predictive capability of the currently used indices, DiaRem, Ad-DiaRem, DiaBetter, and IMS, and compared them with the CNN model. The DiaRem score is calculated using insulin use, age, HbA_{1c} value, and type of antidiabetic drugs [32]. The Ad-DiaRem score is a modification of the DiaRem score, calculated using insulin use, age, HbA_{1c} value, number of antidiabetic drugs, duration of diabetes, and number of antidiabetic drugs [13]. The DiaBetter is calculated using HbA_{1c}, type of antidiabetic drugs, and duration of diabetes [8]. The IMS score is calculated using the number of preoperative diabetes medications, insulin use, duration of diabetes, and HbA_{1c} level [9].

The points on the nonparametric ROC curve of DiaRem, Ad-DiaRem, DiaBetter, and IMS were generated using each value as a classification cutoff point and computing the corresponding sensitivity and one minus specificity. These points were then connected by straight lines, and the AUC was computed using the trapezoidal rule [33].

The same training and testing procedure used for the CNN model was also applied for the 4 scores.

Software and Hardware

The descriptive analysis and evaluation for DiaRem, Ad-DiaRem, DiaBetter, and IMS were conducted in Stata 16.1

XSL•FO
RenderX

(StataCorp LLC). The CNN model was achieved in Python 3.6 (Python Software Foundation) using the Keras 2.4.0 and Scikit-learn 0.23 packages. All the computation was operated on a computer with 64-bit Windows 7 Enterprise operating system (Service Pack 1, Microsoft Corporation), an Intel Core TM i5-4210U 2.40-GHz CPU, and 16.0 GB installed random access memory.

Ethics

The study was approved by the regional ethics committee in Stockholm (reference #2013/535-31/5, #2014/1639-32, and #2017/857-32). The study was conducted according to the guidelines of the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement [34].

Results

Patient Characteristics

In total, 8112 patients met the inclusion criteria; after exclusion of 55 patients who died within the first 2 years after surgery, 8057 patients remained in the analysis. Information on pharmaceutical usage before and after surgery was available for all patients. A postoperative weight was registered for 7268 patients at 1 year after surgery (90.21%), and 4996 patients at 2 years after surgery (62.01%). A postoperative glycosylated HbA_{1c} test result was available for 6989 patients (86.74%). Baseline characteristics of the included patients are shown in Table 1. Statistically significant differences were found for almost all the predictor variables between the remission patients and nonremission patients, except for depression and education (Table 1), which implies the potential for using the predictor variables to predict outcome. Preoperative HbA_{1c} values were missing for about one-seventh of the patients, indicating the need for imputation since the predictive capability otherwise would be significantly reduced and biased by excluding a considerable proportion of the data with missing values. Patients with a missing HbA_{1c} value were more often males of marginally higher age and longer duration of disease, and small differences were also seen in terms of pharmacological treatment, education, and residence (Supplementary Table S1, Multimedia Appendix 1). After multiple imputation, similar distributions of HbA_{1c} values were seen (Supplementary Figure S1, Multimedia Appendix 1).

XSL•FO
RenderX

Table 1. Characteristics of study participants with further stratification on remission of diabetes (N=8057)[a].

Characteristic	Overall (n=8057)	Nonremission (n=1846)	Remission (n=6211)	*P* value[b]
Age (years), mean (SD)	47.7 (10.1)	51.7 (8.7)	46.6 (10.2)	<.001
Sex, n (%)				.001
Women	4970 (61.68)	1079 (58.45)	3891 (62.65)	
Men	3087 (38.32)	767 (41.55)	2320 (37.35)	
BMI (kg/m^2), mean (SD)	42.22 (5.74)	41.16 (5.44)	42.53 (5.80)	<.001
Hemoglobin A_{1c} (mmol/mol) mean, (SD)	59.0 (17.3)	67.4 (17.5)	56.7 (16.5)	<.001
Diabetes duration (years), median (IQR)	2.0 (0.0-6.0)	6.0 (3.0-10.0)	1.0 (0.0-4.0)	<.001
Number of drugs, median (IQR)	1.0 (1.0-2.0)	2.0 (1.0-2.0)	1.0 (0.0-2.0)	<.001
Insulin, n (%)	2313 (28.71)	1184 (64.14)	1129 (18.18)	<.001
Metformin, n (%)	5610 (69.63)	1618 (87.65)	3992 (64.27)	<.001
Other noninsulin treatment, n (%)	1912 (23.73)	745 (40.36)	1167 (18.79)	<.001
Sleep apnea, n (%)	1529 (18.98)	383 (20.75)	1146 (18.45)	.03
Hypertension, n (%)	4546 (56.42)	1287 (69.72)	3259 (52.47)	<.001
Cardiovascular comorbidity, n (%)	917 (11.38)	305 (16.52)	612 (9.85)	<.001
Dyslipidemia, n (%)	2527 (31.36)	864 (46.80)	1663 (26.78)	<.001
Depression, n (%)	1297 (16.10)	311 (16.85)	986 (15.88)	.34
Education, n (%)				.40
Elementary education	1606 (19.93)	392 (21.24)	1214 (19.55)	
Secondary education	4762 (59.10)	1091 (59.10)	3671 (59.10)	
Higher education <3 years	838 (10.40)	179 (9.70)	659 (10.61)	
Higher education >3 years	796 (9.88)	173 (9.35)	623 (10.03)	
Residence, n (%)				.001
Large city	2734 (33.93)	687 (37.22)	2047 (32.96)	
Medium-sized town	3061 (37.99	671 (36.35)	2390 (38.48)	
Small town or rural area	2231 (27.69)	487 (26.38)	1744 (28.08)	
DiaRem, median (IQR)	6.0 (3.0-13.0)	16.0 (8.0-18.0)	5.0 (3.0-8.0)	<.001
Ad-DiaRem, median (IQR)	7.0 (5.0-11.0)	12.0 (9.0-15.0)	7.00 (4.0-9.0)	<.001
DiaBetter, median (IQR)	3.0 (1.0-6.0)	7.0 (5.0-8.0)	3.0 (1.0-4.0)	<.001
IMS[c], median (IQR)	39.8 (16.0-75.2)	87.2 (59.9-107.2)	28.6 (16.0-57.8)	<.001

[a]Including all the baseline variables used in the study.

[b]*P* value comparing remission vs nonremission.

[c]IMS: individualized metabolic surgery.

Surgical Outcome

The mean BMI loss at 1 year after surgery was 12.2 kg/m^2 (SD 4.0 kg/m^2), with an excess BMI loss (100 × [initial BMI – postoperative BMI]/[initial BMI – 25] %) of 74.0% (SD 22.5%), and a total weight loss (100 × weight loss/preoperative weight%) of 28.7% (SD 7.6%). Mean BMI loss at 2 years after surgery was 12.0 kg/m^2 (SD 4.53 kg/m^2) with an excess BMI loss of 73.3% (SD 24.4%) and a total weight loss of 28.4% (SD 8.9%).

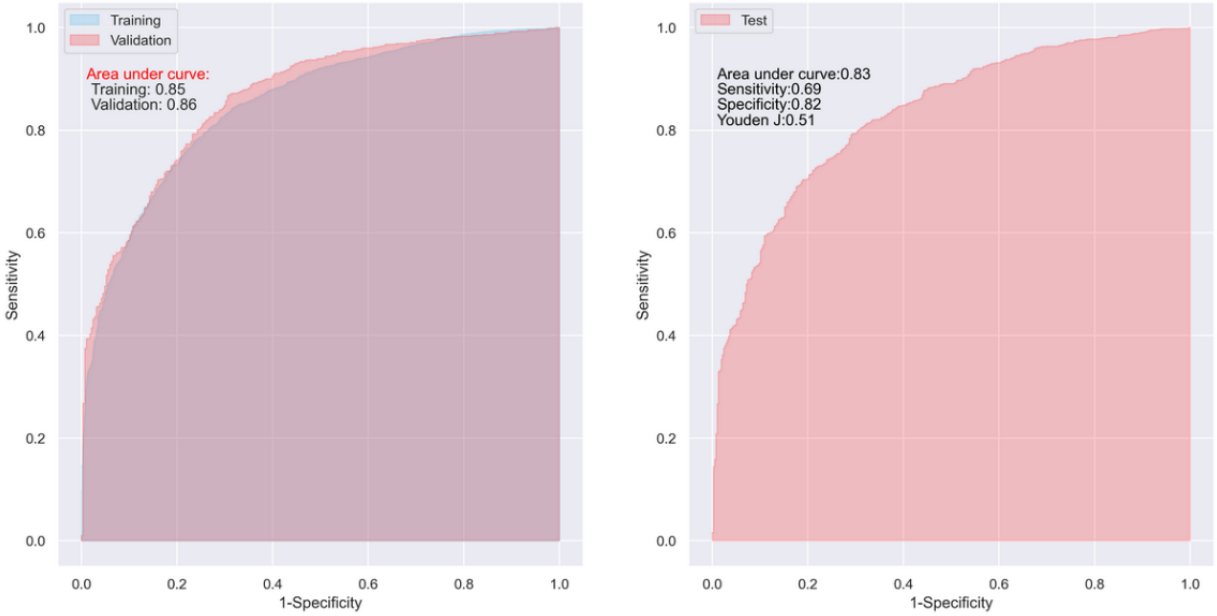
At 2 years after surgery, 77.09% (6211/8057) of the patients were able to discontinue the pharmacological treatment of T2D, while complete T2D remission was seen in 63.07% (n=4004) of the 6348 patients who had been evaluated for complete remission.

Predictive Capability of the CNN Model, DiaRem, Ad-DiaRem, DiaBetter, and IMS

The predictive capability of the CNN model for the major outcome (remission) is shown in Figure 2 and Table 2. In both the training and validation, the CNN model presented good predictive ability, with an AUC of 0.86 (95% CI 0.85-0.87) and 0.85 (95% CI 0.83-0.86), respectively (Table 2).

XSL•FO
RenderX

Figure 2. Receiver operating characteristic (ROC) curves of the convolutional neural network model in one of the 100 trainings and validations (left; because the 2 areas under the ROC curves are almost totally overlapping, the blended red and blue colors appear purple), and tests (right).



The DiaRem, Ad-DiaRem, DiaBetter, and IMS also showed good predictive capability in the training with an AUC >0.8 (Figure 3 left and Table 2) but only acceptable predictive ability in the validation (Table 2), with an AUC of 0.73 (95% CI 0.71-0.75), 0.72 (95% CI 0.69-0.74), 0.75 (95% CI 0.72-0.78), and 0.76 (95% CI 0.73-0.79), respectively. In general, the predictive capability of the CNN model was 16.4%, 18.1%, 13.3%, and 11.8% higher than that of DiaRem, Ad-DiaRem, DiaBetter, and IMS, in terms of AUC, respectively. In the tests, the AUC for the predictive ability of the CNN (AUC=0.83; 95% CI 0.82-0.85) model was 10.6%, 12.2%, 12.2%, and 9.2% higher than that of DiaRem (AUC=0.75; 95% CI 0.73-0.76), Ad-DiaRem (AUC=0.74; 95% CI 0.71-0.77), DiaBetter (AUC=0.74; 95% CI 0.72-0.76), and IMS (AUC=0.76; 95% CI 0.73-0.78), respectively (Figure 2 right and Figure 3 right).

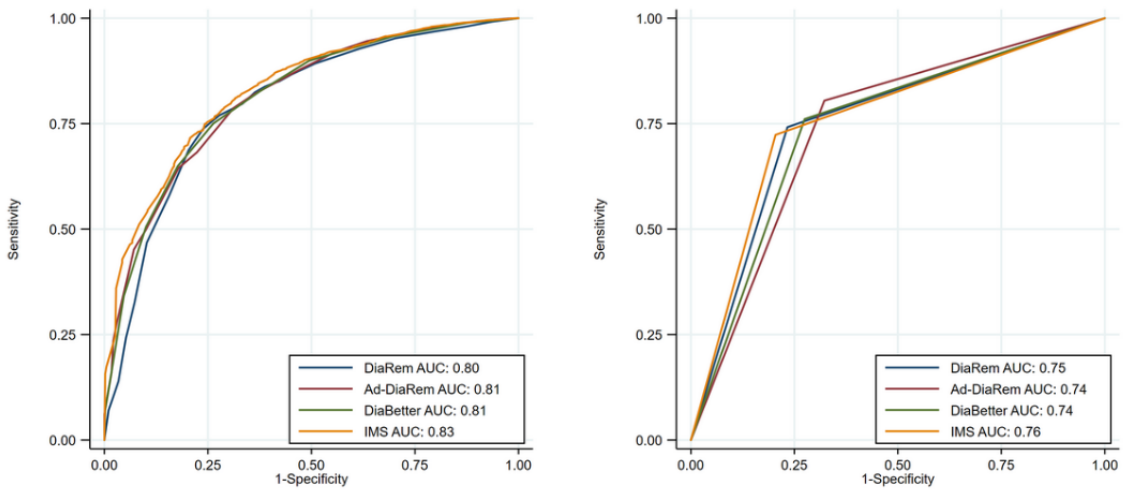
XSL•FO
RenderX

Table 2. Predictive capability of the CNN model and diabetes indices for the major outcome.

Models by index	Value (95% CI)	
	Training	Validation
AUC[a]		
CNN[b]	0.86 (0.85-0.87)	0.85 (0.83-0.86)
DiaRem	0.81 (0.79-0.82)	0.73 (0.71-0.75)
Ad-DiaRem	0.82 (0.81-0.83)	0.72 (0.69-0.74)
DiaBetter	0.82 (0.81-0.83)	0.75 (0.72-0.78)
IMS[c]	0.84 (0.83-0.85)	0.76 (0.73-0.79)
Specificity		
CNN	0.78 (0.74-0.83)	0.78 (0.72-0.85)
DiaRem	0.76 (0.80-0.73)	0.81 (0.78-0.85)
Ad-DiaRem	0.70 (0.68-0.71)	0.75 (0.70-0.79)
DiaBetter	0.76 (0.74-0.78)	0.76 (0.71-0.80)
IMS	0.77 (0.72-0.82)	0.77 (0.72-0.81)
Sensitivity		
CNN	0.77 (0.73-0.82)	0.76 (0.70-0.83)
DiaRem	0.75 (0.71,0.78)	0.65 (0.62-0.67)
Ad-DiaRem	0.79 (0.78-0.80)	0.69 (0.67-0.72)
DiaBetter	0.75 (0.74-0.76)	0.75 (0.72-0.78)
IMS	0.75 (0.70-0.80)	0.75 (0.73-0.77)
Youden J		
CNN	0.56 (0.54-0.57)	0.54 (0.50-0.59)
DiaRem	0.51 (0.50-0.52)	0.46 (0.42-0.50)
Ad-DiaRem	0.48 (0.47-0.49)	0.44 (0.39-0.49)
DiaBetter	0.51 (.049-0.54)	0.51 (0.45-0.56)
IMS	0.52 (0.50-0.54)	0.52 (0.47-0.57)

[a]AUC: area under the receiver operating characteristic curve.
[b]CNN: convolutional neural network.
[c]IMS: individualized metabolic surgery.

Figure 3. Receiver operating characteristic curves of diabetes indices in one of the 100 trainings (left), and tests (right). AUC: area under the curve.



XSL•FO
RenderX

For the secondary outcome, complete remission, the CNN model also presented a good predictive capability in both the training and validation, with an AUC of 0.84 (95% CI 0.83-0.85) and 0.83 (95% CI 0.81-0.85), respectively (Supplementary Table S4, Multimedia Appendix 1). Although DiaRem, Ad-DiaRem, DiaBetter, and IMS showed good predictive ability in the training with an AUC ≥0.80, they only showed acceptable predictive ability in the validation with an AUC of 0.72 (95% CI 0.69-0.75), 0.72 (95% CI 0.69-0.74), 0.74 (95% CI 0.72-0.77), and 0.74 (95% CI 0.72-0.76), respectively (Supplementary Table S4, Multimedia Appendix 1). In general, the predictive capability of the CNN model was 15.3%, 15.3%, 12.2%, and 12.2% higher than that of DiaRem, Ad-DiaRem, DiaBetter, and IMS, in terms of AUC, respectively.

In the tests, the AUC for the predictive capability of the CNN model (AUC=0.82; 95% CI 0.81-0.83) was 9.3%, 10.8%, 10.8%, and 9.3% higher than that of DiaRem (AUC=0.75; 95% CI 0.73-0.78), Ad-DiaRem (AUC=0.74; 95% CI 0.73-0.75), DiaBetter (AUC=0.74; 95% CI 0.71-0.76), and IMS (AUC=0.75; 95% CI 0.73-0.77), respectively (Supplementary Figure S2 right and S3 right, Multimedia Appendix 1).

Discussion

Principal Findings

The CNN model evaluated in this study showed high accuracy for cessation of antidiabetic drugs and complete remission of T2D after gastric bypass surgery, providing 9%-12% better predictive indices compared to available scores.

The currently available and widely accepted predictive indices for diabetes remission, including DiaRem, Ad-DiaRem, DiaBetter, and IMS, were assessed in our study and are all simple and easily available to clinicians for clinical decision support. In addition, one other index, the ABCD score [35], also includes c-peptide. This laboratory measure additionally includes information of endogenous insulin production and could thus potentially further enhance the effectiveness of a prediction model. However, the ABCD score has not been shown to have higher predictive capacity compared to other available models, and it is highly possible that other measures of severity of T2D disease, such as duration of disease, HbA_{1c} value, and type and number of drugs, may provide the same or even better measures for a prediction model [11].

The main benefits of the CNN method, in comparison to the scores based on traditional statistical methods, lie in its ability to include a high number of variables and to learn over time. In contrast to available models designed to offer simple entry and calculations of the most important variables, it offers the ability to handle variables in a more complex way, also including variables of smaller impact. Furthermore, the model construction is not limited by the statistical assumptions and distribution of the data, which usually need to be fulfilled in the traditional regression methods. Exposing the AI to a higher quantity of real-world data also has the potential to further improve it with cumulative learning.

Implications

The use of AI or machine learning techniques in medical research and practice is currently an evolving field with great potential. Although the exact role of AI in this setting remains to be established, one potential area where the AI seems to outperform traditional techniques is indeed in the construction of prediction models for outcomes from surgical procedures [36]. Previous studies on the construction of prediction models for perioperative complications have reported discouraging results, mainly as a direct consequence of the complexity and diversity of causes for perioperative complications [18,27,37]. In contrast with safety outcomes, efficacy outcomes (in particular those of highly standardized surgical methods such as gastric bypass) may be more suited for adequate prediction models since the factors influencing long-term effects are less diverse. Remission of diabetes is one such outcome that is largely influenced by a few specific factors, making prediction models more easily available. The results of our study support the promising results from previous studies with smaller sample sizes using sparse support vector machine, decision tree, and artificial neural networks to predict diabetes remission after bariatric surgery [13,14,17].

Although our CNN model did not include postoperative weight loss, a factor known to be associated with higher remission and reduced relapse of diabetes [8], the model included measures of patient-specific characteristics, information on duration and severity of disease, and a few socioeconomic factors that all should be easily available at the time of consultation before surgery. Although it is likely that the model could have reached a higher precision if postoperative results (such as early weight loss or improvement in glucose homeostasis) were included, these measures are not available in the preoperative setting and their inclusion would therefore reduce the clinical usefulness of the model [1,5,38]. Age, duration of diabetes, preoperative HbA_{1c}, and diabetes medications are all known predictive factors [1,5]. In addition, the model identified sex, BMI, metabolic and cardiovascular comorbidities, and place of residence as factors influencing the chance of diabetes remission.

Although the disposition of adiposity and insulin resistance appears to affect men and women differently [39], differences between sexes may be highly influenced by other covarying factors, such as obesity-related comorbidities, BMI, and age [1]. Indeed, when adjustment is made for other factors, the influence of sex on outcome tends to shift [1]. The influence of BMI on remission rates is also controversial [40]. Patients with higher BMI may have a greater degree of insulin resistance and a higher expected total weight loss [41,42], and may thus benefit more from the favorable metabolic effect of bariatric surgery. However, the influence of BMI on remission can be related to several other factors of relevance for both diabetes remission and postoperative weight loss. Whether or not the influence of BMI is strictly weight dependent or not remains to be answered. Although no difference in remission dependent on educational level was seen, place of residence was associated with the chance of achieving diabetes remission. Residents of larger cities may experience higher life stress and represent a more diverse socioeconomic population [43]. Many socioeconomic factors (such as education, income, profession, and ethnicity)

XSL•FO
RenderX

have been reported to influence other efficacy outcomes, such as weight loss, which in turn may contribute to these differences [42].

Challenges and Limitations

In contrast to traditional regression models, we observed significant improvement with the continuous training process. When increasing amounts of data in the test data set were seen by the CNN model (or more data in the test data set leaked into the training data set), AUC, specificity, and sensitivity increased gradually and eventually approximated 1 (Supplementary Figure S5, Multimedia Appendix 1). From training with more available data and decorrelating data with methods such as principal component analysis, the predictive capability of the CNN model could be improved even further, at least in the Swedish context. To generalize the application of the CNN model, a multinational registration consortium of gastric bypass surgery patients would be needed for improved model training and validation. However, the capacity of memory is also a limitation of the CNN because it reduces the model's flexibility to incorporate the information from external unseen data, which results in overfitting to specific past data or underfitting to the new data and impedes generalization of the model [44]. Teaching neural networks to strategically forget is an important task in ML. This highlights one of the major challenge of ML techniques [45]. To fulfill this task, incorporating long short-term memory units into CNN networks has been attempted to process temporal sequences and reduce model parameters in human face and activity recognition, which has shown consistent superior performance and good generalization [46,47]. Furthermore, the methods of ML are less transparent and more complex than those of traditional regression models, making their exact nature more difficult to scrutinize [44]. In the absence of clear guidelines, we have—to the best of our ability—conducted and reported the study to match the requirements of the TRIPOD statement and suggested modifications [34]. The programming code of the study is available at the repository figshare website [48]. Furthermore, the study was only based on data from a single country. For full use of the model, external validation would also be needed in other parts of the world.

Finally, only Roux-en-Y gastric bypass procedures were included in the model. The effects of sleeve gastrectomy on diabetes remission may be expected to differ [40], and thus the model is presently only suited for gastric bypass surgery. Including other surgical methods in future development of the model would further improve generalizability.

Despite these limitations, the CNN model outperformed the currently available high-quality prediction models. It also demonstrated better predictive ability than that mentioned in a previous report on AI for diabetes remission [49]. The CNN model may therefore find a place in the preoperative setting for surgeons, bariatricians, or endocrinologists looking to quantify the probability of diabetes remission in their decision-making for bariatric surgery in a given patient. After further validation, the AI model could be made available on a webpage or as a mobile app to allow user-friendly and fully available use in the clinical context.

Conclusions

Our CNN-based ML model performed well in identifying morbidly obese patients with T2D who might benefit from Roux-en-Y gastric bypass surgery. We also demonstrated the model had better predictive capability compared with the current widely used 4 comprehensive indices for diabetes remission after gastric bypass surgery. Prospectively identifying this subset of patients using data available at the time of preoperative evaluation provides an opportune time window to intervene and prevent or reduce the risk of morbidity and mortality, and may potentially reduce the total cost of care. However, this model should be further validated in future research using external data in other countries before it is incorporated into clinical practice.

Conflicts of Interest

JO has received reimbursement for participating in the advisory board of Johnson & Johnson and Vifor PHarma. ES received reimbursement by Johnson & Johnson Medical for a lecture on a topic unrelated to the contents of the present work. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 437 KB - medinform_v9i8e25612_app1.docx]

References

1. Jans A, Näslund I, Ottosson J, Szabo E, Näslund E, Stenberg E. Duration of type 2 diabetes and remission rates after bariatric surgery in Sweden 2007-2015: A registry-based cohort study. PLoS Med 2019 Nov;16(11):e1002985 [FREE Full text] [doi: 10.1371/journal.pmed.1002985] [Medline: 31747392]
2. Sjöström L, Peltonen M, Jacobson P, Ahlin S, Andersson-Assarsson J, Anveden, et al. Association of bariatric surgery with long-term remission of type 2 diabetes and with microvascular and macrovascular complications. JAMA 2014 Jun 11;311(22):2297-2304. [doi: 10.1001/jama.2014.5988] [Medline: 24915261]
3. Buchwald H, Estok R, Fahrbach K, Banel D, Jensen MD, Pories WJ, et al. Weight and type 2 diabetes after bariatric surgery: systematic review and meta-analysis. Am J Med 2009 Mar;122(3):248-256.e5. [doi: 10.1016/j.amjmed.2008.09.041] [Medline: 19272486]

XSL•FO
RenderX

4. Hofsø D, Fatima F, Borgeraas H, Birkeland KI, Gulseth HL, Hertel JK, et al. Gastric bypass versus sleeve gastrectomy in patients with type 2 diabetes (Oseberg): a single-centre, triple-blind, randomised controlled trial. Lancet Diabetes Endocrinol 2019 Dec;7(12):912-924. [doi: 10.1016/S2213-8587(19)30344-4] [Medline: 31678062]
5. Arterburn DE, Bogart A, Sherwood NE, Sidney S, Coleman KJ, Haneuse S, et al. A multisite study of long-term remission and relapse of type 2 diabetes mellitus following gastric bypass. Obes Surg 2013 Jan;23(1):93-102 [FREE Full text] [doi: 10.1007/s11695-012-0802-1] [Medline: 23161525]
6. Craig Wood G, Horwitz D, Still CD, Mirshahi T, Benotti P, Parikh M, et al. Performance of the DiaRem score for predicting diabetes remission in two health systems following bariatric surgery procedures in Hispanic and non-Hispanic White patients. Obes Surg 2018 Jan;28(1):61-68 [FREE Full text] [doi: 10.1007/s11695-017-2799-y] [Medline: 28717860]
7. Dicker D, Golan R, Aron-Wisnewsky J, Zucker J, Sokolowska N, Comaneshter DS, et al. Prediction of long-term diabetes remission after RYGB, sleeve gastrectomy, and adjustable gastric banding using DiaRem and Advanced-DiaRem scores. Obes Surg 2019 Mar;29(3):796-804. [doi: 10.1007/s11695-018-3583-3] [Medline: 30467708]
8. Pucci A, Tymoszuk U, Cheung WH, Makaronidis JM, Scholes S, Tharakan G, et al. Type 2 diabetes remission 2 years post Roux-en-Y gastric bypass and sleeve gastrectomy: the role of the weight loss and comparison of DiaRem and DiaBetter scores. Diabet Med 2018 Mar;35(3):360-367 [FREE Full text] [doi: 10.1111/dme.13532] [Medline: 29055156]
9. Aminian A, Brethauer SA, Andalib A, Nowacki AS, Jimenez A, Corcelles R, et al. Individualized metabolic surgery score: procedure selection based on diabetes severity. Ann Surg 2017 Oct;266(4):650-657. [doi: 10.1097/SLA.0000000000002407] [Medline: 28742680]
10. Chen J, Hsu N, Lee W, Chen S, Ser K, Lee Y. Prediction of type 2 diabetes remission after metabolic surgery: a comparison of the individualized metabolic surgery score and the ABCD score. Surg Obes Relat Dis 2018 May;14(5):640-645. [doi: 10.1016/j.soard.2018.01.027] [Medline: 29526672]
11. Sjöholm K, Carlsson LMS, Taube M, le Roux CW, Svensson P, Peltonen M. Comparison of preoperative remission scores and diabetes duration alone as predictors of durable type 2 diabetes remission and risk of diabetes complications after bariatric surgery: a post hoc analysis of participants from the Swedish obese subjects study. Diabetes Care 2020 Nov;43(11):2804-2811. [doi: 10.2337/dc20-0157] [Medline: 32873586]
12. Koliaki C, Tzeravini E, Papachristoforou E, Severi I, El Deik E, Karaolia M, et al. Eligibility and awareness regarding metabolic surgery in patients with type 2 diabetes mellitus in the real-world clinical setting; estimate of possible diabetes remission. Front Endocrinol (Lausanne) 2020;11:383 [FREE Full text] [doi: 10.3389/fendo.2020.00383] [Medline: 32582036]
13. Aron-Wisnewsky J, Sokolovska N, Liu Y, Comaneshter DS, Vinker S, Pecht T, et al. The advanced-DiaRem score improves prediction of diabetes remission 1 year post-Roux-en-Y gastric bypass. Diabetologia 2017 Oct;60(10):1892-1902. [doi: 10.1007/s00125-017-4371-7] [Medline: 28733906]
14. Hayes MT, Hunt LA, Foo J, Tychinskaya Y, Stubbs RS. A model for predicting the resolution of type 2 diabetes in severely obese subjects following Roux-en Y gastric bypass surgery. Obes Surg 2011 Jul;21(7):910-916. [doi: 10.1007/s11695-011-0370-9] [Medline: 21336560]
15. Razzaghi T, Safro I, Ewing J, Sadrfaridpour E, Scott JD. Predictive models for bariatric surgery risks with imbalanced medical datasets. Ann Oper Res 2019 Feb 1;280(1-2):1-18. [doi: 10.1007/s10479-019-03156-8]
16. Thomas DM, Kuiper P, Zaveri H, Surve A, Cottam D. Neural networks to predict long-term bariatric surgery outcomes. Bariatric Times 2017;14(12):14-17.
17. Pedersen HK, Gudmundsdottir V, Pedersen MK, Brorsson C, Brunak S, Gupta R. Ranking factors involved in diabetes remission after bariatric surgery using machine-learning integrating clinical and genomic biomarkers. NPJ Genom Med 2016;1:16035 [FREE Full text] [doi: 10.1038/npjgenmed.2016.35] [Medline: 29263820]
18. Cao Y, Fang X, Ottosson J, Näslund E, Stenberg E. A comparative study of machine learning algorithms in predicting severe complications after bariatric surgery. J Clin Med 2019 May 12;8(5):668 [FREE Full text] [doi: 10.3390/jcm8050668] [Medline: 31083643]
19. Johnston SS, Morton JM, Kalsekar I, Ammann EM, Hsiao C, Reps J. Using machine learning applied to real-world healthcare data for predictive analytics: an applied example in bariatric surgery. Value Health 2019 May;22(5):580-586. [doi: 10.1016/j.jval.2019.01.011] [Medline: 31104738]
20. Hedenbro JL, Näslund E, Boman L, Lundegårdh G, Bylund A, Ekelund M, et al. Formation of the Scandinavian Obesity Surgery Registry, SOReg. Obes Surg 2015 Oct;25(10):1893-1900. [doi: 10.1007/s11695-015-1619-5] [Medline: 25703826]
21. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care 2014 Jan;37 Suppl 1:S81-S90. [doi: 10.2337/dc14-S081] [Medline: 24357215]
22. Buse JB, Caprio S, Cefalu WT, Ceriello A, Del Prato S, Inzucchi SE, et al. How do we define cure of diabetes? Diabetes Care 2009 Nov;32(11):2133-2135 [FREE Full text] [doi: 10.2337/dc09-9036] [Medline: 19875608]
23. Tang F, Ishwaran H. Random forest missing data algorithms. Stat Anal Data Min 2017 Dec;10(6):363-377 [FREE Full text] [doi: 10.1002/sam.11348] [Medline: 29403567]
24. Zheng A, Casari A. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. Sebastopol, CA: O'Reilly Media, Inc; 2018.
25. Lantz B. Machine Learning with R: Expert Techniques for Predictive Modeling. Birmingham, UK: Packt Publishing Ltd; 2019.

XSL•FO
RenderX

26. Ketkar N. Convolutional Neural Networks. Deep Learning with Python. New York: Springer; 2017:63-78.
27. Cao Y, Montgomery S, Ottosson J, Näslund E, Stenberg E. Deep learning neural networks to predict serious complications after bariatric surgery: analysis of Scandinavian Obesity Surgery Registry data. JMIR Med Inform 2020 May 08;8(5):e15992 [FREE Full text] [doi: 10.2196/15992] [Medline: 32383681]
28. Kingma D, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv 2014:1-15.
29. Yin J, Tian L. Joint confidence region estimation for area under ROC curve and Youden index. Stat Med 2014 Mar 15;33(6):985-1000. [doi: 10.1002/sim.5992] [Medline: 24123069]
30. Marzban C. The ROC curve and the area under it as performance measures. Weather Forecast 2004;19(6):1106-1114. [doi: 10.1175/825.1]
31. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 2010 Sep;5(9):1315-1316 [FREE Full text] [doi: 10.1097/JTO.0b013e3181ec173d] [Medline: 20736804]
32. Still CD, Wood GC, Benotti P, Petrick AT, Gabrielsen J, Strodel WE, et al. Preoperative prediction of type 2 diabetes remission after Roux-en-Y gastric bypass surgery: a retrospective cohort study. Lancet Diabetes Endocrinol 2014 Jan;2(1):38-45 [FREE Full text] [doi: 10.1016/S2213-8587(13)70070-6] [Medline: 24579062]
33. StataCorp LLC. Stata Base Reference Manual Release. College Station, TX: Stata Press Publication; 2007.
34. Collins GS, Reitsma JB, Altman DG, Moons KGM, TRIPOD Group. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. Circulation 2015 Jan 13;131(2):211-219 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.114.014508] [Medline: 25561516]
35. Lee W, Hur KY, Lakadawala M, Kasama K, Wong SKH, Chen S, et al. Predicting success of metabolic surgery: age, body mass index, C-peptide, and duration score. Surg Obes Relat Dis 2013;9(3):379-384. [doi: 10.1016/j.soard.2012.07.015] [Medline: 22963817]
36. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. Ann Surg 2018 Jul;268(1):70-76 [FREE Full text] [doi: 10.1097/SLA.0000000000002693] [Medline: 29389679]
37. Geubbels N, de Brauw LM, Acherman YIZ, van de Laar AWJM, Bruin SC. Risk stratification models: how well do they predict adverse outcomes in a large Dutch bariatric cohort? Obes Surg 2015 Dec;25(12):2290-2301. [doi: 10.1007/s11695-015-1699-2] [Medline: 25937046]
38. Yan Y, Wang G, Xu N, Wang F. Correlation between postoperative weight loss and diabetes mellitus remission: a meta-analysis. Obes Surg 2014 Nov;24(11):1862-1869. [doi: 10.1007/s11695-014-1285-z] [Medline: 24831461]
39. Geer EB, Shen W. Gender differences in insulin resistance, body composition, and energy balance. Gend Med 2009;6 Suppl 1:60-75 [FREE Full text] [doi: 10.1016/j.genm.2009.02.002] [Medline: 19318219]
40. Rubino F, Nathan DM, Eckel RH, Schauer PR, Alberti KGMM, Zimmet PZ, Delegates of the 2nd Diabetes Surgery Summit. Metabolic surgery in the treatment algorithm for type 2 diabetes: a joint statement by international diabetes organizations. Diabetes Care 2016 Jun;39(6):861-877. [doi: 10.2337/dc16-0236] [Medline: 27222544]
41. Martinez KE, Tucker LA, Bailey BW, LeCheminant JD. Expanded normal weight obesity and insulin resistance in US adults of the National Health and Nutrition Examination Survey. J Diabetes Res 2017;2017:9502643 [FREE Full text] [doi: 10.1155/2017/9502643] [Medline: 28812029]
42. Stenberg E, Näslund I, Persson C, Szabo E, Sundbom M, Ottosson J, et al. The association between socioeconomic factors and weight loss 5 years after gastric bypass surgery. Int J Obes (Lond) 2020 Nov;44(11):2279-2290 [FREE Full text] [doi: 10.1038/s41366-020-0637-0] [Medline: 32651450]
43. Zarzycka D, Ślusarska B, Marcinowicz L, Wrońska I, Kózka M. Assessment of differences in psychosocial resources and state of health of rural and urban residents--based on studies carried out on students during examination stress. Ann Agric Environ Med 2014;21(4):882-887 [FREE Full text] [doi: 10.5604/12321966.1129952] [Medline: 25528939]
44. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019 Apr 20;393(10181):1577-1579. [doi: 10.1016/S0140-6736(19)30037-6] [Medline: 31007185]
45. Bourtoule L, Chandrasekaran V, Choquette-Choo C, Jia H, Travers A, Zhang B. Machine unlearning. arXiv preprint arXiv 2019:191203817. [doi: 10.1090/mbk/121/79]
46. Xu Z, Li S, Deng W. Learning temporal features using LSTM-CNN architecture for face anti-spoofing. 2015 Presented at: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR); 3-6 Nov; Kuala Lumpur, Malaysia. [doi: 10.1109/acpr.2015.7486482]
47. Xia K, Huang J, Wang H. LSTM-CNN architecture for human activity recognition. IEEE Access 2020;8:56855-56866. [doi: 10.1109/access.2020.2982225]
48. Cao Y, Näslund I, Näslund E, Ottosson J, Montgomery S, Stenberg E. Python code for: Using convolutional neural network to predict remission of diabetes after gastric bypass surgery – a machine learning study from the Scandinavian Obesity Surgery Register 2020. figshare. URL: https://doi.org/10.6084/m9.figshare.13078943.v1 [accessed 2021-08-06]
49. Johnston SS, Morton JM, Kalsekar I, Ammann EM, Hsiao C, Reps J. Using machine learning applied to real-world healthcare data for predictive analytics: an applied example in bariatric surgery. Value Health 2019 May;22(5):580-586. [doi: 10.1016/j.jval.2019.01.011] [Medline: 31104738]

XSL•FO
RenderX

Abbreviations

ABCD: age, body mass index, C-peptide level, and duration of type 2 diabetes
AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
CNN: convolutional neural network
HbA_{1c}: hemoglobin A_{1c}
IMS: individualized metabolic surgery
ML: machine learning
SOReg: Scandinavian Obesity Surgery Registry
T2D: type 2 diabetes
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by C Lovis; submitted 10.11.20; peer-reviewed by J Hjelmesæth, R Krukowski; comments to author 24.11.20; revised version received 07.12.20; accepted 10.07.21; published 19.08.21.

Please cite as:
Cao Y, Näslund I, Näslund E, Ottosson J, Montgomery S, Stenberg E
Using a Convolutional Neural Network to Predict Remission of Diabetes After Gastric Bypass Surgery: Machine Learning Study From the Scandinavian Obesity Surgery Register
JMIR Med Inform 2021;9(8):e25612
URL: https://medinform.jmir.org/2021/8/e25612
doi:10.2196/25612
PMID:34420921

©Yang Cao, Ingmar Näslund, Erik Näslund, Johan Ottosson, Scott Montgomery, Erik Stenberg. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Patient-Level Cancer Prediction Models From a Nationwide Patient Cohort: Model Development and Validation

Eunsaem Lee[1*], MSc; Se Young Jung[2*], MD, MPH; Hyung Ju Hwang[1], PhD; Jaewoo Jung[3], PhD

[1]Department of Mathematics, Pohang University of Science and Technology, Pohang-si, Republic of Korea

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

[3]AMSquare Corporation, Pohang-si, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Hyung Ju Hwang, PhD
Department of Mathematics
Pohang University of Science and Technology
77 Cheongam-ro
Nam-gu
Pohang-si, 37673
Republic of Korea
Phone: 82 054 279 2056
Fax: 82 054 279 2799
Email: hjhwang@postech.ac.kr

Abstract

Background: Nationwide population-based cohorts provide a new opportunity to build automated risk prediction models at the patient level, and claim data are one of the more useful resources to this end. To avoid unnecessary diagnostic intervention after cancer screening tests, patient-level prediction models should be developed.

Objective: We aimed to develop cancer prediction models using nationwide claim databases with machine learning algorithms, which are explainable and easily applicable in real-world environments.

Methods: As source data, we used the Korean National Insurance System Database. Every Korean in ≥40 years old undergoes a national health checkup every 2 years. We gathered all variables from the database including demographic information, basic laboratory values, anthropometric values, and previous medical history. We applied conventional logistic regression methods, light gradient boosting methods, neural networks, survival analysis, and one-class embedding classifier methods to effectively analyze high dimension data based on deep learning–based anomaly detection. Performance was measured with area under the curve and area under precision recall curve. We validated our models externally with a health checkup database from a tertiary hospital.

Results: The one-class embedding classifier model received the highest area under the curve scores with values of 0.868, 0.849, 0.798, 0.746, 0.800, 0.749, and 0.790 for liver, lung, colorectal, pancreatic, gastric, breast, and cervical cancers, respectively. For area under precision recall curve, the light gradient boosting models had the highest score with values of 0.383, 0.401, 0.387, 0.300, 0.385, 0.357, and 0.296 for liver, lung, colorectal, pancreatic, gastric, breast, and cervical cancers, respectively.

Conclusions: Our results show that it is possible to easily develop applicable cancer prediction models with nationwide claim data using machine learning. The 7 models showed acceptable performances and explainability, and thus can be distributed easily in real-world environments.

(JMIR Med Inform 2021;9(8):e29807) doi:10.2196/29807

KEYWORDS

prediction; model; claim data; cancer; machine learning; development; cohort; validation; database; algorithm

XSL•FO
RenderX

Introduction

Cancer is a major cause of death, accounting for nearly 10 million deaths worldwide in 2020 [1]. It is a preventable disease requiring major lifestyle modifications [2], for which screening is important because it can help health care professionals with early detection and treatment of several types of cancer before they become aggravated [3]. In the early stages, cancer is normally indolent and symptomless. Thus, nationwide cancer screening programs for the general population have been adopted in many countries [4-8]. A national cancer control program (NCCP) framework, a public health program designed to mitigate the number of cancer cases and deaths and improve quality of life of patients, was proposed by the World Health Organization [6,9]. In South Korea, the NCCP was designed in 1996 and implemented in 1999 to provide free screening services for low-income Medical Aid patients. Beginning in 2000, the NCCP has expanded its target population to include all National Health Insurance (NHI) recipients. Since that time, the survival rate of cancer patients has continued to improve. According to cancer registration statistics in 2013, the relative survival rate of cancer patients has increased to 70.3% [10]. For 7 major cancer, namely, stomach, colorectal, breast, lung, cervical, pancreas, and liver cancer, every NHI beneficiary receives cancer screening tests mainly based on his or her age and gender. For instance, everyone ≥40 years old is examined by upper gastrointestinography or gastrointestinal endoscopy every 2 years to screen for stomach cancer. However, concerns have been raised about this one-size-fits-all cancer screening program because every procedure for cancer screening has its own risks for false-positive cases. For instance, false-positive cases of mammograms for screening breast cancer have resulted in many unnecessary invasive breast excisional biopsies, which reduce the quality of life in women [11,12]. Thus, personalized cancer screening protocols based on patient's individual risks have been in need since the NCCP was introduced [13,14]. The National Health Insurance System (NHIS) has collected health checkup data since 2003 under a structured data format and made it available for researchers [15]. There are two types of NHIS cohort data: a 1-million-person cohort sampled randomly from all NHI beneficiaries reflecting general characteristics of the entire South Korean population and a 500-thousand-person cohort sampled from those who received national health checkup services. All data include every diagnosis code and medications of each patient in all hospitals and clinics. For beneficiaries of national health checkup services, data include basic anthropometric measurements, laboratory values, past medical history, and family history. Despite the limited number of variables for the development of machine learning algorithms compared to electronic health records (EHRs) in hospitals, this type of data has the substantial advantages of a well-refined structured format and large sample size [16]. The data structure of the NHIS cohort and the monthly claim data from every EHR in hospitals are the same; therefore, the developed patient-level prediction models can be implemented in any EHR system in South Korea. In this study, we aimed to develop practical patient-level prediction models of 7 major cancers with acceptable performances and explainability, which can be distributed easily in real-world environments.

Methods

Data Description

We used an NHIS database to develop our cancer prediction models. The NHIS, a mandatory social insurance system, has collected health screening data at the national population level since the mid-1970s [15]. As this is a centralized system, Korean health screening data can be centralized, while paid health care providers act on a per-service basis [17]. The NHIS database consists of 2 different data sets: a health checkup cohort and a national sample cohort [18]. We used the health checkup cohort in the learning process and included training and internal validation and the remaining national sample cohort for external validation.

The NHIS provides a free health checkup program to all NHI members every 2 years. The health checkup cohort contains a total of 514,866 patients' health checkup records randomly extracted from health insurance members who have undergone a heath checkup program. The national sample cohort contains about 1 million patient records corresponding to about 2.2% of the Korean population in 2002. This data set was collected by considering demographics, such as population, age, and geographic factors. Both data sets include social and economic eligibility variables, health resource utilization status, description, treatment details, disease type, prescription details, and clinic status. The NHIS data set statistics are presented in Table 1.

XSL•FO
RenderX

Table 1. Statistics of the National Health Insurance Service data sets (2002-2013).

Description	Health checkup cohort, n	National sample cohort, n
Hospital	51,920	52,483
Patients	514,866	1,113,656
Prescriptions	83,935,395	83,935,395
Visits	96,534,359	119,362,188
Diagnostic codes (full code name)	17,385	19,626
Diagnostic codes (first 3 digits)	2160	2319
Annual patient visits, mean	15.6	8.9
Diagnostic codes/visit, mean	2.4	2.5
Drug/prescription, mean	4.4	4.4

Study Population Definition

It is mandatory that all cancer patients in South Korea be enrolled into a national cancer management program in the hospital where the cancer is diagnosed so that cancer patients only pay 5% of the total medical cost [19]. This means that almost all cancer patients in South Korea can be identified by diagnosis codes registered in the NHIS database [20].

We used the Korean Classification of Disease version 7, which is compatible with International Classification of Disease (ICD)-9 and defined the following 7 major cancers [21]: liver cancer (malignant neoplasm of the liver and intrahepatic bile ducts), C22; lung cancer (malignant neoplasm of the bronchus and lung), C34; colorectal cancer (malignant neoplasm of the colon, rectosigmoid junction, and rectum), C18, C19, and C20; pancreatic cancer (malignant neoplasm of the pancreas), C25; stomach cancer (malignant neoplasm of the stomach), C16; and breast cancer (malignant neoplasm of the breast), C50; and cervical cancer (malignant neoplasm of the cervix uteri), C53.

The prevalence of each cancer is presented in Table 2.

Table 2. The number of cancer-free patients and the number of cancer patients diagnosed for each cancer.

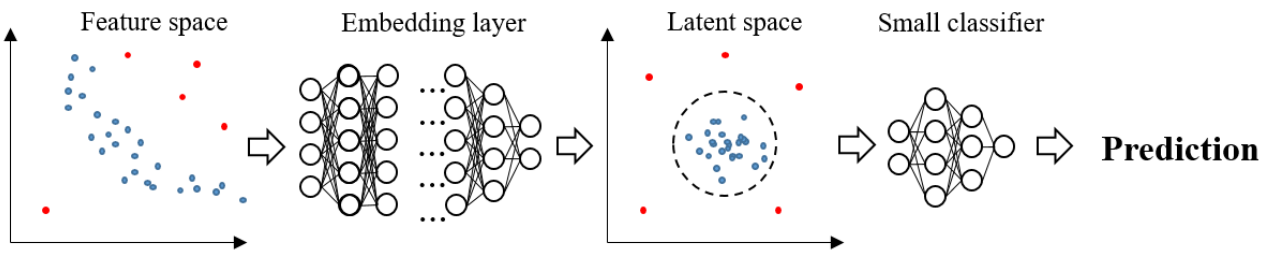
Patient type	Liver	Lung	Colorectal	Pancreatic	Stomach	Breast	Cervical
Free, n	234,659	233,931	233,203	235,633	232,493	91,982	92,736
Diagnosed, n	1587	2335	2845	551	3679	1029	306

Input Features and Algorithms

First, we used basic features consisting of simple demographic information, including age and gender, health examination, and survey results (18 features, level 1). Second, we added 11 more features obtained from a questionnaire, including the patient's medical history and family medical history (29 features, level 2). Third, we included 10 specific disease diagnostic records that appeared significant through univariate analysis for each cancer (39 features, level 3). The specific codes for each of the 10 cancers are provided in Multimedia Appendix 1.

To predict future cancers, we focused on cancer incidence within the next 5 years based on the time of screening. We first trained our predictive model with 4 common machine learning models: logistic regression (LR), random forest (RF), Light Gradient Boosting Machine (LGBM; a tree-based gradient boosting model), and multilayer perceptron (MLP). Further, we built a one-class embedding classifier (OCEC), which is a deep anomaly detection-based model (Figure 1). This method assumes that the data have one large class and several types of small anomalies not included in that class. This is an appropriate assumption because, while most people have normal screening records, few have cancer. To build our OCEC structure, we modified a deep one-class classification, the first deep learning–based anomaly detection model [22]. We then added a small classifier to the latent space to predict future cancer. The hyperparameters used for training models are shown in Multimedia Appendix 2.

Figure 1. Concept of one-class embedding classifier.



XSL•FO
RenderX

Model Evaluation Strategy

We divided an entire health checkup cohort, with 80% placed into a training set and 20% placed into a validation set. The model was trained only with the training set while the internal validation set was not used in the learning process. After training, the model output a prediction score for the probability of developing cancer in the next 5 years after the input year.

A cancer prediction problem is heavily imbalanced because the proportion of cancer-diagnosed patients is too small. In our data, the proportions of cancer-diagnosed patients were <2% for all 7 cancers. Thus, we used the area under the receiver operating characteristic curve (AUROC) and area under the precision recall curve (AUPRC) score to evaluate our models. The AUROC is an evaluation metric with values between 0 and 1 that is widely used as an evaluation metric for the imbalance problem, while the AUPRC combines recall and precision and corresponds to the average of the precision according to the precision recall curve. The baseline for AUROC is always 0.5, meaning a random classifier would produce an AUROC of 0.5. However, with AUPRC, the baseline is equal to the fraction of positive cancer cases (number of positive examples/total number of examples). The baseline AUPRC for each cancer in both the internal and external validation sets is shown in Table 3.

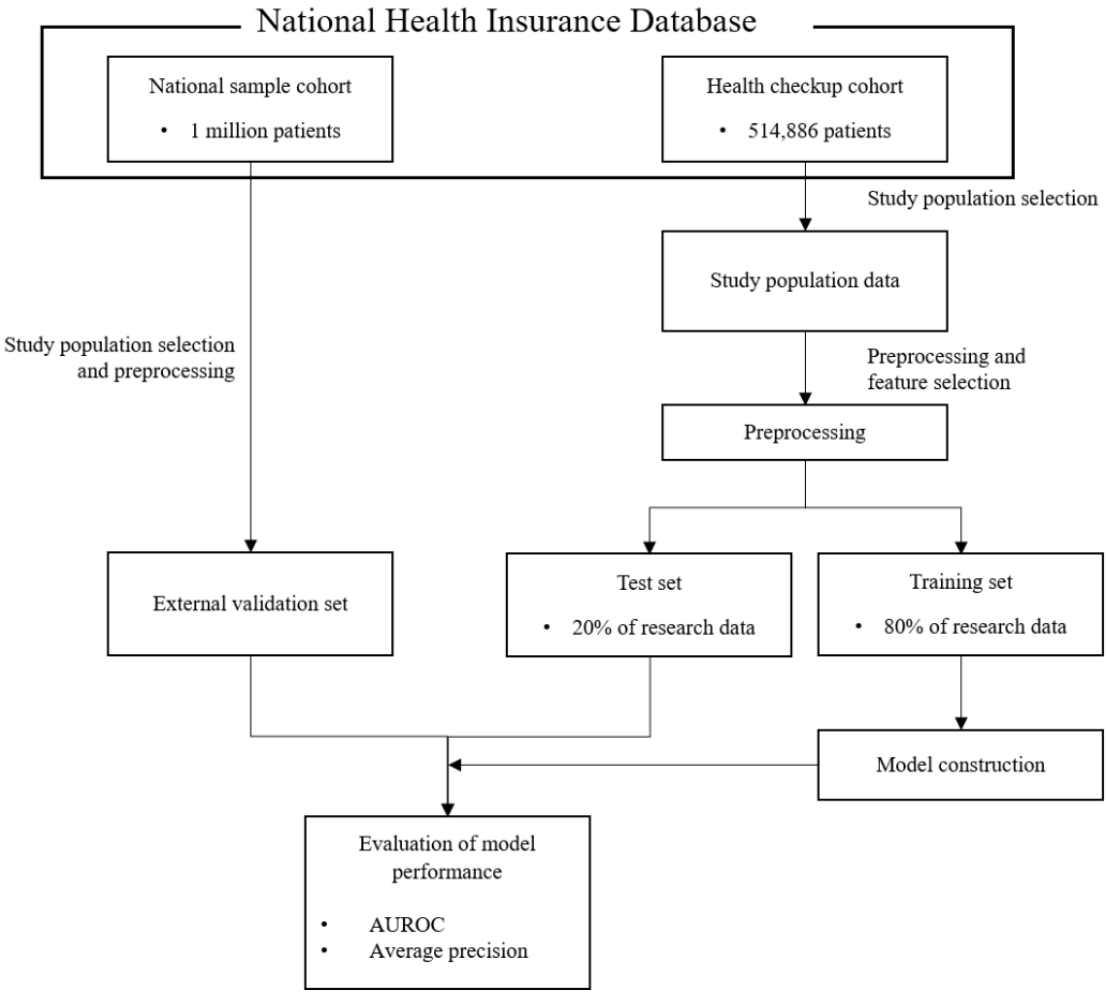
Table 3. The baseline area under the precision recall curve for the internal and external validation sets.

Validation set	Liver	Lung	Colorectal	Pancreatic	Stomach	Breast	Cervical
Internal validation	4.45×10^{-3}	6.03×10^{-3}	7.72×10^{-3}	1.50×10^{-3}	1.04×10^{-2}	7.65×10^{-3}	2.39×10^{-3}
External validation	2.96×10^{-3}	3.86×10^{-3}	5.52×10^{-3}	1.01×10^{-3}	6.65×10^{-3}	7.97×10^{-3}	2.22×10^{-3}

We evaluated the above metrics for both internal and external validation sets and compared the results. Additionally, for the external data set, we used the survival analysis method. We plotted Kaplan-Meier cumulative density curves to see the actual effectiveness of the predictive score. The study flow chart for learning and verification of the overall process is shown in Figure 2.

The NHIS institutional review board approved all data requests for research purposes (NHIS-2017-2-326). Because this public database is fully anonymized, institutional approval of Seoul National University Bundang Hospital (SNUBH) was waived by the institutional review board (X-2009-634-902).

Figure 2. Flow chart of the overall process. AUROC: area under the receiver operating characteristic curve.



XSL•FO
RenderX

Results

Performance of Cancer Prediction Models

Table 4 shows the internal validation results for each cancer across the 5 models. Overall, the LGBM and deep learning models performed better than did LR and RF. The former models performed well in terms of AUROC and AUPRC scores. LR, the most widely used classic model, showed low AUPRC scores, while RF had a low AUROC.

Notably, more than half of the OCEC AUROC scores were top rated compared to other models. Two models, OCEC and MLP, are both deep learning structured models. However, OCEC uses dense dimension reduction and performed better for both AUROC and AUPRC score compared to the MLP model. This shows that the anomaly-based one-class classification model can be a suitable deep learning structure for rare disease prediction.

When looking at the internal validation results of each cancer, liver and lung cancers showed the best results (AUROC>0.8), followed by stomach, pancreatic, and colorectal cancers (0.8>AUROC>0.7). Cervical and breast cancers (both female cancers) showed the lowest results (0.7>AUROC>0.6). The same findings also appeared in the external validation (Table 5).

According to feature level, the results tended to improve as feature level increased from level 1 to 3, but this was not significant. However, in some cases, the opposite tendency was observed.

The findings for the external validation score were similar to those of the internal score. Interestingly, the external validation scores (Table 5) were higher than the internal ones overall.

XSL•FO
RenderX

Table 4. Internal validation performance of outcome prediction across models.

Cancer type	Feature level	LGBM[a]		LR[b]		RF[c]		MLP[d]		OCEC[e]	
		AUROC[f]	AUPRC[g]	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Liver											
	Level 1	0.858	0.359	0.836	0.045	0.748	0.359	0.858	0.296	0.857	0.313
	Level 2	0.868	0.363	0.841	0.048	0.770	0.342	0.856	0.297	0.860	0.301
	Level 3	0.871	0.383	0.852	0.080	0.788	0.361	0.860	0.315	0.868	0.334
Lung											
	Level 1	0.845	0.396	0.823	0.106	0.735	0.366	0.845	0.360	0.849	0.382
	Level 2	0.845	0.395	0.822	0.110	0.750	0.366	0.832	0.338	0.841	0.338
	Level 3	0.845	0.401	0.829	0.130	0.754	0.367	0.841	0.345	0.843	0.343
Colorectal											
	Level 1	0.790	0.385	0.764	0.055	0.707	0.366	0.794	0.347	0.795	0.371
	Level 2	0.792	0.387	0.767	0.063	0.701	0.363	0.790	0.321	0.798	0.342
	Level 3	0.794	0.385	0.769	0.075	0.704	0.360	0.791	0.322	0.796	0.342
Pancreatic											
	Level 1	0.723	0.300	0.724	0.017	0.676	0.316	0.744	0.234	0.746	0.259
	Level 2	0.720	0.281	0.727	0.018	0.669	0.309	0.725	0.240	0.745	0.240
	Level 3	0.723	0.271	0.730	0.018	0.682	0.311	0.730	0.225	0.743	0.231
Stomach											
	Level 1	0.787	0.385	0.768	0.086	0.713	0.353	0.793	0.348	0.798	0.367
	Level 2	0.790	0.382	0.770	0.092	0.704	0.351	0.796	0.345	0.800	0.345
	Level 3	0.791	0.383	0.772	0.108	0.715	0.351	0.787	0.329	0.795	0.329
Breast											
	Level 1	0.684	0.344	0.689	0.077	0.666	0.343	0.705	0.325	0.713	0.332
	Level 2	0.696	0.345	0.696	0.083	0.681	0.346	0.706	0.324	0.711	0.327
	Level 3	0.722	0.357	0.733	0.129	0.689	0.353	0.734	0.339	0.749	0.345
Cervical											
	Level 1	0.647	0.268	0.667	0.013	0.656	0.273	0.671	0.263	0.690	0.265
	Level 2	0.672	0.271	0.669	0.012	0.632	0.274	0.660	0.266	0.670	0.266
	Level 3	0.653	0.296	0.612	0.027	0.679	0.301	0.638	0.275	0.645	0.279

[a]LGBM: Light Gradient Boosting Model.
[b]LR: logistic regression.
[c]RF: random forest.
[d]MLP: multilayer perceptron.
[e]OCEC: one-class embedding classifier.
[f]AUROC: area under receiver operator characteristics curve.
[g]AUPRC: area under precision recall curve.

XSL•FO
RenderX

Table 5. External performance of outcome prediction across models.

Cancer type	Feature level	LGBM[a]		LR[b]		RF[c]		MLP[d]		OCEC[e]	
		AUROC[f]	AUPRC[g]	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Liver											
	Level 1	0.910	0.485	0.893	0.065	0.815	0.502	0.911	0.433	0.912	0.442
	Level 2	0.909	0.485	0.895	0.067	0.826	0.488	0.900	0.391	0.911	0.433
	Level 3	0.915	0.514	0.907	0.120	0.838	0.527	0.910	0.463	0.919	0.471
Lung											
	Level 1	0.896	0.465	0.875	0.097	0.789	0.468	0.898	0.431	0.897	0.450
	Level 2	0.895	0.463	0.875	0.104	0.788	0.465	0.886	0.296	0.894	0.401
	Level 3	0.897	0.464	0.879	0.118	0.794	0.471	0.887	0.402	0.894	0.408
Colorectal											
	Level 1	0.872	0.455	0.858	0.070	0.776	0.482	0.883	0.426	0.887	0.449
	Level 2	0.874	0.453	0.858	0.076	0.780	0.481	0.874	0.394	0.887	0.423
	Level 3	0.877	0.455	0.859	0.085	0.776	0.473	0.882	0.393	0.884	0.415
Pancreatic											
	Level 1	0.891	0.420	0.884	0.029	0.753	0.456	0.898	0.360	0.904	0.336
	Level 2	0.888	0.405	0.884	0.030	0.747	0.450	0.883	0.335	0.902	0.337
	Level 3	0.885	0.407	0.886	0.039	0.759	0.450	0.883	0.323	0.897	0.336
Stomach											
	Level 1	0.889	0.481	0.863	0.088	0.795	0.478	0.891	0.457	0.894	0.440
	Level 2	0.891	0.480	0.864	0.095	0.793	0.479	0.887	0.422	0.893	0.436
	Level 3	0.889	0.478	0.864	0.109	0.792	0.473	0.885	0.401	0.890	0.413
Breast											
	Level 1	0.763	0.485	0.704	0.108	0.750	0.492	0.686	0.406	0.753	0.421
	Level 2	0.771	0.488	0.716	0.106	0.745	0.492	0.678	0.396	0.697	0.410
	Level 3	0.780	0.497	0.759	0.143	0.757	0.491	0.730	0.411	0.745	0.429
Cervical											
	Level 1	0.729	0.364	0.742	0.021	0.722	0.375	0.671	0.293	0.735	0.336
	Level 2	0.721	0.370	0.744	0.018	0.715	0.377	0.710	0.338	0.732	0.334
	Level 3	0.749	0.386	0.760	0.058	0.731	0.400	0.744	0.349	0.744	0.354

[a]LGBM: Light Gradient Boosting Model.
[b]LR: logistic regression.
[c]RF: random forest.
[d]MLP: multilayer perceptron.
[e]OCEC: one-class embedding classifier.
[f]AUROC: area under receiver operator characteristics curve.
[g]AUPRC: area under precision recall curve.

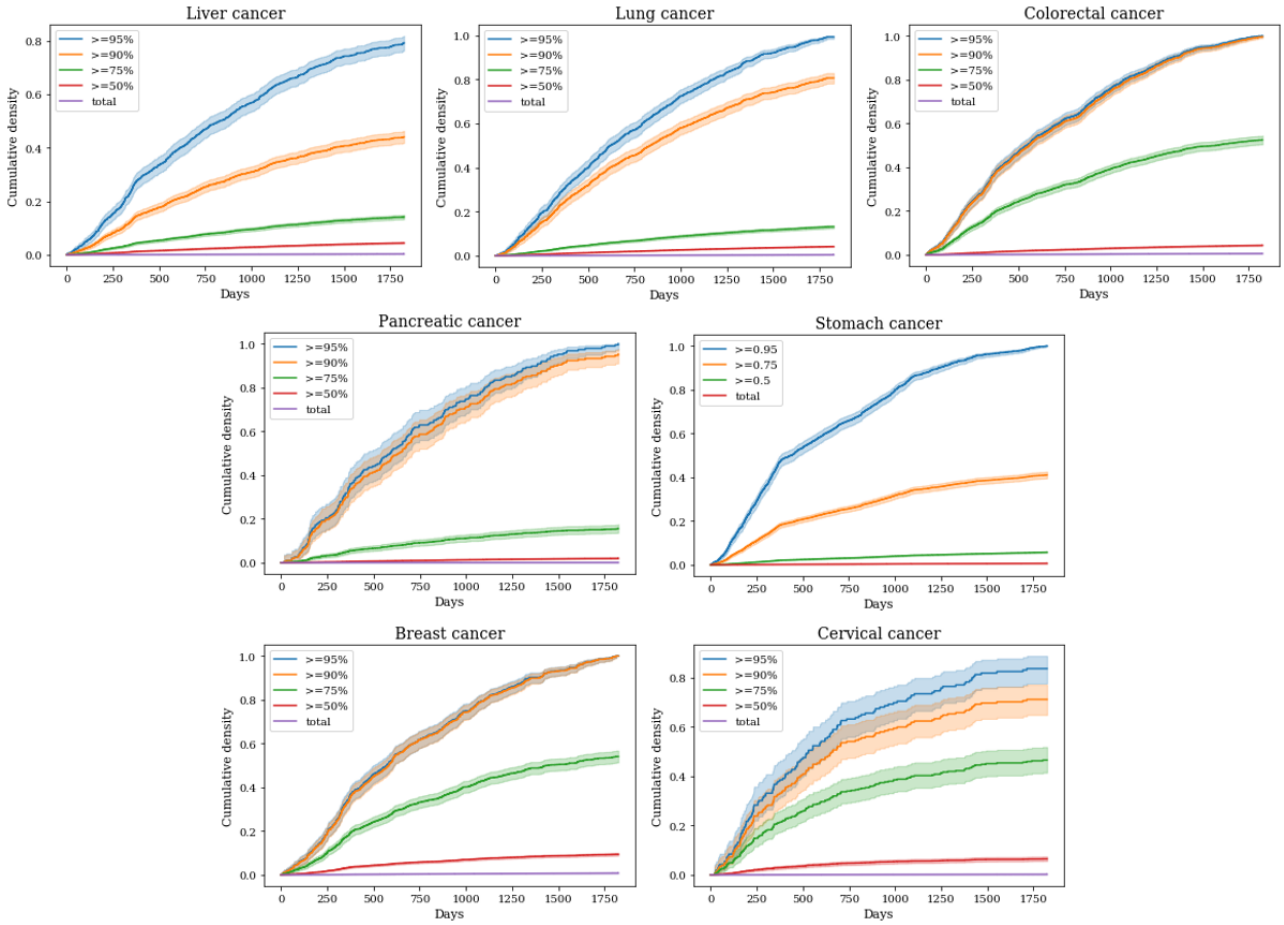
Survival Analysis

To unveil the actual cancer incidence according to the predicted value, we use a survival analysis method. We analyzed the prediction scores of the LGBM model, one of the best performing of the aforementioned models. The prediction score indicates the probability of developing cancer within 5 years from the screening date. Therefore, the closer the prediction score is to 1, the likelier it is that cancer will actually occur after a certain time. We analyzed 5 groups of patients by prediction scores: group 1 (prediction score ≥0.95), group 2 (prediction score ≥0.90), group 3 (prediction score ≥0.75), group 4 (prediction score ≥0.50), and total patient groups. We drew Kaplan-Meier cumulative density curves for each group and compared them. In Figure 3, the x-axis represents time from the screening date, and the y-axis the rate of cancer incidence within the group. All these analyses were performed with

XSL•FO
RenderX

external validation data. As the proportion of cancer patients is <1% for all cancers, the cumulative density curves are attached to the x-axis. The density curve of the group with the higher probability score is located at the higher cumulative density value (y-axis). These trends were collectively observed in all cancers and show the reliability of our models. Significantly, >80% of patients in group 1 actually developed cancer within 5 years.

Figure 3. Kaplan-Meier cumulative density curves.



Model Explainability

With the LGBM and Shapley Additive Explanations (SHAP) method we can explain how the model outputs cancer prediction scores [23]. We can evaluate which features are the most important to predicting future cancer. Moreover, it is possible to know whether a feature has a positive effect or a negative effect.

Table 6 shows the top 5 features for predicting cancer incidence for each type of cancer. Overall, age was the most important variable as was gender except for women's cancers. In addition, drinking frequency, alcohol consumption, and total cholesterol levels were all relevant factors.

In particular, aspartate aminotransferase and gamma-glutamyl transferase levels are important for liver cancer. Smoking frequency is an important variable in lung cancer but not in other cancers. Similarly, drinking is the third most important feature for stomach cancer. In breast and pancreatic cancers, blood glucose levels were a more important variable than they were for other cancers. For further details on SHAP values including correlations between each variable and cancer prediction, see Multimedia Appendix 3.

XSL•FO
RenderX

Table 6. Top 5 features by Shapley Additive Explanations.

Liver	Lung	Colorectal	Pancreatic	Stomach	Breast	Cervical
Age	Age	Age	Age	Age	Age	Age
GTP[a]	Smoking	Sex	Hemoglobin	Sex	BMI	Fasting glucose
AST[b]	Sex	BMI	Total cholesterol	BMI	Total cholesterol	BMI
Total cholesterol	BMI	Total cholesterol	BP[c] (high)	Drinking habit	Fasting glucose	Conjunctivitis
BMI	GTP	Fasting glucose	BMI	Hemoglobin	BP (high)	Total cholesterol

[a]GTP: guanosine triphosphate.

[b]AST: aspartate aminotransferase.

[c]BP: blood pressure.

Discussion

In this study, we used nationwide population-based health care data to construct a machine learning model to predict the future incidence of 7 common types of cancer: liver, stomach, colorectal, lung, pancreatic, breast, and cervical cancer.

Among the 5 distinct models, the LGBM and OCEC, which is our original structure, performed best. Both models had a higher AUROC and AUPRC than did the other models. Interestingly, OCEC scored best in terms of AUROC score and outperformed the normal deep learning method (MLP). Our dense dimension reduction method with one-class anomaly insights was the best model structure.

All models performed well on the external validation set; therefore, it was a success in terms of generalization. Actually, the external validation results were even better than those of the internal validation, thus ensuring the generalizability of our models. We believe that this result was obtained due to the different sampling methods use between the training and validation cohort: the training data set consisted of only those with health checkup information, whereas the validation data set was sampled based on patients' demographic information. As such, the national sample cohort has a similar distribution to the health checkup cohort. In addition, the national sample cohort has a sufficient number of data samples, thus producing good external validation results.

We drew a Kaplan-Meier cumulative density curve for the LGBM model, which is the traditional way to determining whether the marker (prediction score in this case) is suitable to predict cancer occurrence. More than 80% of the people with a prediction score ≥0.95 actually developed cancer within 5 years from the screening date. This is a significant result, which shows that our model can be a powerful tool for identifying high-risk groups. These high-risk groups could then take precautions before the cancer develops. In female cancers, such as breast and cervical cancer, the predictive power was lower than in other cancers. This is probably because both the size of the total female data sample and the number of cancer patients were relatively small. On the other hand, the predictive power for liver and lung cancer was very high. Our data set included liver-related features such as glutamic oxaloacetic transaminase and glutamate pyruvate transaminase. Moreover, we believed that smoking- and drinking-related features also helped predict these cancers. Accordingly, we can conclude that securing high-quality features and a large amount of data can improve predictive power.

There have been previous attempts to develop cancer prediction models with various input features. Japanese researchers developed a prediction model for the 10-year risk of hepatocellular carcinoma in middle-aged Japanese people using data obtained from 17,654 Japanese aged 40 to 69 years who participated in regular health checkups [24]. They obtained a higher AUROC (0.933) than did our models (0.912 in level 1 feature set). However, they did not provide AUPRC, which is important in real-world settings. Furthermore, they used viral markers of hepatitis virus B and C, which are not commonly checked in the normal population. Compared to the previous model, our model used general input features that are easily obtainable, and we acquired a comparable AUROC to the previous model. A Korean research group developed a risk prediction model using Cox proportional hazard regression models for colorectal cancer with a population of 846,559 men and 479,449 women who participated in health examinations by the National Health Insurance Corporation, and they obtained C statistics between 0.69 and 0.78 [25]. They used a similar data set with a different timespan (from 1997 to1997) from our data set and obtained a similar performance to our model (0.730 vs 0.780) This means the performance of classifiers tends to depend on the training data set characteristics rather than the data and time windows. In another study, a multivariable lung cancer risk prediction model including low-dose computed tomography screening results from 22,229 participants obtained an AUROC of 0.761, which is lower than that of our model (0.898 in the MLP model) [26]. Importantly, our model showed a higher performance with an AUROC of 0.875 in a simple linear model (logistic regression with level 1 input features).

In terms of real-world implementation, this study has several implications. Thus far, many studies using machine learning have been conducted on EHR time sequence data. One study aimed to predict heart failure from EHR data [27], and others focused on diabetes development [28-30] or hypertension [31,32]. Furthermore, a few studies have used nationwide claim health checkup data to create a cancer prediction model [33-36]. To solve the overdiagnosis problem of cancer screening programs resulting in unnecessary intervention, accurate, easy-to-implement, patient-level models should be developed. Applying the developed algorithms in previous studies to

XSL•FO
RenderX

hospital sites requires considerable effort because the data structure of the developed model differs from that of hospitals. However, our models have the same data structure as the national health care claim data generated on a monthly basis, which means that our models can be directly applied to EHR and makes this study meaningful in terms of its easy applicability. In addition, since we applied an explainable model to LGBM, every doctor can access the modifiable risk factors from the predicted results.

Our research has several limitations. First, this study used only South Korean nationwide claim data. Depending on the country, the performance of the developed algorithms can differ. The value of NHIS data is well-known, and the data have been used in previous epidemiologic studies. Furthermore, we validated the developed algorithms using another database. Future additional external model validations using claim data from other countries can provide robustness to the models. Second, comparative effectiveness research is needed to prove the usefulness of the developed models. Conventional screening models can be compared to new patient-level prediction models in terms of cost and the number of false-positives avoided by the new models.

Acknowledgments

This research was supported by the SNUBH Research Fund (grant #14-2017-0018), the National Research Foundation of Korea grant funded by the Korea government (NRF-2017R1E1A1A03070105 and NRF-2019R1A5A1028324), and the Institute for Information & Communications Technology Promotion grant funded by the Korea government (Artificial Intelligence Graduate School Program [POSTECH]; #2019-0-01906).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Ten disease codes used as features for each cancer.
[DOCX File , 13 KB - medinform_v9i8e29807_app1.docx]

Multimedia Appendix 2
Hyperparameters used for training models.
[DOCX File , 13 KB - medinform_v9i8e29807_app2.docx]

Multimedia Appendix 3
Shapley Additive Explanations (SHAP) summary plot for each cancer.
[DOCX File , 378 KB - medinform_v9i8e29807_app3.docx]

References

1. Global Cancer Observatory. World Health Organization. URL: https://gco.iarc.fr/ [accessed 2021-04-13]
2. Anand P, Kunnumakkara AB, Kunnumakara AB, Sundaram C, Harikumar KB, Tharakan ST, et al. Cancer is a preventable disease that requires major lifestyle changes. Pharm Res 2008 Sep;25(9):2097-2116 [FREE Full text] [doi: 10.1007/s11095-008-9661-9] [Medline: 18626751]
3. Centers for Disease ControlPrevention (CDC). Cancer screening - United States, 2010. MMWR Morb Mortal Wkly Rep 2012 Jan 27;61(3):41-45 [FREE Full text] [Medline: 22278157]
4. Fracheboud J, de Koning H, Boer R, Groenewoud J, Verbeek A, Broeders M, National Evaluation Team for Breast cancer screening in The Netherlands. Nationwide breast cancer screening programme fully implemented in The Netherlands. Breast 2001 Feb;10(1):6-11. [doi: 10.1054/brst.2000.0212] [Medline: 14965550]
5. de Koning H. Assessment of nationwide cancer-screening programmes. The Lancet 2000 Jan;355(9198):80-81. [doi: 10.1016/s0140-6736(99)00419-5]
6. Romero Y, Trapani D, Johnson S, Tittenbrun Z, Given L, Hohman K, et al. National cancer control plans: a global analysis. The Lancet Oncology 2018 Oct;19(10):e546-e555. [doi: 10.1016/s1470-2045(18)30681-8]
7. Suh Y, Lee J, Woo H, Shin D, Kong S, Lee H, et al. National cancer screening program for gastric cancer in Korea: Nationwide treatment benefit and cost. Cancer 2020 Jan 01;126(9):1929-1939 [FREE Full text] [doi: 10.1002/cncr.32753] [Medline: 32031687]
8. Geller AC, Greinert R, Sinclair C, Weinstock MA, Aitken J, Boniol M, et al. A nationwide population-based skin cancer screening in Germany: proceedings of the first meeting of the International Task Force on Skin Cancer Screening and Prevention (September 24 and 25, 2009). Cancer Epidemiol 2010 Jun;34(3):355-358. [doi: 10.1016/j.canep.2010.03.006] [Medline: 20381443]

XSL•FO
RenderX

9. Published Online First:3 February 2017. WHO | National Cancer Control Programmes (NCCP). URL: https://www.who.int/cancer/nccp/en/ [accessed 2021-04-13]
10. Kim Y, Jun JK, Choi KS, Lee HY, Park EC. Overview of the National Cancer screening programme and the cancer screening status in Korea. Asian Pac J Cancer Prev 2011;12(3):725-730 [FREE Full text] [Medline: 21627372]
11. Lewis S, Huang K, Nguyen T, Gandomkar Z, Norsuddin N, Thoms C. Characteristics of frequently recalled false positive cases in screening mammography. 2020 Presented at: The 15th International Workshop on Breast Imaging (IWBI2020); 24-27 May 2020; Leuven, Belgium. [doi: 10.1117/12.2560290]
12. Le MT, Mothersill CE, Seymour CB, McNeill FE. Is the false-positive rate in mammography in North America too high? Br J Radiol 2016 Sep;89(1065):20160045 [FREE Full text] [doi: 10.1259/bjr.20160045] [Medline: 27187600]
13. Walker R, Enderling H. A new paradigm for personalized cancer screening. bioRxiv. 2018. URL: https://www.biorxiv.org/content/10.1101/265959v1 [accessed 2021-04-13]
14. Román M, Sala M, Domingo L, Posso M, Louro J, Castells X. Personalized breast cancer screening strategies: A systematic review and quality assessment. PLoS One 2019 Dec 16;14(12):e0226352 [FREE Full text] [doi: 10.1371/journal.pone.0226352] [Medline: 31841563]
15. Seong SC, Kim Y, Park SK, Khang YH, Kim HC, Park JH, et al. Cohort profile: the National Health Insurance Service-National Health Screening Cohort (NHIS-HEALS) in Korea. BMJ Open 2017 Sep 24;7(9):e016640 [FREE Full text] [doi: 10.1136/bmjopen-2017-016640] [Medline: 28947447]
16. Shen W, Zhou M, Yang F, Dong D, Yang C, Zang Y, et al. Learning from experts: developing transferable deep features for patient-level lung cancer prediction. 2016 Presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention; 17-21 Oct 2016; Athens. [doi: 10.1007/978-3-319-46723-8_15]
17. Kwon S. Thirty years of national health insurance in South Korea: lessons for achieving universal health care coverage. Health Policy Plan 2009 Jan 12;24(1):63-71. [doi: 10.1093/heapol/czn037] [Medline: 19004861]
18. Lee YH, Han K, Ko SH, Ko KS, Lee KU, Taskforce Team of Diabetes Fact Sheet of the Korean Diabetes Association. data analytic process of a nationwide population-based study using national health information database established by national health insurance service. Diabetes Metab J 2016 Feb;40(1):79-82 [FREE Full text] [doi: 10.4093/dmj.2016.40.1.79] [Medline: 26912157]
19. Hong S, Won YJ, Park YR, Jung KW, Kong HJ, Lee ES, Community of Population-Based Regional Cancer Registries. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2017. Cancer Res Treat 2020 Apr;52(2):335-350 [FREE Full text] [doi: 10.4143/crt.2020.206] [Medline: 32178489]
20. NHI program. h-well NHIS. URL: https://www.nhis.or.kr/static/html/wbd/g/a/wbdga0405.html [accessed 2021-04-16]
21. Statistics Korea news. Statistics Korea. URL: http://kostat.go.kr/portal/eng/news/3/index.board?bmode=read&aSeq=71706 [accessed 2021-04-16]
22. Ruff L, Vandermeulen R, Goernitz N, Deecke L, Siddiqui S, Binder A, et al. Deep one-class classification. 2018 Presented at: The 35th International Conference on Machine Learning; 10-15 July 2018; Stockholm.
23. Lundberg S, Su-In L. A unified approach to interpreting model predictions. arXiv. URL: http://arxiv.org/abs/1705.07874 [accessed 2021-04-13]
24. Michikawa T, Inoue M, Sawada N, Iwasaki M, Tanaka Y, Shimazu T, Japan Public Health Center-based Prospective Study Group. Development of a prediction model for 10-year risk of hepatocellular carcinoma in middle-aged Japanese: the Japan Public Health Center-based Prospective Study Cohort II. Prev Med 2012 Aug;55(2):137-143. [doi: 10.1016/j.ypmed.2012.05.017] [Medline: 22676909]
25. Shin A, Joo J, Yang H, Bak J, Park Y, Kim J, et al. Risk prediction model for colorectal cancer: National Health Insurance Corporation study, Korea. PLoS One 2014 Feb 12;9(2):e88079 [FREE Full text] [doi: 10.1371/journal.pone.0088079] [Medline: 24533067]
26. Tammemägi MC, Ten Haaf K, Toumazis I, Kong CY, Han SS, Jeon J, et al. Development and validation of a multivariable lung cancer risk prediction model that includes low-dose computed tomography screening results: a secondary analysis of data from the national lung screening trial. JAMA Netw Open 2019 Mar 01;2(3):e190204 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.0204] [Medline: 30821827]
27. Ng K, Steinhubl SR, deFilippi C, Dey S, Stewart W. Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. Circ Cardiovasc Qual Outcomes 2016 Nov;9(6):649-658 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.116.002797] [Medline: 28263940]
28. Lai H, Huang H, Keshavjee K, Guergachi A, Gao X. Predictive models for diabetes mellitus using machine learning techniques. BMC Endocr Disord 2019 Oct 15;19(1):101 [FREE Full text] [doi: 10.1186/s12902-019-0436-6] [Medline: 31615566]
29. Badholia A. Predictive modelling and analytics for diabetes using a machine learning approach. ITII 2021 Feb 28;9(1):215-223. [doi: 10.17762/itii.v9i1.121]
30. Daanouni O, Cherradi B, Tmiri A. Type 2 diabetes mellitus prediction model based on machine learning approach. 2019 Presented at: Fourth International Conference on Smart City Applications (SCA2019); 2-4 October 2019; Casablanca, Morocco. [doi: 10.1007/978-3-030-37629-1_33]

XSL•FO
RenderX

31. Kanegae H, Suzuki K, Fukatani K, Ito T, Harada N, Kario K. Highly precise risk prediction model for new-onset hypertension using artificial intelligence techniques. J Clin Hypertens (Greenwich) 2020 Mar 09;22(3):445-450 [FREE Full text] [doi: 10.1111/jch.13759] [Medline: 31816148]
32. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak 2019 Jul 29;19(1):146 [FREE Full text] [doi: 10.1186/s12911-019-0874-0] [Medline: 31357998]
33. Sihto H, Lundin J, Lundin M, Lehtimäki T, Ristimäki A, Holli K, et al. Breast cancer biological subtypes and protein expression predict for the preferential distant metastasis sites: a nationwide cohort study. Breast Cancer Res 2011 Sep 13;13(5):R87. [doi: 10.1186/bcr2944] [Medline: 21914172]
34. Lee T, Wang C, Chen T, Kuo KN, Wu M, Lin J, Taiwan Gastrointestinal Disease Helicobacter Consortium. A tool to predict risk for gastric cancer in patients with peptic ulcer disease on the basis of a nationwide cohort. Clin Gastroenterol Hepatol 2015 Feb;13(2):287-293.e1. [doi: 10.1016/j.cgh.2014.07.043] [Medline: 25083561]
35. Zelic R, Garmo H, Zugna D, Stattin P, Richiardi L, Akre O, et al. Corrigendum re "Predicting prostate cancer death with different pretreatment risk stratification tools: a head-to-head comparison in a nationwide cohort study" [Eur Urol 2020;77:180-8]. Eur Urol 2020 Jul;78(1):e45-e47. [doi: 10.1016/j.eururo.2020.03.016] [Medline: 32386780]
36. Ali Khan U, Fallah M, Sundquist K, Sundquist J, Brenner H, Kharazmi E. Risk of colorectal cancer in patients with diabetes mellitus: A Swedish nationwide cohort study. PLoS Med 2020 Nov 13;17(11):e1003431 [FREE Full text] [doi: 10.1371/journal.pmed.1003431] [Medline: 33186354]

Abbreviations

AUPRC: area under precision recall curve
AUROC: area under receiver operator characteristics curve
EHR: electronic health record
LGBM: Light Gradient Boosting Machine
LR: logistic regression
MLP: multilayer perceptron
NCCP: national cancer control program
NHI: National Health Insurance
NHIS: National Health Insurance System
OCEC: one-class embedding classifier
RF: random forest
SHAP: Shapley Additive Explanations
SNUBH: Seoul National University Bundang Hospital

Edited by G Eysenbach; submitted 21.04.21; peer-reviewed by X Cheng, N Hardikar; comments to author 12.05.21; revised version received 07.07.21; accepted 26.07.21; published 30.08.21.

Please cite as:
Lee E, Jung SY, Hwang HJ, Jung J
Patient-Level Cancer Prediction Models From a Nationwide Patient Cohort: Model Development and Validation
JMIR Med Inform 2021;9(8):e29807
URL: https://medinform.jmir.org/2021/8/e29807
doi:10.2196/29807
PMID:34459743

©Eunsaem Lee, Se Young Jung, Hyung Ju Hwang, Jaewoo Jung. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 30.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning

Pei-Fu Chen[1,2], MD; Ssu-Ming Wang[1], MSc; Wei-Chih Liao[1], MSc; Lu-Cheng Kuo[3], MD; Kuan-Chih Chen[1,4], MD, MSc; Yu-Cheng Lin[5,6], MD, PhD; Chi-Yu Yang[7,8], MD; Chi-Hao Chiu[9], MS; Shu-Chih Chang[10], MA; Feipei Lai[1,11,12], PhD

[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, National Taiwan University Hospital, National Taiwan University College of Medicine, Taipei, Taiwan
[4]Department of Internal Medicine, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[5]Department of Medical Affairs, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[6]Department of Healthcare Administration, Oriental Institute of Technology, New Taipei City, Taiwan
[7]Department of Information Technology, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[8]Section of Cardiovascular Medicine, Cardiovascular Center, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[9]Section of Health Insurance, Department of Medical Affairs, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[10]Medical Records Department, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[11]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[12]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan

Corresponding Author:
Feipei Lai, PhD
Department of Computer Science and Information Engineering
National Taiwan University
No 1, Sec 4, Roosevelt Road
Taipei, 10617
Taiwan
Phone: 886 0911126526
Email: flai@ntu.edu.tw

Abstract

Background: The International Classification of Diseases (ICD) code is widely used as the reference in medical system and billing purposes. However, classifying diseases into ICD codes still mainly relies on humans reading a large amount of written material as the basis for coding. Coding is both laborious and time-consuming. Since the conversion of ICD-9 to ICD-10, the coding task became much more complicated, and deep learning– and natural language processing–related approaches have been studied to assist disease coders.

Objective: This paper aims at constructing a deep learning model for ICD-10 coding, where the model is meant to automatically determine the corresponding diagnosis and procedure codes based solely on free-text medical notes to improve accuracy and reduce human effort.

Methods: We used diagnosis records of the National Taiwan University Hospital as resources and apply natural language processing techniques, including global vectors, word to vectors, embeddings from language models, bidirectional encoder representations from transformers, and single head attention recurrent neural network, on the deep neural network architecture to implement ICD-10 auto-coding. Besides, we introduced the attention mechanism into the classification model to extract the keywords from diagnoses and visualize the coding reference for training freshmen in ICD-10. Sixty discharge notes were randomly selected to examine the change in the F_1-score and the coding time by coders before and after using our model.

Results: In experiments on the medical data set of National Taiwan University Hospital, our prediction results revealed F_1-scores of 0.715 and 0.618 for the ICD-10 Clinical Modification code and Procedure Coding System code, respectively, with a *bidirectional encoder representations from transformers* embedding approach in the Gated Recurrent Unit classification model. The well-trained

XSL•FO
RenderX

models were applied on the ICD-10 web service for coding and training to ICD-10 users. With this service, coders can code with the F_1-score significantly increased from a median of 0.832 to 0.922 (P<.05), but not in a reduced interval.

Conclusions: The proposed model significantly improved the F_1-score but did not decrease the time consumed in coding by disease coders.

(JMIR Med Inform 2021;9(8):e23230) doi:10.2196/23230

KEYWORDS

natural language processing; deep learning; International Classification of Diseases; Recurrent Neural Network; text classification

Introduction

The International Classification of Diseases (ICD) is a medical classification list released by the World Health Organization, which defines the universe of diseases, disorders, injuries, and other related health conditions and the classifying standard of diagnosis [1]. Since the first publication in 1893, the ICD has become one of the most important indexes in medical management systems, health insurance, or literature research.

At present, in most medical institutions, ICD-10 codes that are used in diagnostic related group subsidy for inpatients mainly rely on manual coding from a group of licensed and professional disease coders on a case-by-case basis, who spend a lot of time reading a multitude of medical materials. On the other hand, some other cases—especially outpatients—are coded by physicians.

Since the conversion from ICD-9 to ICD-10 in 2014, Taiwan has used the ICD-10 as the reference for diagnostic-related group subsidy. However, because of the complexity of the ICD-10 structure and coding rules such as the code orders, the inclusion and exclusion criteria, and the enormously increasing number of ICD-10 codes, ICD-10 coding work became much more laborious and time-consuming, even if a disease coder with professional abilities takes approximately 30 minutes per case on average. According to the analysis from *Handbook of Research on Informatics in Healthcare and Biomedicine*, the cost for adopting the ICD-10 system, including training of disease coders, physicians, and code users; initial and long-term loss of productivity among providers; and sequential conversion, is estimated to range from a 1-time cost of US $425 million to US $1.15 billion in addition to US $5-40 million per year in lost productivity [2].

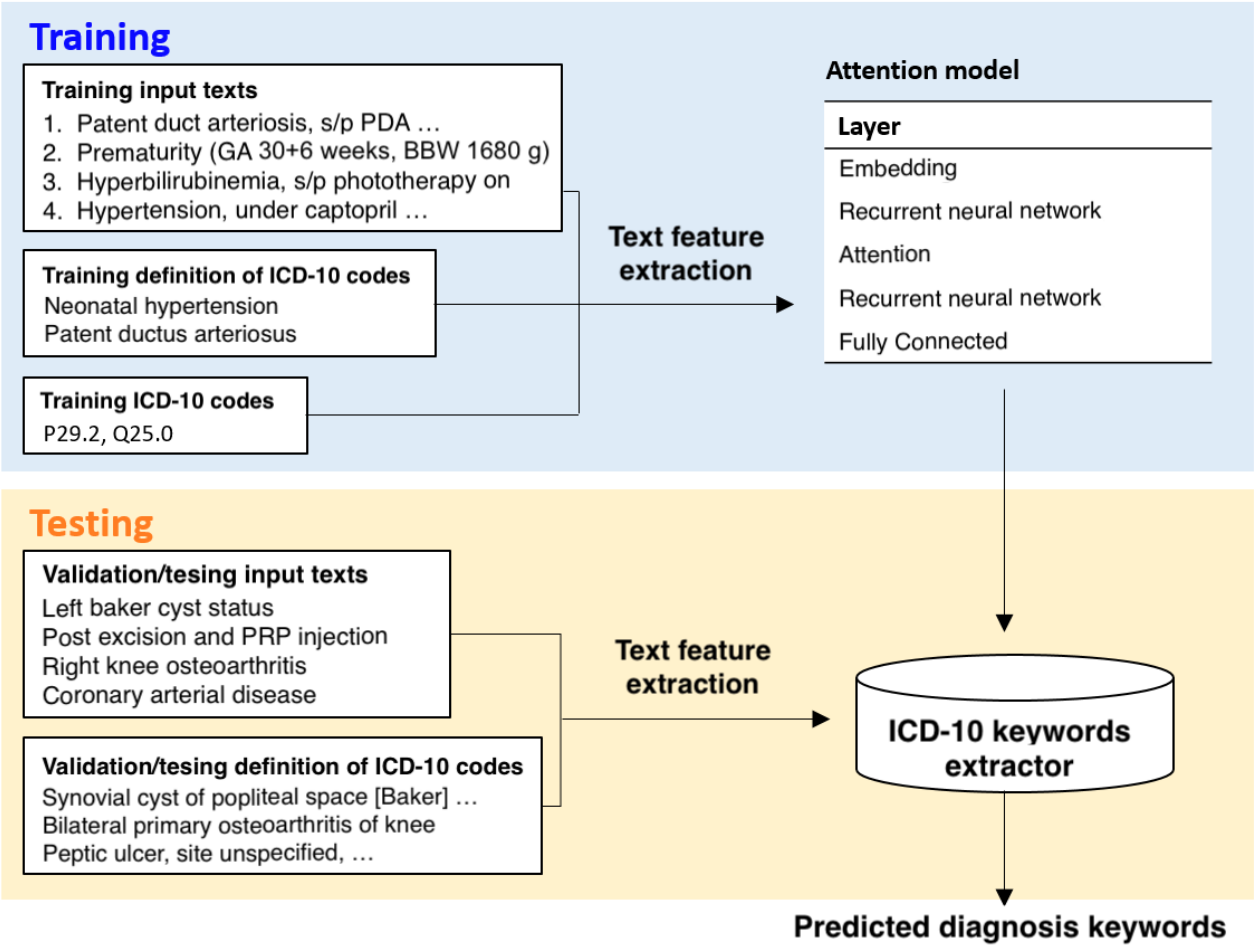
Previous studies had built a model for the ICD-9 system. In 2008, Farkas and Szarvas [3] utilized a rule-based approach querying other reference tools to implement the ICD auto-coding task. However, compared to ICD-9, ICD-10 contains more than 60,000 codes. Building a rule-based automatic system is labor-intensive and time-consuming. In addition, the entirety of the rules of the ICD-10 system is complicated even for disease coders. For the aforementioned reasons, recent studies have emphasized on deep learning– and natural language processing (NLP)–related approaches; for instance, Zhang et al [4] used a gated recurrent unit (GRU) network with content-based attention to predict medication prescriptions on the basis of the disease codes, and Wang et al [5] applied and compared NLP techniques such as Global Vectors (GloVe) in an electronic health record (EHR) data classification task.

In previous studies, we have already applied word to vectors (Word2Vec), an NLP method, in an ICD-10 auto-coding task and achieved an F_1-score of 0.67/0.58 in Clinical Modification (CM)/Procedure Coding System (PCS). Furthermore, we also built an ICD-10 code recommendation system for ICD-10 users [6,7]. In this study, we made a comparison on most of the recent NLP approaches such as Word2Vec, embeddings from language models (ELMo), and bidirectional encoder representations from transformers (BERT). Furthermore, we introduced the attention mechanism to our classification model to visualize the word importance for training new coders in ICD-10 coding.

In the ICD classification framework illustrated in Figure 1, the left panel denotes the large amounts of free-text data written by physicians, which would be read and learned by the classifier in the right panel of the graph with supervised learning. Well-trained classifiers would be applied to predict the ICD-10 codes accurately for each patient. Furthermore, to distinguish the primary, secondary, or additional diagnosis, a sequential correction was conducted by coding the ICD-10 codes in a sequential format, using a sequence-to-sequence model followed by combining the classification coding results with the sequential order outcome.

XSL•FO
RenderX

Figure 1. Training and validation process for the ICD-10 classification and attention models. BBW: birth body weight; GA: gestational age; PRP: platelet-rich plasma.



The attention framework for paragraph highlighting is also illustrated in Figure 1. Different from the classification framework, the input data in the left panel include both the diagnoses and the corresponding ICD-10 definitions from the National Health Insurance Administration rather than using merely the diagnoses, and the output data in the right panel is the attention weight matrix extracted from the predicting process rather than the classification result. With a combination of these 2 methods, we constructed an ICD-10 auto-coding and training system to assist ICD-10 code users.

Our study aims at building an automatic ICD-10 coding and training system based on NLP technology, attention mechanism, and Deep Neural Network (DNN) models, which are applied for extracting information from EHR data, highlighting the key points from the extracted features, and implementing an ICD-10 classification task with sequential correction, respectively, for assisting all ICD-10 users.

Methods

Data Description

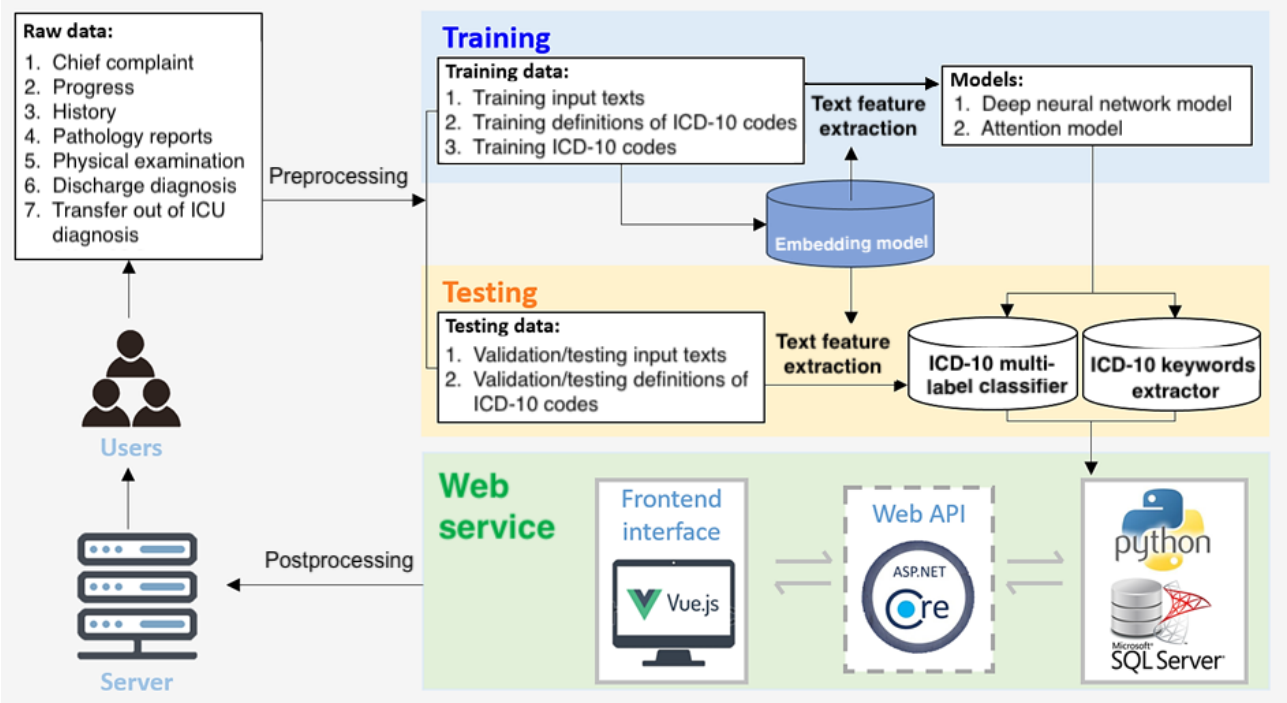
Our data were acquired from patients at National Taiwan University Hospital (NTUH) from January 2016 to July 2018. The ground-truth ICD-10 codes were annotated by the coders at NTUH. Data attributes and types include account IDs, type contents, course and treatment, and discharge diagnoses. The distribution of ICD-10 codes is shown in our previous study [7].

System Architecture

The entire process of the system constructing framework is composed of data processing, feature extracting, model constructing, model training, and web service building. To detail and visualize the ICD-10 web service clearly in this study, the complete workflow of the ICD-10 coding and training system is illustrated in Figure 2.

XSL•FO
RenderX

Figure 2. Complete framework of the ICD-10 auto-coding and training system. API: application programming interface; ICU: intensive care unit.



Data Processing

Preprocessing

Preprocessing, including the removal of Chinese words, null or duplicate elements, punctuation, stop words, and infrequent words, was applied before tokenization of the texts. The basic preprocessing methods were applied using the Natural Language Toolkit [8] and Scikit-Learn [9] library. We then randomly split the data set at a 9:1 ratio into training and validation sets with the Scikit-Learn library.

Postprocessing

In ICD-10 coding, combination codes remain an intractable issue because, in some cases, disease coders cannot—and should not—assign multiple diagnosis codes when a single combination code clearly identifies all aspects of the patient's diagnosis [10]. In this study, a user-defining panel is provided in the auto-coding system to deal with combination codes by replacing the incorrect outcomes, where the combination codes were either predicted incorrectly or separated into 2 different codes on the basis of the given codes.

Feature Extraction

During feature extraction, we applied NLP techniques, including GloVe [11], Word2Vec [12], ELMo [13], BERT [14], and single head attention recurrent neural network (SHA-RNN), to convert the word contexts to numerical data and extract word and contextual information. Except for the BERT-based pretrained weight, we also attempted clinicalBERT [15] and BioBERT [16], which were trained with clinical notes from MIMIC-III, PubMed, and PubMed Central. Hyperparameters of the embedding models are attached in Table 1.

XSL•FO
RenderX

Table 1. Hyperparameters of word-embedding models.

Hyperparameters	Size/number
Global Vector	
Word embedding size	100
Word to Vectors	
Word embedding size	300
Embeddings from Language Models	
Convolutional neural network char embedding size	50
Convolutional neural network word embedding size	100
Highway number	2
Intermediate size	512
Bidirectional encoder representations from transformers[a]	
Word embedding size	768
Sentence embedding size	768
Position embedding size	768
Intermediate size	3072
Attention head number	12
Hidden layer number	12
Dropout	0.1
Single head attention recurrent neural network	
Word embedding size	1024
Hidden size	1024
Layer number	4

[a]Clinical bidirectional encoder representations from transformers (BERT) and BERT for biomedical text mining shared the same hyperparameters with BERT.

Classification Model

The classification model was constructed with 4 neural network layers, including RNN and fully connected neural network (FCNN), where the hyperparameters are shown in Table 2 and the architecture is shown in Figure 3. The first layer is the word embedding layer, which transforms the tokenized word list input into word vectors. The second layer is a bidirectional GRU (BiGRU) layer [17]. The remaining 2 layers are fully connected layers, where the final fully connected layer should be set to the size of the dimension we expect to predict. In our case, we conducted 2 classification tasks, including whole label classification for CM and PCS with 14,602/9780 labels of CM/PCS in NTUH data records in total. Hence, the final fully connected layer size should be set to 14,602 and 9780 dimensions, respectively. To make a comparison, a classification model with only 1 fully connected layer— fully connected layer 2—was used as the baseline model. In addition, the attention mechanism based on the Bahdanau [18] attention model was introduced to our classification model to further extract the keywords for ICD-10 coding by computing the weight information of context—ICD title–vector pairs; that is, the importance of the information with respect to the current target word.

Table 2. Hyperparameters of the classification models.

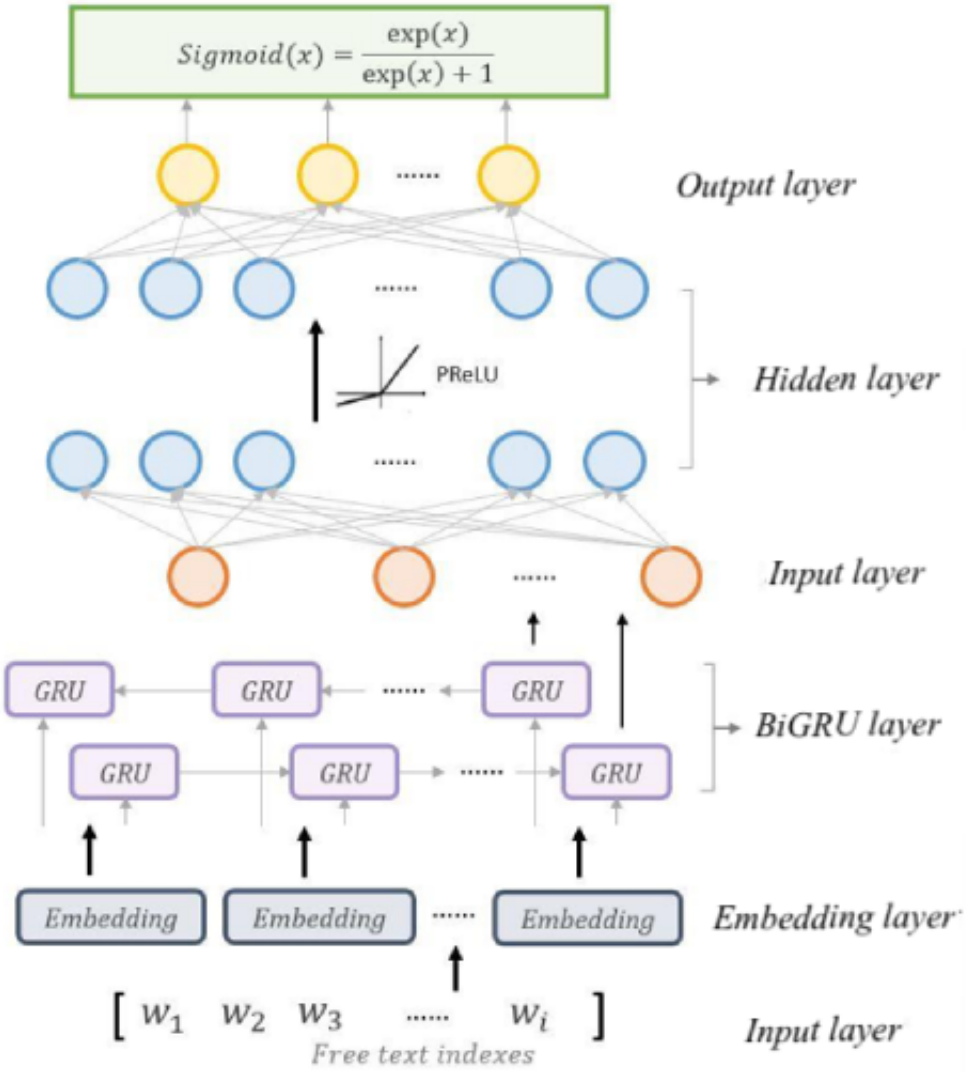
Hyperparameters	Size
Bidirectional GRU[a] layer	256
Fully connected layer 1	700
Fully connected layer 2 CM/PCS[b]	14,602/9780
Dropout	0.2

[a]GRU: Gated Recurrent Unit.

[b]CM/PCS: Clinical Modification/Procedure Coding System.

XSL•FO
RenderX

Figure 3. Architecture of the Deep Neural Network classification model. BiGRU: Bidirectional Gated Recurrent Unit; GRU: Gated Recurrent Unit; PReLU: Parametric Rectified Linear Unit.



Model Assessment

Micro F_1-score is the harmonic mean of recall and precision, which are the sum of the number of true-positive results divided by sum of the number of all positive results and the sum of the number of true-positive results divided by the sum of the number of all relevant samples, respectively. The micro F_1-score considers the number for each label while calculating the recall and precision; hence, it is appropriate for evaluating the performance of a multi-label classification task with imbalanced data set.

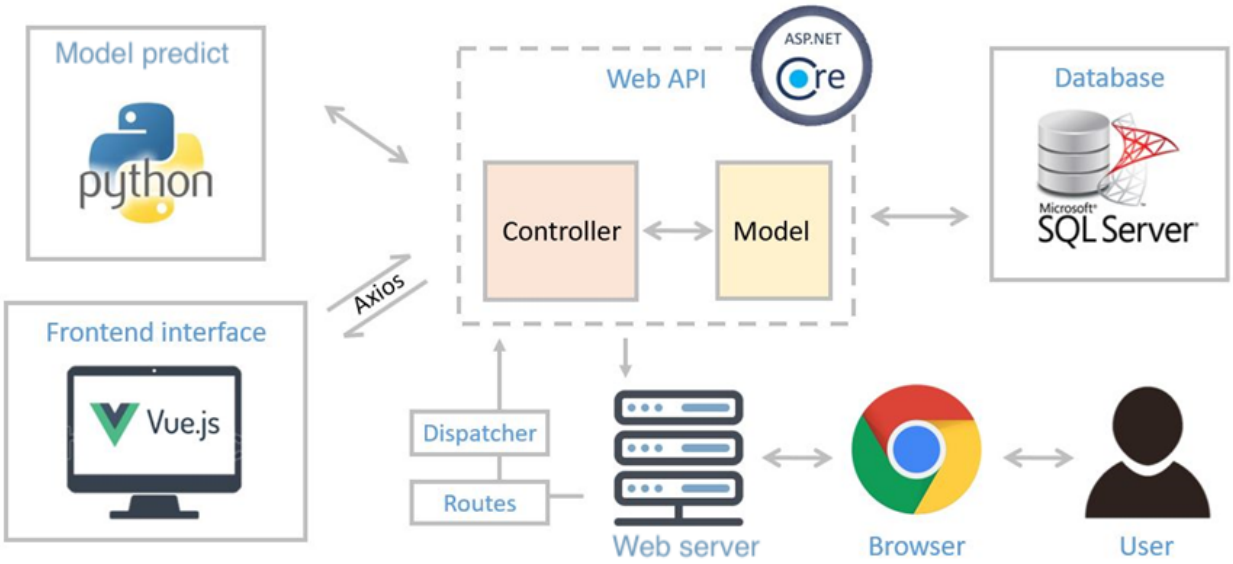
For realistic application in the auto-coding system, recall@K, which calculates the proportion of correct answers in the first K prediction results returned by the classifier, was also applied for validating the model's performance. In our case, considering the limitation of the quantity of CM and PCS codes, 20 was chosen as the K value.

ICD-10 Coding and Training System Framework

An ICD-10 auto-coding and training system prototype was constructed with python3, ASP.NET Core 2.2 MVC, SQL Server, and Vue.js. Whenever a user performs an action, such as typing a discharge diagnosis or retrieving information from a database on the frontend interface built with Vue.js, the axios, a promise-based HTTP client for the browser and node.js, would call for the Web application programming interface in the backend built with ASP.NET Core 2.2 MVC to send the case information to the backend for predicting and processing via python3 or to the database for data preservation in SQL Server. The complete system framework is illustrated in Figure 4. In ICD-10 Coder and Trainer, with the discharge diagnosis as the data input, the top 20 related ICD-10-CM/PCS codes and the importance of each word related to the corresponding code would be returned to all ICD-10 users for auxiliary.

XSL•FO
RenderX

Figure 4. System architecture of the ICD-10 auto coding and training web service. API: application programming interface.



Comparing the Time Consumed and the F_1-Score With and Without the Auto-Coding System

We collected 60 discharge notes from February 2021 from the Far Eastern Memorial Hospital (New Taipei City, Taiwan) randomly. Nine coders participated in this experiment. The most experienced coder provided the ground truth. The other 8 coders were divided into 4 groups, and each case assigned to each group could be coded by 2 coders. There are 2 parts in this experiment. In part 1, we only provided medical record numbers, and the coders coded the randomly assigned medical records on a daily basis. Each group was assigned a different set of 10 cases. In part 2, we provided medical record numbers and ICD codes predicted by our best DNN classification model. Each group was randomly assigned 5 cases. We compared the time consumed and the F_1-score between parts 1 and 2 and performed a paired samples Wilcoxon signed-rank test. A 2-tailed $P<.05$ was considered significant. Furthermore, a questionnaire was designed to collect coders' opinions on this system.

Results

ICD-10-CM Whole Label Classification

In the NTUH data set, the complete ICD-10-CM codes (ie, CM codes with 3-7 characters) corresponding to the discharge diagnosis records comprise 14,602 labels in total. The best DNN classification model based on BERT embedding and FCNN with BiGRU could achieve an F_1-score of 0.715 and recall@20 of 0.873. Table 3 shows all comparisons of the whole label classification. Classification results with different BERT pretrained models show no significant effect on performance in both of baseline and BiGRU models.

Table 3. F_1-score and Recall@20 of all embedding models in the International Classification of Diseases-10 Clinical Modification.

Embedding model	Baseline F_1-score	F_1-score	Recall@20
Word to Vectors	0.355	0.680	0.873
Global Vectors	0.220	0.635	0.836
Embeddings from Language Models	0.633	0.631	0.852
Bidirectional encoder representations from transformers–based	0.715	0.710	0.869
Clinical bidirectional encoder representations from transformers model	0.712	0.714	0.869
Bidirectional encoder representations from transformers for biomedical text mining	0.709	0.701	0.863
Single Head Attention Recurrent Neural Network	0.402	0.570	0.835

ICD-10-PCS Whole Label Classification

In the ICD-10-PCS whole label classification task, the complete ICD-10-PCS code (ie, PCS codes with 7 characters) corresponding to discharge diagnosis records comprised 9513 labels. Progress and discharge diagnosis were applied for training the DNN model. The results summarized in Table 4 imply that our best DNN classification model based on BERT embedding and FCNN with BiGRU could achieve an F_1-score of 0.618 and a recall@20 of 0.887.

XSL•FO
RenderX

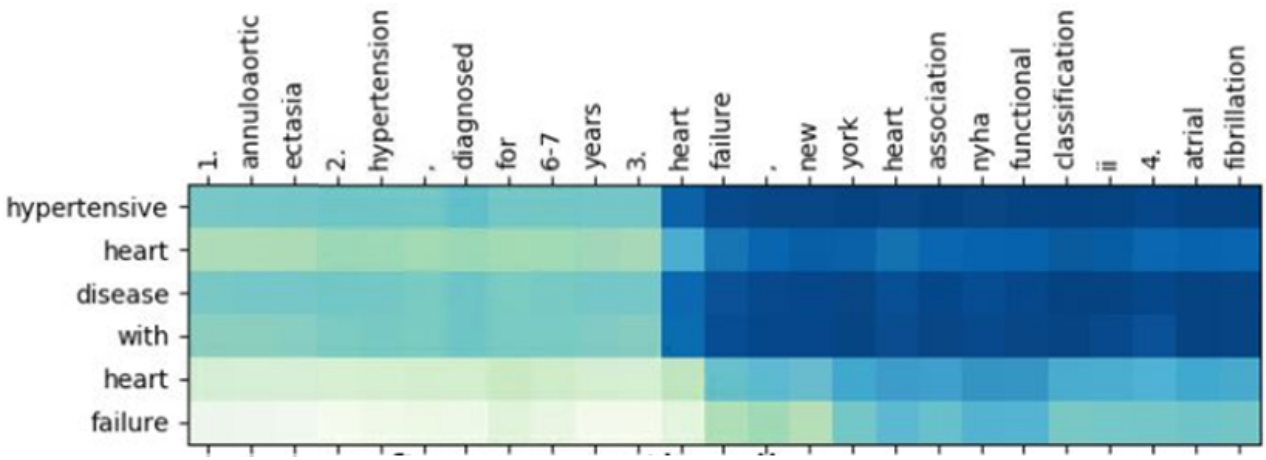
Table 4. F_1-score and recall@20 of all embedding models in the International Classification of Diseases-10 Procedure Coding System.

Embedding model	Baseline F_1-score	F_1-score	Recall@20
Word to Vectors	0.278	0.580	0.850
Global Vectors	0.120	0.520	0.841
Embeddings from Language Models	0.547	0.557	0.874
Bidirectional encoder representations from transformers–based	0.618	0.611	0.880
Clinical bidirectional encoder representations from transformers model	0.596	0.615	0.887
Bidirectional encoder representations from transformers for biomedical text mining	0.611	0.613	0.880
Single Head Attention Recurrent Neural Network	0.269	0.527	0.879

ICD-10 Classification With Attention

By introducing the attention mechanism into the classification model, the relation and importance between word pairs could be computed and visualized. For instance, for 2 sentences, "He had coronary artery disease. Also, he got fever." and "A heart disease," weight information for the word "heart" might focus on "coronary" or "artery." Hence, by extracting the attention weights of the diagnoses and ICD-10 definitions, how coders focus on the words within diagnoses during the ICD-10 coding process could be well understood (Figure 5). Furthermore, the extracted diagnosis attention weights and the corresponding ICD-10 code could be visualized by highlighting the key words, the weight of which would be higher than a certain threshold, for training a new coder in disease coding. By considering all positive cases and negative sampling up to 40 cases in total, the classification model with the attention mechanism could achieve an F_1-score of 0.86.

Figure 5. Visualization of attention weights.



ICD-10 Coding and Training System Framework

The objective of this study is to build an ICD-10 auto-coding and training system for assisting disease coders to elevate their work efficiency and coding accuracy. An ICD-10 auto-predicting interface with discharge diagnosis as the reference is available on the internet [19] for accelerating the coding efficiency. The DNN model executed by the python script would return the top 20 ICD-10-CM and ICD-10-PCS codes with a recall@20 of 0.87 and 0.88, respectively. The predicting process of each case takes less than 30 seconds, which drastically shortens the coding time of 30 minutes per case on average. In addition, training for ICD-10 coding is also provided under the training tab. Given a paragraph of discharge diagnosis, the key words to support the code could be highlighted by clicking on the target code.

To make the prediction more flexible and adaptable to disease coders in different hospitals, postprocessing rules for dealing with exceptions, such as combination codes and hospital consensus, could be defined under the rule definition panel. Users could apply the default setting or build their own setting to apply the specific coding style. The ICD-10 auto-coding, training, and rule defining panels are shown in Figures 6, 7, and 8 respectively.

XSL•FO
RenderX

Figure 6. ICD-10 auto-coding panel.

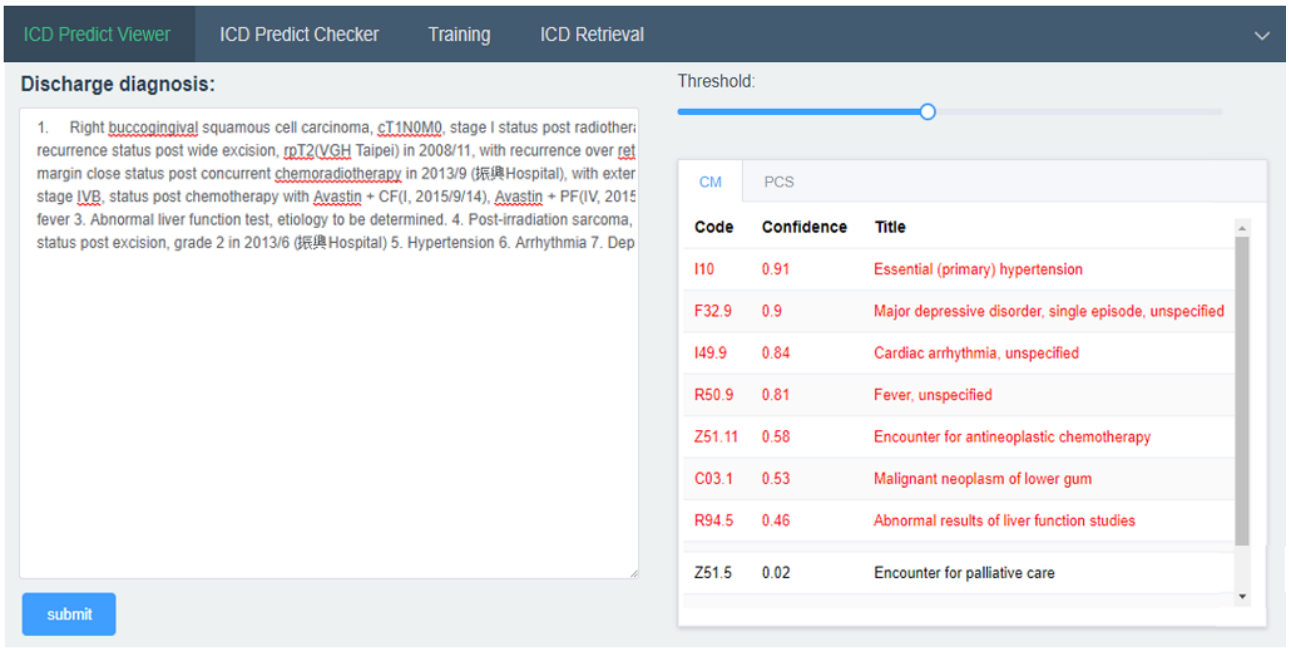
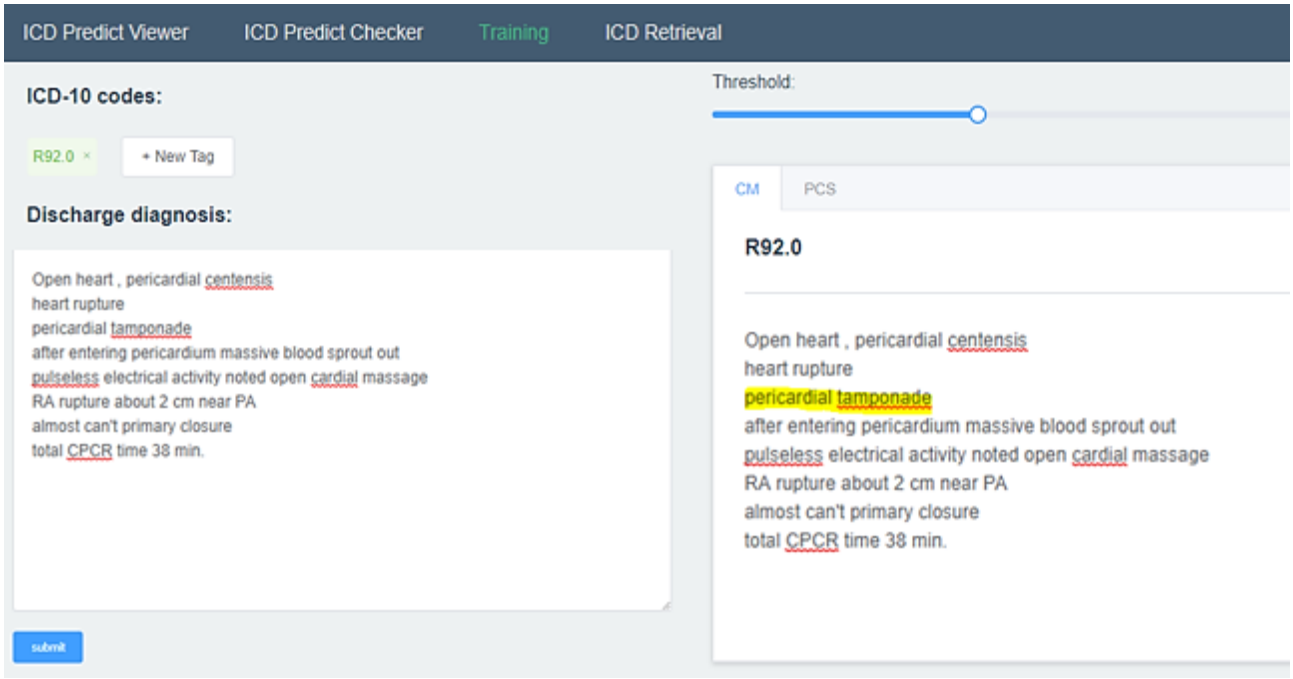
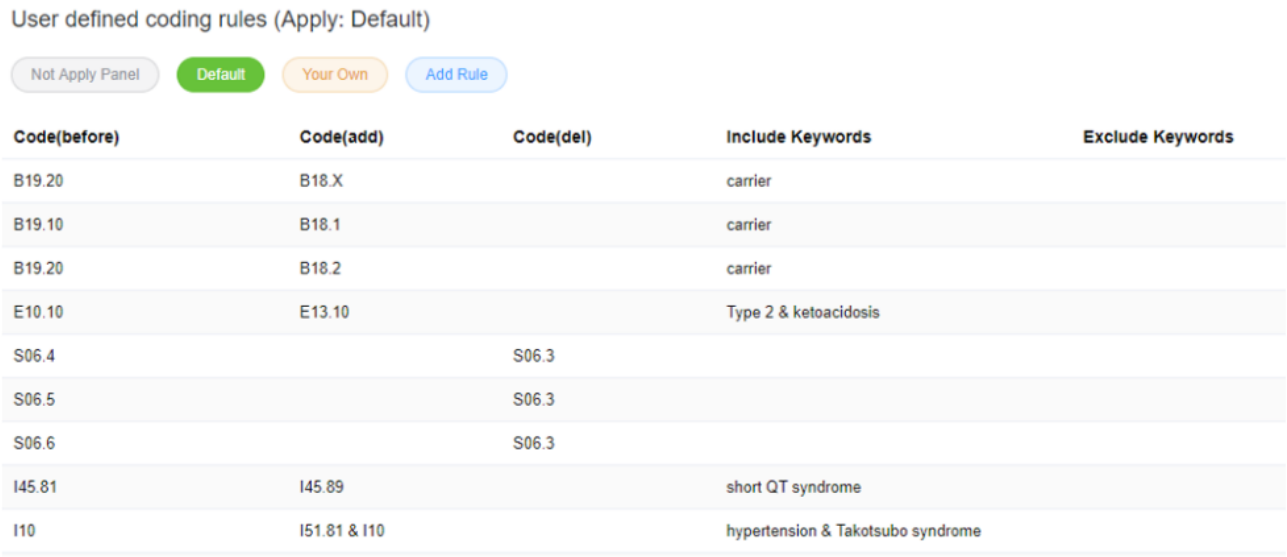


Figure 7. ICD-10 auto-training panel.



XSL•FO
RenderX

Figure 8. Postprocessing user defining panel.



Code(before)	Code(add)	Code(del)	Include Keywords	Exclude Keywords
B19.20	B18.X		carrier	
B19.10	B18.1		carrier	
B19.20	B18.2		carrier	
E10.10	E13.10		Type 2 & ketoacidosis	
S06.4		S06.3		
S06.5		S06.3		
S06.6		S06.3		
I45.81	I45.89		short QT syndrome	
I10	I51.81 & I10		hypertension & Takotsubo syndrome	

Time Consumed and F_1-Score With and Without the Auto-Coding System

The ICD-10 auto-coding system with our best DNN classification model significantly improved the coders' mean F_1-score from a median of 0.832 to 0.922 (P<.05) but did not decrease their mean coding time (P=.64), as shown in Table 5. The questionnaire revealed that a coder took approximately 20-40 minutes on average to code a case, and 62.5% of coders are willing to use this system in their work. This system might potentially help them not only increase the accuracy of ICD-coding but also save their time.

Table 5. Time consumed and the F_1-score with and without the auto-coding system.

Coder	Mean time consumed in part 1[a,b] (minutes:seconds)	Mean time consumed in part 2[c,d,e] (minutes:seconds)	Mean F_1-score in part 1[a,f]	Mean F_1-score in part 2[c,g,h]
1	07:49	05:11	0.801	0.893
2	08:19	06:01	0.900	0.960
3	04:57	06:16	0.980	0.951
4	05:02	07:32	0.867	0.950
5	06:23	05:18	0.766	0.978
6	05:23	03:53	0.652	0.892
7	05:45	05:25	0.815	0.838
8	05:33	06:43	0.848	0.827

[a]Without the auto-coding system.

[b]Median time consumed in part 1=5 minutes 39 seconds (95% CI 5 minutes 1 second to 7 minutes 54 seconds).

[c]With the auto-coding system.

[d]Median time consumed in part 2=5 minutes 43 seconds (95% CI 4 minutes 56 seconds to 6 minutes 52 seconds).

[e]Nonsignificant difference in the mean time consumed by coders between parts 1 and 2 of the study (2-tailed P=.64 derived from a paired samples Wilcoxon signed-rank test).

[f]Median F_1-score in part 1=0.832 (95% CI 0.744-0.915).

[g]Median F_1-score in part 2=0.922 (95% CI 0.836-0.963).

[h]Significant difference in mean F_1-scores between parts 1 and 2 (2-tailed P<.05 derived from a paired samples Wilcoxon rank sum test).

XSL•FO
RenderX

Discussion

Principal Findings

Compared to a previous study on ICD-9 classification with 85,522 training data and an F_1-score of 0.41 [20], our best DNN classification model based on the BERT embedding method and FCNN with BiGRU achieved an F_1-score of 0.715 and recall@20 of 0.873. Comparing to the baseline model with only 1 fully connected layer, models with BiGRU showed better performance within the embedding approaches using fixed word embedding vectors. However, within embedding methods that are more flexible, such as BERT, the BiGRU classification model shows no significant effect on performance. This indicates that higher-level embedding techniques such as ELMo and BERT could certainly be able to sequentially consider the contextual semantics information; since they widely introduce the BiGRU and BiLSTM layers or other contextual information extraction methods within their model architectures. On the other hand, among all the embedding methods, BERT showed the best performance; however, it seems that initializing with different BERT pretrained weights has no significant influence on the classification results. However, the simplified BERT model SHA-RNN could only achieve 0.57 on the classification task and could not achieve over 0.41 on the baseline model. This might result from the lack of the corpus on training of the embedding model, comparing to BERT models which were trained with millions of articles from Bookcorpus, Wikipedia, etc; we only used our own discharge diagnosis records on SHA-RNN training. This implies the ability of the BERT model to learn and extract the information well in a specific field via only the fine-tuning process; thus, there is no need to train our BERT model from scratch with our own data set, but rather only to initialize with the pretrained weight and fine-tune with our own data set.

Another previous study compared BERT with other DNNs in ICD-10 auto-coding in nontechnical summaries of animal experiments. They achieved a micro F_1-score of 73.02% with BioBERT, which is comparable to our results [21]. However, nontechnical summaries of animal experiments are not as complicated as the medical records we worked on and BioBERT could perform better than BERT in their data set, but no significant difference was observed in the medical records, as shown herein. Another study found that contextualized deep learning representation models including BERT and ELMo outperform noncontextualized representation models in discovering medical synonyms [22], which is consistent with our findings.

Our system improved the coder's mean F_1-score ($P<.05$) but did not decrease the mean coding time ($P=.64$). One of the explanations is that coders had not become familiar with this system yet, and the other explanation is that relatively simple cases were included in this experiment, which led them to take less than 20-40 minutes per case during their daily work, as they indicated in their questionnaire responses. The long-term effect of the ICD-10 auto-coding system should be investigated in the future to determine whether the coding time can be saved.

Limitations

Our study has some limitations. First, our training data are derived from only 1 medical center. The performance in other hospitals could be affected by different writing habits, and different disease prevalence. Second, combination codes remain an intractable issue because in some cases, disease coders cannot and should not assign multiple diagnosis codes in cases where a single combination code clearly identifies all aspects of the patient's diagnosis. In our results, the combination codes were either predicted incorrectly or separated into 2 different codes. In addition, there are multiple diagnoses that corresponded to multiple codes in order; that is, primary diagnosis, secondary diagnosis, tertiary diagnosis, etc [10]. However, the classification model could only give the probability of each code rather than the corresponding order. To resolve the problem while maintaining high performance in the classification task, we proposed a novel approach by combining the Seq2Seq model, which gives the code order. Finally, our system is still new to coders, and few coders have used it. After more users' responses are collected, further analysis and modification can be performed to improve our system.

Conclusions

In this study, an ICD-10 classification model developed using NLP and a deep learning model without any background knowledge from EHR data yielded an F_1-score of 0.715 and 0.618 for CM and PCS, respectively. In addition, we built and released the platform for automated ICD-10 prediction and training based on our well-trained models for free to ICD-10 users worldwide and further shortened the coding time from 20-40 minutes to 30 seconds per case. Our platform can be found on the internet [19]. Our system can significantly improve coders' F_1-score in ICD-10 coding.

In future studies, we shall attempt to develop and provide other functions such as user feedback and auto-training with new input data to our model. ICD-10 codes in different hospitals with different coding styles will also be constructed in accordance with the amount of user information and prediction history records to improve the automated ICD-10 coding and training system further.

Acknowledgments

This study was supported by grants from the Ministry of Science and Technology, Taiwan (MOST 110-2634-F-002-032-).

XSL•FO
RenderX

Authors' Contributions

FL and SMW designed the study. SMW and WCL designed and developed the system. PFC, LCK, KCC, YCL, CYY, CHC, and SCC collected the data. PFC and KCC conducted the experiment. SMW, WCL, and PFC conducted statistical analyses and drafted the manuscript. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

References

1. The International Classification of Diseases, 10th Revision. World Health Organization. 2015. URL: https://icd.who.int/browse10/2015/en [accessed 2021-08-04]
2. Lazakidou AA. Handbook of Research on Informatics in Healthcare and Biomedicine. Hershey, PA: IGI Global; 2006.
3. Farkas R, Szarvas G. Automatic construction of rule-based ICD-9-CM coding systems. BMC Bioinformatics 2008 Apr 11;9 Suppl 3:S10 [FREE Full text] [doi: 10.1186/1471-2105-9-S3-S10] [Medline: 18426545]
4. Zhang Y, Chen R, Tang J, Stewart WF, Sun J. LEAP: Learning to Prescribe Effective and Safe Treatment Combinations for Multimorbidity. 2017 Presented at: The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2017; Halifax, NS p. 1315-1324. [doi: 10.1145/3097983.3098109]
5. Wang Y, Afzal N, Fu S, Wang L, Shen F, Rastegar-Mojarad M, et al. MedSTS: a resource for clinical semantic textual similarity. Lang Resour Eval 2018 Oct 24;54(1):57-72. [doi: 10.1007/s10579-018-9431-1]
6. Wang S, Lai F, Sung C, Chen Y. ICD-10 Auto-coding System Using Deep Learning. 2020 Presented at: 10th International Workshop on Computer Science and Engineering; June 19-21, 2020; Shanghai p. 46-51. [doi: 10.18178/wcse.2020.02.008]
7. Wang SM, Chang YH, Kuo LC, Lai F, Chen YN, Yu FY, et al. Using Deep Learning for Automatic Icd-10 Classification from FreeText Data. Eur J Biomed Inform 2020;16(1):1-10 [FREE Full text] [doi: 10.24105/ejbi.2020.16.1.1]
8. Loper E, Bird S. NLTK: the Natural Language Toolkit. 2002 Presented at: ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics; July 7, 2002; Philadelphia, PA. [doi: 10.3115/1118108.1118117]
9. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
10. International Classification of Diseases, Tenth Revision, Clinical Modification. Qeios. URL: https://www.qeios.com/read/SA6DYU [accessed 2021-08-04]
11. Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation. 2014 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 2014; Doha. [doi: 10.3115/v1/d14-1162]
12. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. arXiv. Preprint posted online October 16, 2013. [FREE Full text]
13. Gardner M, Grus J, Neumann M, Tafjord O, Dasigi P, Liu N, et al. AllenNLP: A Deep Semantic Natural Language Processing Platform. 2018 Presented at: Workshop for NLP Open Source Software (NLP-OSS); July 2018; Melbourne. [doi: 10.18653/v1/w18-2501]
14. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019 Presented at: 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019); June 2-7, 2019; Minneapolis, MN. [doi: 10.18653/v1/N19-1423]
15. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. 2019 Presented at: ACM Conference on Health, Inference, and Learning; April 2-4, 2020; Toronto, ON. [doi: 10.1090/mbk/121/79]
16. 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]
17. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv. Preprint posted online December 11, 2014. [FREE Full text]
18. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv. Preprint posted online September 1, 2014. [FREE Full text]
19. ICD Web. URL: https://nets.csie.ntu.edu.tw/ [accessed 2021-08-05]
20. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
21. Amin S, Neumann G, Dunfield K, Vechkaeva A, Chapman KA, Wixted MK. MLT-DFKI at CLEF eHealth 2019: Multi-label classification of ICD-10 codes with BERT. 2019 Presented at: 10th Conference and Labs of the Evaluation Forum; September 9-12, 2019; Lugano URL: http://ceur-ws.org/Vol-2380/paper_67.pdf
22. Schumacher E, Dredze M. Learning unsupervised contextual representations for medical synonym discovery. JAMIA Open 2019 Dec;2(4):538-546 [FREE Full text] [doi: 10.1093/jamiaopen/ooz057] [Medline: 32025651]

XSL•FO
RenderX

Abbreviations

BERT: bidirectional encoder representations from transformers
BiGRU: Bidirectional Gated Recurrent Unit
BioBERT: bidirectional encoder representations from transformers for biomedical text mining
CM: Clinical Modification
DNN: Deep Neural Network
EHR: Electronic Health Records
ELMo: Embeddings from Language Models
FCNN: fully-connected neural network
GloVe: Global Vectors
GRU: Gated Recurrent Unit
ICD: International Classification of Diseases
NTUH: National Taiwan University Hospital
NLP: natural language processing
PCS: Procedure Coding System
SHA-RNN: Single Head Attention Recurrent Neural Network
Word2Vec: Word to Vectors

Edited by C Lovis; submitted 05.08.20; peer-reviewed by A Kimia, G Lim, E Frontoni; comments to author 19.01.21; revised version received 15.03.21; accepted 25.07.21; published 31.08.21.

Please cite as:
Chen PF, Wang SM, Liao WC, Kuo LC, Chen KC, Lin YC, Yang CY, Chiu CH, Chang SC, Lai F
Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning
JMIR Med Inform 2021;9(8):e23230
URL: https://medinform.jmir.org/2021/8/e23230
doi:10.2196/23230
PMID:34463639

©Pei-Fu Chen, Ssu-Ming Wang, Wei-Chih Liao, Lu-Cheng Kuo, Kuan-Chih Chen, Yu-Cheng Lin, Chi-Yu Yang, Chi-Hao Chiu, Shu-Chih Chang, Feipei Lai. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 31.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

An Artificial Neural Network–Based Pediatric Mortality Risk Score: Development and Performance Evaluation Using Data From a Large North American Registry

Niema Ghanad Poor[1,2], MSc; Nicholas C West[3], MSc; Rama Syamala Sreepada[1,3], PhD; Srinivas Murthy[1,4], MD; Matthias Görges[1,3], PhD

[1]Research Institute, BC Children's Hospital, Vancouver, BC, Canada

[2]Department of Electrical Engineering and Computer Science, Technische Hochschule Lübeck, Lübeck, Germany

[3]Department of Anesthesiology, Pharmacology & Therapeutics, The University of British Columbia, Vancouver, BC, Canada

[4]Department of Pediatrics, The University of British Columbia, Vancouver, BC, Canada

Corresponding Author:
Matthias Görges, PhD
Department of Anesthesiology, Pharmacology & Therapeutics
The University of British Columbia
Rm V3-324, 950 West 28th Avenue
Vancouver, BC, V5Z 4H4
Canada
Phone: 1 6048752000 ext 5616
Fax: 1 6048752668
Email: mgorges@bcchr.ubc.ca

Abstract

Background: In the pediatric intensive care unit (PICU), quantifying illness severity can be guided by risk models to enable timely identification and appropriate intervention. Logistic regression models, including the pediatric index of mortality 2 (PIM-2) and pediatric risk of mortality III (PRISM-III), produce a mortality risk score using data that are routinely available at PICU admission. Artificial neural networks (ANNs) outperform regression models in some medical fields.

Objective: In light of this potential, we aim to examine ANN performance, compared to that of logistic regression, for mortality risk estimation in the PICU.

Methods: The analyzed data set included patients from North American PICUs whose discharge diagnostic codes indicated evidence of infection and included the data used for the PIM-2 and PRISM-III calculations and their corresponding scores. We stratified the data set into training and test sets, with approximately equal mortality rates, in an effort to replicate real-world data. Data preprocessing included imputing missing data through simple substitution and normalizing data into binary variables using PRISM-III thresholds. A 2-layer ANN model was built to predict pediatric mortality, along with a simple logistic regression model for comparison. Both models used the same features required by PIM-2 and PRISM-III. Alternative ANN models using single-layer or unnormalized data were also evaluated. Model performance was compared using the area under the receiver operating characteristic curve (AUROC) and the area under the precision recall curve (AUPRC) and their empirical 95% CIs.

Results: Data from 102,945 patients (including 4068 deaths) were included in the analysis. The highest performing ANN (AUROC 0.871, 95% CI 0.862-0.880; AUPRC 0.372, 95% CI 0.345-0.396) that used normalized data performed better than PIM-2 (AUROC 0.805, 95% CI 0.801-0.816; AUPRC 0.234, 95% CI 0.213-0.255) and PRISM-III (AUROC 0.844, 95% CI 0.841-0.855; AUPRC 0.348, 95% CI 0.322-0.367). The performance of this ANN was also significantly better than that of the logistic regression model (AUROC 0.862, 95% CI 0.852-0.872; AUPRC 0.329, 95% CI 0.304-0.351). The performance of the ANN that used unnormalized data (AUROC 0.865, 95% CI 0.856-0.874) was slightly inferior to our highest performing ANN; the single-layer ANN architecture performed poorly and was not investigated further.

Conclusions: A simple ANN model performed slightly better than the benchmark PIM-2 and PRISM-III scores and a traditional logistic regression model trained on the same data set. The small performance gains achieved by this two-layer ANN model may not offer clinically significant improvement; however, further research with other or more sophisticated model designs and better imputation of missing data may be warranted.

XSL•FO
RenderX

(JMIR Med Inform 2021;9(8):e24079) doi:10.2196/24079

KEYWORDS

artificial intelligence; risk assessment; decision support techniques; intensive care unit; pediatric; decision making; computer-assisted

Introduction

Background

The use of risk models in medicine enables timely and more targeted interventions for a given patient and facilitates benchmarking quality of care and conduct of clinical studies [1]. It is often necessary to quantify the severity of illness in the pediatric intensive care unit (PICU). Estimating the probability of mortality or expected length of stay from early admission data with such risk models is mainly used for quality improvement and benchmarking; however, it might enable a clinician to make objective medical decisions regarding the state of the patient, the necessary level of care, possible treatments, discharge plans, or expected costs [2-4].

PICUs are data-rich environments with a wide range of physiological variables that are responsive to interventions over short periods and outcomes that are well-defined and generally quantifiable [5]. Thus, the PICU provides fertile ground to develop and test prediction models of risks and outcomes. A score, which is quick and pragmatic to use, can enable the timely identification of adverse conditions and may be used to tailor appropriate interventions [6]. Two commonly encountered pediatric risk scores are the pediatric index of mortality 2 (PIM-2) [2] and pediatric risk of mortality III (PRISM-III) [1]. Both are derived from logistic regression models, which estimate mortality risk and have been validated with respective areas under the receiver operating characteristic curves (AUROCs) of 0.90 and 0.89 [1,7].

Increased computing capabilities, big data, and machine learning algorithms enable the application of artificial intelligence (AI) for clinical decision support [8]. Artificial neural networks (ANNs), a subtype of AI, can be used in different medical areas and have been shown to outperform physicians in diagnosis based on medical imaging or data from electronic medical records [9-12]. A recurrent neural network is a type of ANN that is most commonly used for sequential data. An ANN-based cardiac risk score, which used the recurrent neural network approach, was able to detect small changes in an electrocardiogram segment, which cannot be found by visual inspection [11]; another was used to classify clinical time series data for pediatric patients in critical care [12].

The clinical adoption of ANN-based risk models relies on gaining physicians' trust in the use of AI [13,14], which may include, but is not limited to, demonstrating better performance than traditional regression approaches.

Objectives

The primary aim of this study is to examine the performance of an ANN-based approach compared to that of traditional approaches based on logistic regression models when applied to estimating the risk of mortality in children admitted to PICU with suspected sepsis. We developed an ANN model using features required in the PIM-2 and PRISM-III models to predict mortality outcomes (died or survived) in a large North American registry data set and evaluated the ANN's performance using the AUROC. We compared its performance with the benchmark PIM-2 and PRISM-III scores, as well as a logistic regression model, trained on the study data set, which used the same features as PIM-2 and PRISM-III.

Methods

Study Design and Approval

In this study, we used data from a North American PICU registry to compare the performance of an ANN model with PIM-2 and PRISM-III scores. The data set was obtained from Virtual Pediatric Systems (VPS), LLC, a registry of prospectively collected records from 130 PICUs in the United States and Canada. This is a secondary analysis of data obtained for a different purpose—to develop a simple risk stratification score for children with sepsis [6]. Ethical approval for the study was obtained from the University of British Columbia/Children's and Women's Health Centre of British Columbia Research Ethics Board (H15-01398). The requirement for written informed consent was waived by the research ethics board, as this study was a secondary analysis of registry data. This manuscript has been prepared in accordance with the guidelines for Transparent Reporting if a multivariable prediction risk model for Individual Prognosis or Diagnosis.

As sepsis diagnosis might not necessarily be known or documented *at the time of admission* to the PICU, we identified all children in the VPS data set whose diagnostic codes *at discharge* exhibited evidence of an infection, and combined with their admission to the PICU, this provides a reasonably strong indication for sepsis. This allowed us to create a representative data set of children with a high likelihood of sepsis.

Study Data Set

Data Available for Analysis

The analyzed data set included data on PICU admissions between January 1, 2009, and December 31, 2014. Data were available from 102,945 children, of whom 4068 died (mortality rate 3.95%). Each entry included a variety of vital signs, laboratory tests, and other clinical information, including the variables required to calculate the PIM-2 and PRISM-III scores. The clinical data used in this analysis were solely from early admission to the PICU. Hence, the longer the length of stay, the less associated these predictors were with the outcome under investigation: mortality or survival at PICU discharge.

Although the variables for PIM-2 and PRISM-III were collected from the same source, these models captured data from different sampling windows. For any given PICU admission, the VPS data set provides a single measurement for each variable used

XSL•FO
RenderX

by these 2 risk scores as required for their respective calculations.

PRISM-III Variables and Sampling Window

PRISM-III uses the highest or lowest values of systolic blood pressure, heart rate, temperature, mental status, pupillary reflexes, acidosis, pH, P_{CO2}, total carbon dioxide (CO_2), Pa_{O2}, glucose, potassium, creatinine, blood urea nitrogen, white blood cell count, platelet count, and prothrombin time or partial thromboplastin time [1]. Values included were measured in the first 12 hours of PICU care; laboratory variables were also considered up to 2 hours before PICU admission.

PIM-2 Variables and Sampling Window

PIM-2 uses the first recorded values of systolic blood pressure, pupillary reaction to light, Pa_{O2}, base excess, early mechanical ventilation (yes or no), elective PICU admission (yes or no), admission following surgery (yes or no), admission following cardiopulmonary bypass, high-risk diagnoses (nine options: cardiac arrest preceding intensive care unit (ICU) admission, severe combined immune deficiency, leukemia or lymphoma after first induction, spontaneous cerebral hemorrhage, cardiomyopathy or myocarditis, hypoplastic left heart syndrome, HIV infection, liver failure as the main reason for ICU admission, or neurodegenerative disorder), and low-risk diagnoses (five options: main reason for ICU admission of asthma, bronchiolitis, croup, obstructive sleep apnea, or diabetic keto-acidosis) [2]. Values included were measured in the first hour of PICU care starting at the time of the first face-to-face meeting of the patient with a PICU team member.

Not all vital signs were collected routinely for every patient, so the data set was only sparsely populated, and the vital signs used for calculating PIM-2 and PRISM-III scores were incomplete in some cases, for example, the Glasgow Coma Score (mental status) was missing from 60.2% (61,976/102,945) of cases. In the calculation of both PIM-2 and PRISM-III scores, missing vital signs are taken as a sign of being normal, that is, healthy, as such tests were not ordered or performed by the PICU team [1,2]. For example, a missing Glasgow Coma Score is interpreted as indicating a normal mental status and is input to the model as such. This assumption is discussed further in the *Limitations* section.

Preprocessing

Preprocessing was performed in Python (v3.8.5; Python Software Foundation) to perform three tasks: (1) generate the training and test sets, (2) address missing values in the data set, and (3) generate new variables through data transformation.

Generation of Training and Test Sets

The total data set was initially divided into training and test sets using a stratified approach to ensure that the class ratio for mortality remained approximately equal for the training, test, and full data sets. ANN and logistic regression models were built on the training sets and evaluated on the test sets, and the results were compared against the PIM-2 and PRISM-III models. The overall data set was bootstrapped 100 times to generate the training and test sets.

Addressing Data Missingness Through Simple Substitution

The data set was examined for missing entries, and the missing values were imputed based on the feature type; specifically, the missing values in categorical features, such as pupillary reaction and coma status, were imputed using the most common value (mode). The missing values in numerical features, such as glucose or P_{CO2}, were imputed using the median value, as most of these features did not follow a normal distribution. Median and mode approaches were used to build imputation models and fill the missing values in the training set, and these imputation models were applied to the test set separately to avoid a data leakage problem.

Generation of New Variables Through Data Transformation

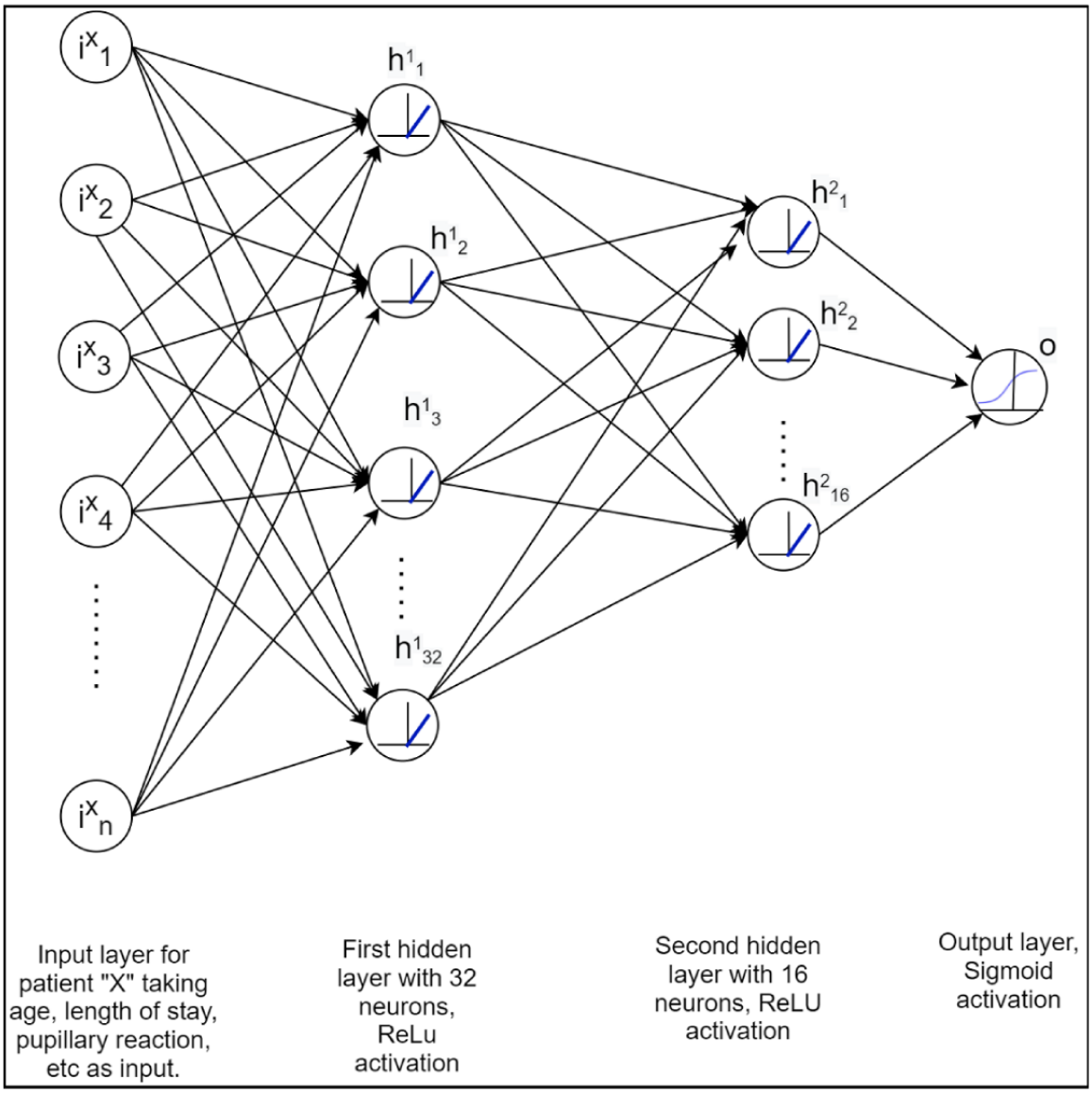
We performed minimum-maximum normalization to normalize numerical data for the ANN and logistic regression models. The minimum and maximum values of each feature from the training set were used to normalize the data in the training and test sets to avoid data leakage. Dummy encoding was performed on categorical features that contained more than 2 distinct values, such as pupillary reaction, but all categorical features with only 2 distinct values were dichotomized to accommodate them in the machine learning models. We used thresholds defined by PRISM-III to define normal and abnormal values. PIM-2 does not have defined thresholds; however, it penalizes any diversion of a vital sign from its normal value continuously.

Model Training

We built an ANN model using the Keras framework on top of TensorFlow (Google Brain Team) in Python (Python Software Foundation), with training conducted in Jupyter notebook (IPython). The Python code files that were used to build the models and generate results are available in Multimedia Appendix 1. We used a grid approach to determine the optimum configuration while designing the neural network. We tested various configurations between 1 and 3 hidden layers and 8 and 32 neurons per hidden layer, with a rule-of-thumb approach to limit the number of hidden layer neurons to the neurons in the input layer. Through experimentation, we identified that a 2-layer ANN, with 32 nodes in the first hidden layer and 16 nodes in the second hidden layer, performed better than the other configurations we tested. Our final model consisted of 32 input features (consisting of the variables used in the PRISM-III [1] and PIM-2 [2] models; see the *Study Data Set* section), and the 2 hidden layers, with each node using rectifier linear unit activation functions; finally, a sigmoid activated dense layer was used to predict the mortality for each instance (Figure 1). The model was compiled using an *adam* optimizer with a binary cross-entropy loss function. While keeping the main network the same, we also evaluated the model with unnormalized data as well as a model with only a single hidden layer. We conducted training with a batch size of 32 and observed that the loss remained constant after 100 epochs.

XSL•FO
RenderX

Figure 1. Artificial neural network architecture with two hidden layers: the node in input layer "iXy" processes data from pediatric intensive care unit admission "X" with feature "y" (such as age, length of stay, pupillary reaction, etc). The total number of features in the data set is denoted by "n." The first and second hidden layers are represented by h1 and h2, respectively, with a subscript to denote the node number. The output layer has a single node (o), which shows probability of mortality for patient "X." ReLU: rectifier linear unit.



The ANN model was trained with features used in the PIM-2 and PRISM-III models to predict the outcome (*died* or *survived*); AUROC was used as an evaluation metric while training the model. Finally, we developed a logistic regression model for comparison using the same features from PIM-2 and PRISM-III.

Model Evaluation

The empirical range of AUROC scores was computed for each test set (obtained from bootstrap) using the *sklearn.metrics* function in Python. The test set that resulted in the median AUROC value was used to determine the optimum Youden index value. This threshold was then used to calculate the false positive rate (FPR) and false negative rate (FNR) for each test set, and the 95% empirical CIs were reported by pooling the results from all the test sets [15-17]; median and ranges of pooled results were reported for all other indices. We also reported the area under the precision recall curve (AUPRC) and its empirical 95% CI for each model. A Welch 2-sided *t* test was used to compare AUROC and AUPRC for model pairs.

To compare how the models performed at specific true positive rate (TPR) and FPR levels, we fixed the TPR values at 95%, 90%, and 85% and computed the corresponding median FPR values (from all the test sets) for ANN, logistic regression, PIM-2, and PRISM-III. Similarly, we also reported the median TPR results by fixing the FPR at 5%, 10%, and 15%.

Results

Data Set Characteristics

The data set included 102,945 children with infection admitted between 2009 and 2014, of whom 4068 died (3.95% mortality rate). The training sets contained 72,061 children, of whom a median of 2852 (range 2790-2903) died, equivalent to a 3.96% mortality rate; the test sets contained 30,884 children, of whom a median of 1216 (range 1165-1278) died, equivalent to a 3.94% mortality rate (Table 1).

XSL•FO
RenderX

Table 1. Overview of study population with demographics and risk factors split by outcome (N=102,945)[a].

Characteristic	All (n=102,945)	Died (n=4068)	Survived (n=98,877)	Training		Testing	
				Died (n=2852)	Survived (n=69,209)	Died (n=1216)	Survived (n=29,668)
Males, n (%)	58,058 (56.39)	2186 (53.73)	55,872 (56.5)	1531 (53.68)	39,075 (56.46)	655 (53.87)	16,797 (56.62)
Age							
Age (months), median (IQR)	28.9 (7.5-100.3)	39.3 (7.0-137.7)	28.6 (7.5-98.7)	37.8 (6.9-138.4)	28.4 (7.3-98.4)	42.45 (7.2-135.25)	29.3 (7.9-99.4)
<1 month, n (%)	4733 (4.6)	342 (8.41)	4391 (4.44)	245 (8.59)	3108 (4.49)	97 (7.98)	1283 (4.32)
1-23 months, n (%)	42,935 (41.71)	1400 (34.41)	41,535 (42.01)	980 (34.36)	29,195 (42.18)	420 (34.54)	12,340 (41.59)
2-5 years, n (%)	22,264 (21.63)	719 (17.67)	21,545 (21.79)	510 (17.88)	14,967 (21.63)	209 (17.19)	6578 (22.17)
6-12 years, n (%)	18,652 (18.12)	776 (19.07)	17,876 (18.08)	542 (19)	12,474 (18.02)	234 (19.24)	5402 (18.21)
13-18 years, n (%)	14,353 (13.94)	829 (20.39)	13,524 (13.68)	574 (20.13)	9461 (13.67)	255 (20.97)	4063 (13.69)
>18 years, n (%)	8 (0.01)	2 (0.05)	6 (0.01)	1 (0.04)	4 (0.01)	1 (0.08)	2 (0.01)
Primary diagnosis category, n (%)							
Respiratory	63,928 (62.1)	1404 (34.51)	62,524 (63.23)	968 (33.94)	43,696 (63.14)	436 (35.86)	18,828 (63.46)
Infectious	12,288 (11.94)	1387 (34.1)	10,901 (11.02)	979 (34.33)	7588 (10.96)	408 (33.56)	3313 (11.17)
Neurological	3589 (3.49)	162 (3.98)	3427 (3.47)	120 (4.21)	2447 (3.54)	42 (3.45)	980 (3.3)
Gastrointestinal	2248 (2.18)	103 (2.53)	2145 (2.17)	79 (2.77)	1518 (2.19)	24 (1.97)	627 (2.11)
Dermatologic	1769 (1.72)	30 (0.74)	1739 (1.76)	18 (0.63)	1219 (1.76)	12 (0.99)	520 (1.75)
Location before PICU[b] admission, n (%)							
Inpatient	30,691 (29.81)	1752 (43.07)	28,939 (29.27)	1238 (43.41)	20,280 (29.3)	514 (42.27)	8659 (29.19)
Postoperative admission	18,435 (17.91)	576 (14.16)	17,859 (18.06)	412 (14.45)	12,447 (17.98)	164 (13.49)	5412 (18.24)
Resuscitation procedures							
Cardiac massage before PICU, n (%)	1863 (1.81)	588 (14.45)	1275 (1.29)	419 (14.69)	864 (1.25)	169 (13.9)	411 (1.39)
Mechanical ventilation within 24 hours, n (%)	53,903 (52.36)	3417 (84)	50,486 (51.06)	2405 (84.33)	35,258 (50.94)	1012 (83.22)	15,228 (51.33)
Mechanical ventilation within 1 hour, n (%)	42,940 (41.71)	2658 (65.34)	40,282 (40.74)	1863 (65.32)	28,161 (40.69)	795 (65.38)	12,121 (40.86)
Length of stay (days), median (IQR)	3.5 (1.7-8.0)	7.2 (2.2-21.4)	3.4 (1.7-7.8)	7.1 (2.2-21.0)	3.4 (1.7-7.8)	7.5 (2.4-22.7)	3.4 (1.6-7.8)
PRISM-III[c] probability of death (%), median (IQR)	0.63 (0.3-1.6)	10 (1.7-44.6)	0.5 (0.3-1.4)	10.2 (1.7-47.8)	0.5 (0.3-1.4)	8.3 (1.6-39.2)	0.5 (0.3-1.4)
PIM-2[d] probability of death (%), median (IQR)	1 (0.4-3.5)	5.2 (2.8-18)	1 (0.3-3.3)	5.3 (2.9-18.2)	0.9 (0.3-3.3)	4.8 (2.1-17.6)	0.9 (0.3-3.3)
Died, n (%)	4068 (3.95)	4068 (100)	0 (0)	2852 (100)	0 (0)	1216 (100)	0 (0)

[a]Data are reported separately for the complete population and the training and test cohorts. Note that only the top 5 primary diagnosis categories are reported. In addition, note that only the initial 3 columns are true results; the remaining 4 are median values over the 100 data sets created.

[b]PICU: pediatric intensive care unit.

[c]PRISM-III: pediatric risk of mortality III.

[d]PIM-2: pediatric index of mortality 2.

As is commonly encountered in large clinical registries using clinical availability of routinely collected data, between 0.41% (424/102,945) and 80.27% (82,636/102,945) of entries were missing per feature required for PRISM-III: more commonly measured vital signs, such as systolic blood pressure and heart rate, had fewer missing values (424/102,945, 0.41%-517/102,945, 0.5%), whereas others were missing many entries, such as CO_2 (49,837/102,945, 48.41% missing) and partial thromboplastin time (82,636/102,945, 80.27% missing). For PIM-2, only 2.05% (2115/102,945) of entries were missing

XSL•FO
RenderX

the numerical feature systolic blood pressure, whereas base excess was missing in 84.73% (87,230/102,945) of entries, and both fraction of inspired oxygen and Pa_{O2} were missing from 94.18% (96,958/102,945) of the entries. On the other hand, there was no missing information in any of the binary features, such as high- or low-risk diagnosis and recovery from surgery, which are features required for the PIM-2 calculation.

Model Performance: ANN Trained Using Imputed and Normalized Data

With the ANN trained on normalized data, the median FPR was mostly close to 18.4% (range 12.5%-30.8%) and the median FNR value was 24% (range 12.7%-33.2%; Table 2), with a median accuracy of 81.3% (range 69.9%-86.7%) on the test set.

Table 2. Performance characteristics of 4 different mortality prediction models[a].

Prediction model	Threshold trigger (%)	False positive detections (FPR[b]), n (%)	Missed cases (FNR[c]), n (%)
PIM-2[d]	3.36	7278 (24.5)	377 (31.5)
PRISM-III[e]	2.21	1052 (3.5)[f]	651 (54.3)
Logistic regression	0.50	5142 (17.3)	317 (26.8)
ANN[g]	0.04	5467 (18.4)	289 (24.0)[f]

[a]Comparison of the pediatric index of mortality 2, pediatric risk of mortality III, a traditional logistic regression model, and artificial neural network–based approach. For each model, the threshold was selected by optimizing the Youden index.

[b]FPR: false positive rate.

[c]FNR: false negative rate.

[d]PIM-2: pediatric index of mortality 2.

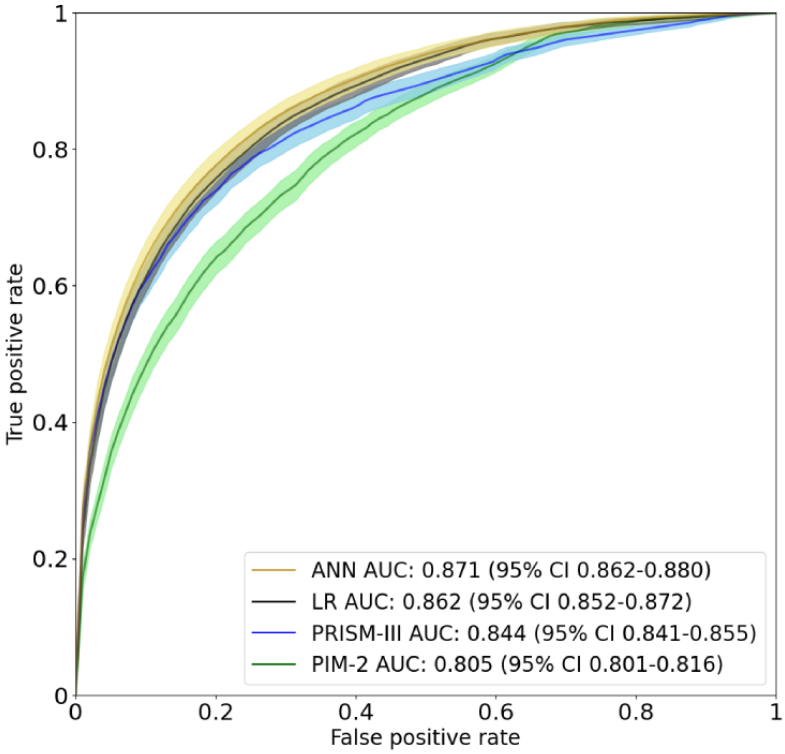
[e]PRISM-III: pediatric risk of mortality III.

[f]The best value in this category.

[g]ANN: artificial neural network.

The AUROCs for PIM-2 and PRISM-III were 0.805 (95% CI 0.801-0.816) and 0.844 (95% CI 0.841-0.855), respectively. The ANN (AUROC 0.871, 95% CI 0.862-0.880) performed better than both PIM-2 (P<.001) and PRISM-III (P<.001; Figure 2).

Figure 2. Receiver operating characteristic curves for four different mortality prediction models: pediatric index of mortality 2, pediatric risk of mortality III, logistic regression, and our best artificial neural network–based approach. The areas under the receiver operating characteristics curve and their 95% CI are indicated in the bottom-right corner. ANN: artificial neural network; AUC: area under the receiver operating characteristic curve; LR: logistic regression; PIM-2: pediatric index of mortality 2; PRISM-III: pediatric risk of mortality III.



XSL•FO
RenderX

Similar results were observed using AUPRC, which indicated that the ANN (AUPRC 0.372, 95% CI 0.345-0.396) performed better than PIM-2 (AUPRC 0.234, 95% CI 0.213-0.255; *P*<.001) and PRISM-III (AUPRC 0.348, 95% CI 0.322-0.367; *P*<.001; Figure 3). The ANN achieved the highest TPR compared with the logistic regression, PIM-2, and PRISM-III when FPR was fixed at 5%, 10%, or 15%. Similarly, FPR was lowest for the ANN when TPR was fixed at 85% or 90% (Table 3). However, the logistic regression model showed the smallest FPR when TPR was fixed at 95%.

Figure 3. Precision recall curves for four different mortality prediction models: pediatric index of mortality 2, pediatric risk of mortality III, logistic regression, and our best artificial neural network–based approach. The areas under the precision recall curves and their 95% CI, are indicated in the top right corner. ANN: artificial neural network; AUPRC: area under the precision recall curve; LR: logistic regression; PIM-2: pediatric index of mortality 2; PRISM-III: pediatric risk of mortality III.

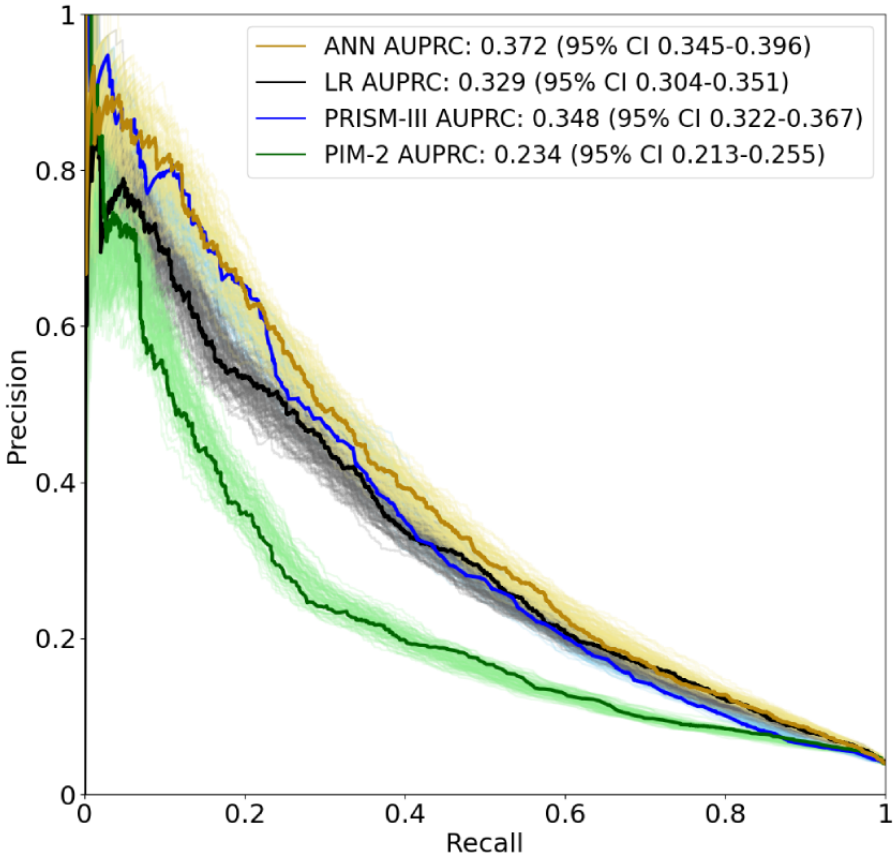


Table 3. Median true positive rate and median false positive rate of 4 different mortality prediction models[a].

Prediction model	TPR[b] (%)			FPR[c] (%)		
	FPR fixed at 5%	FPR fixed at 10%	FPR fixed at 15%	TPR fixed at 95%	TPR fixed at 90%	TPR fixed at 85%
PIM-2[d]	35.6	48.1	56.8	64.6	54	44.2
PRISM-III[e]	48.8	60.5	68.4	66.8	49.4	36.7
Logistic regression	48.4	61.2	69.8	55.1[f]	41.6	31.4
ANN[g]	49.7[f]	62.6[f]	70.8[f]	56	41.1[f]	30.7[f]

[a]Comparison of the pediatric index of mortality 2, pediatric risk of mortality III, a traditional logistic regression model, and artificial neural network–based approach.

[b]TPR: true positive rate.

[c]FPR: false positive rate

[d]PIM-2: pediatric index of mortality 2.

[e]PRISM-III: pediatric risk of mortality III.

[f]The best value in this category.

[g]ANN: artificial neural network.

XSL•FO
RenderX

The lowest FPR observed at the Youden-optimized threshold point for any of the models evaluated was 3.5% using PRISM-III, with a corresponding FNR of 54.3% (Table 2). If we target an FPR of 3.5%, the corresponding FNRs for the other models were 68.7% for PIM-2, 55.8% for the logistic regression, and 54% for the ANN.

Model Performance: Logistic Regression Using Imputed and Nonnormalized Data

The accuracy of the logistic regression model was 81.9% (range 81.4%-82.5%), with an FPR of 17.3% (range 17%-18.3%) and an FNR of 26.8% (range 23.7%-29.9%; Table 2). The AUROC was 0.862 (95% CI 0.852-0.872) and the AUPRC was 0.329 (95% CI 0.304-0.351). The logistic regression model also showed better performance, as measured by AUROC, than both PIM-2 (P<.001) and PRISM-III (P<.001; Figure 2), but PRISM-III performed better than the logistic regression model when evaluated using AUPRC (Figure 3).

Although the AUROCs of the ANN and logistic regression overlap, it was found that ANN performed better than logistic regression (P<.001).

Model Performance: ANN Trained Using Imputed and Nonnormalized Data

The accuracy of the ANN model trained using the nonnormalized data set with imputed data was 82.5% (range 69.6%-89.2%). The FPR value was 17.9% (range 9.7%-31.1%), and the FNR value was 26.7% (range 13.9%-39.3%; Table 2). The AUROC was 0.865 (95% CI 0.856-0.874), which was lower than that of the model with normalized data (P<.001). The AUPRC value was 0.355 (95% CI 0.328-0.376).

Using nonnormalized data, the ANN model had an FPR of 16.8% (95% CI 14.8%-18.2%) at a TPR of 73.3% and achieved its highest TPR of 73.4% (95% CI 71.4%-75.6%) for an FPR of 17.3%.

Discussion

Principal Findings

Summary of Results

We created an ANN-based pediatric risk prediction score using the features included in PIM-2 and PRISM-III scores, which we trained on patients from a large North American multicenter pediatric cohort with presumed sepsis as identified by a discharge diagnosis of infection. The overall performance of the ANN model with binary cross-entropy loss was better than the PIM-2 and PRISM-III scores, with median AUROCs of 0.871 (ANN) versus 0.805 (PIM-2; P<.001) and 0.844 (PRISM-III; P<.001). It also performed better than a traditional logistic regression model that used the same features required by PIM-2 and PRISM-III. However, these performance gains may not represent a clinically significant improvement. Our evaluation of the ANN approach with a single hidden layer and nonnormalized data returned poorer results than the other models evaluated.

Improved Performance, but Is It Relevant?

Our highest performing ANN was significantly better, statistically, than PIM-2 and PRISM-III using the AUROC and AUPRC measures of performance. The ANN missed fewer cases than PIM-2, PRISM-III, and the logistic regression model (ie, the ANN had a lower FNR; Table 2) at their respective ideal thresholds, as determined by optimizing their respective Youden indices; however, its rate of false positive detections was higher than that of PRISM-III and marginally higher than that of the logistic regression model (ie, the ANN had a higher FPR) at these Youden-optimized thresholds. This may suggest an opportunity for further optimization and evaluation, but it should be noted that the ANN did not miss more cases than PRISM-III (ie, the ANN had an equivalent FNR) when the FPR was fixed at the value of 3.5% (PRISM-III's Youden-optimized threshold). A direct comparison between models is challenging given that model selection will depend to a large extent on the clinical context; in some settings, a single objective (eg, to minimize FPR) may be the overriding concern, whereas in other cases, a balance of multiple objectives may be required (eg, to minimize both FPR and FNR).

Despite limited performance gains and increased robustness, the improvement may not be clinically relevant and is unlikely to overcome the initial concerns that physicians might have about the new model. The limited performance gains were not surprising. Although studies have proposed that ANNs outperform logistic regression models [12,18] or offer at least partially better performance [19], a recent systematic review of 71 studies found no superior performance of ANN over logistic regression models [20]. However, ANN-based models allow for the tuning of performance characteristics, which offers a potential advantage.

Trust Issues as a Barrier to ANN Use in Risk Modeling

The successful acceptance of AI-based risk models requires physicians' willingness to accept AI models and the interpretability of those models. Although clinically improved performance might help this case, trust is a key element in acceptance, which is built (or lost) in a dynamic and evolving process [13,21]. Our failure to demonstrate a significant improvement in clinical performance will not help overcome the barriers to adoption.

Future AI-based risk models may need to become more interpretable to find acceptance [14], and the higher the risk, the more interpretability is needed to earn the trust. Including clinicians and patients in the development of AI models may be a step toward promoting acceptability and interpretability. Certification and licensure for AI models might also help build trust in model-based risk scores [22,23]. Finally, it may be useful to assure the user that the model is a tool and not a replacement for the clinician [13].

Challenges With Skewed Data

The working data set was skewed: only 3.95% (4068/102,945) of instances had the outcome as *died*. Local minima are a problem frequently associated with imbalanced data sets, and customized learning algorithms, cost functions, or external approaches (ie, resampling the data set) can be used to help

XSL•FO
RenderX

overcome this problem [24]. Some ANNs tended to predict (mostly) everyone as a survivor; given the overall mortality rate of the population (4068/102,945, 3.95%), even assuming every patient will survive results in an accuracy of approximately 96%, but with an FPR of zero and an FNR of one. A traditional experimental setup with accuracy as an evaluation metric fails when building models with skewed data, as the models tend to be biased toward the majority class (here *survived*) [25]. This challenge can be addressed by modifying the cost function to maximize the AUROC of the model [25].

Limitations

The main limitation of this work is the fact that out of several ANN-based models evaluated, only 1 type learned to discriminate between survival and death of patients effectively. Despite attempts to address the root cause (imbalance of outcomes in the data set), this suggests that the approaches selected may not have been optimal and that further network types and designs should be considered in future approaches. Following the initial positive outcomes with this model, secondary training on a data set can be used to fine-tune the ANN model.

The information included in the new models was limited to risk factors from PIM-2 and PRISM-III. By creating new features such as vital sign combinations or ratios [26], which in principle can be emulated by adding hidden layers, one might be able to provide another significant performance boost to the model. However, this did not seem to be the case in a recent sepsis prediction competition [27], where novel methods or applications seemed to be more promising than the creation of new features.

Another limitation was the relatively low number of complete patient entries in the VPS data set. Given that VPS is a curated data set, the potential reasons for this likely stem from local practices, such as tests not being required for clinical management in particular cases or it being generally decided that recording the results of these tests is optional. Although it makes the creation and use of some modeling techniques more difficult, this is an unavoidable feature of real-world clinical data. Characterizing the missingness to inform modeling might offer a valuable approach, but such features may not be generalizable because they represent local patterns of practice. To use models without the filtering layer, simple imputation approaches were used; however, data were likely not missing at random, which invalidates some of the (median or mode imputation) approaches used. More sophisticated approaches for handling data missingness, such as multivariate imputation by chained equations, may yield better performance [28,29], as the substituted values are likely closer to specific cases than the overall population. Importantly, physicians should inform the treatment of missing values, which might boost confidence in the methods used. It might be possible to use a complete time series in an ANN instead of extreme values observed in a certain window, which could improve performance.

This study explored only a limited range of ANN design techniques. For example, we used rectifier linear unit activation in the hidden layers but did not evaluate the effect of other activation functions on model performance; similarly, we used the *adam* optimizer to identify the optimal ANN architecture but did not evaluate alternative optimizers. Thus, more exhaustive experimentation may yield improved performance results. Similarly, Youden index was used as a pragmatic approach to identify the optimal cut off by maximizing the models' true positive and true negative rates. However, selecting the appropriate operating point for clinical implementation should consider alternative approaches to finding the optimal threshold and would also require a more nuanced evaluation of clinical priorities, which might, for example, penalize missed cases over false positives.

A major limitation to the development of a new risk score is the lack of recognized clinically acceptable performance criteria to assess the utility of integrating ANN-based risk scores into daily clinical routines. In their absence, it is difficult to make a clear statement on the clinical utility of models with slightly better performance compared with existing risk scores.

Conclusions

This study examined the performance of ANN models over logistic regression-based models to estimate the risk of mortality in the PICU. A simple 2-layer ANN demonstrated better performance than traditional logistic regression, PIM-2, and PRISM-III; the statistically significant improvement in performance may not be clinically significant. Further work, including involvement of physicians in defining performance thresholds, better handling of data missingness, and possibly the use of more sophisticated ANN-modeling methods, will be required to achieve meaningful advances to guide decision-making in the care of critically ill children.

Acknowledgments

This study formed part of author NGP's master's thesis, for which he would like to thank Professor Jens Ehlers (Department of Electrical Engineering and Computer Science, Technische Hochschule Lübeck, Lübeck, Germany) for his guidance and input in supervising the thesis work. Data were provided by VPS, LLC. No endorsement or editorial restriction of the interpretation of these data or the opinions of the authors has been implied or stated. The authors wish to thank the VPS team for their support in obtaining the data. MG holds a Michael Smith Foundation for Health Research Scholar salary award, and RSS holds a Mitacs Postdoctoral Fellowship. This work was supported by a 2020 BC Children's Hospital Research Institute External Salary Recognition Award (to MG).

XSL•FO
RenderX

Authors' Contributions

SM and MG designed the study and obtained ethical approval to conduct the research. NGP initially analyzed the data with guidance from MG, and RSS revised the analysis. NGP, NCW, RSS, and MG interpreted the findings and drafted the manuscript. All authors critically reviewed the manuscript and read and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Zipped Python code files were used to build the models and generate the results.
[ZIP File (Zip Archive), 29 KB - medinform_v9i8e24079_app1.zip]

References

1. Pollack MM, Patel KM, Ruttimann UE. PRISM III: an updated Pediatric Risk of Mortality score. Crit Care Med 1996 May;24(5):743-752. [doi: 10.1097/00003246-199605000-00004] [Medline: 8706448]
2. Slater A, Shann F, Pearson G, Paediatric Index of Mortality (PIM) Study Group. PIM2: a revised version of the Paediatric Index of Mortality. Intensive Care Med 2003 Mar 23;29(2):278-285. [doi: 10.1007/s00134-002-1601-2] [Medline: 12541154]
3. Ong ME, Lee Ng CH, Goh K, Liu N, Koh Z, Shahidah N, et al. Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score. Crit Care 2012 Jul 21;16(3):R108 [FREE Full text] [doi: 10.1186/cc11396] [Medline: 22715923]
4. Marcin JP, Pollack MM. Review of the acuity scoring systems for the pediatric intensive care unit and their use in quality improvement. J Intensive Care Med 2007 Jun 30;22(3):131-140. [doi: 10.1177/0885066607299492] [Medline: 17562737]
5. Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton D, Clifford GD. Machine learning and decision support in critical care. Proc IEEE 2016 Feb;104(2):444-466. [doi: 10.1109/jproc.2015.2501978]
6. Peters C, Murthy S, Brant R, Kissoon N, Görges M. Mortality risk using a pediatric quick sequential (Sepsis-related) organ failure assessment varies with vital sign thresholds. Pediat Crit Care Med 2018;19(8):394-402. [doi: 10.1097/pcc.0000000000001598]
7. Shann F, Pearson G, Slater A, Wilkinson K. Paediatric index of mortality (PIM): a mortality prediction model for children in intensive care. Intensive Care Med 1997 Mar;23(2):201-207. [doi: 10.1007/s001340050317] [Medline: 9069007]
8. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan 7;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
9. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec 21;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
10. 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]
11. Myers PD, Scirica BM, Stultz CM. Machine learning improves risk stratification after acute coronary syndrome. Sci Rep 2017 Oct 04;7(1):12692 [FREE Full text] [doi: 10.1038/s41598-017-12951-x] [Medline: 28978948]
12. Aczon M, Ledbetter D, Ho L, Gunny A, Flynn A, Williams J. Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks preprint. arXiv : Statistics - Machine Learning. 2017. URL: https://arxiv.org/abs/1701.06675 [accessed 2021-07-31]
13. Siau K, Wang W. Building trust in artificial intelligence, machine learning, and robotics. Cut Bus Technol J 2018;31(2):53 [FREE Full text]
14. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018 Apr 04;15(141):20170387 [FREE Full text] [doi: 10.1098/rsif.2017.0387] [Medline: 29618526]
15. Bouckaert R, Frank E. Evaluating the replicability of significance tests for comparing learning algorithms. In: Dai H, Srikant R, Zhang C, editors. Advances in Knowledge Discovery and Data Mining. Berlin: Springer; 2004:3-12.
16. Kuhn M, Johnson K. Applied Predictive Modeling. New York, NY: Springer; 2013:1-600.
17. Cohen P. Empirical Methods for Artificial Intelligence. Cambridge, MA: MIT Press; 1995:1-422.
18. Shi H, Lee K, Lee H, Ho W, Sun D, Wang J, et al. Comparison of artificial neural network and logistic regression models for predicting in-hospital mortality after primary liver cancer surgery. PLoS One 2012 Apr 26;7(4):e35781 [FREE Full text] [doi: 10.1371/journal.pone.0035781] [Medline: 22563399]
19. Eftekhar B, Mohammad K, Ardebili HE, Ghodsi M, Ketabchi E. Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC Med Inform Decis Mak 2005 Mar 15;5(1):3 [FREE Full text] [doi: 10.1186/1472-6947-5-3] [Medline: 15713231]

XSL•FO
RenderX

20. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019 Jun;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
21. Hengstler M, Enkel E, Duelli S. Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. Technol Forecast Soc Change 2016 Apr;105:105-120. [doi: 10.1016/j.techfore.2015.12.014]
22. LaRosa E, Danks D. Impacts on trust of healthcare AI. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018 Presented at: AIES '18: AAAI/ACM Conference on AI, Ethics, and Society; February 2 - 3, 2018; New Orleans LA USA p. 210-215. [doi: 10.1145/3278721.3278771]
23. Proposed regulatory framework for modifications to Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device (SaMD). US Food and Drug Administration. 2019. URL: https://www.fda.gov/media/122535/download [accessed 2021-07-31]
24. Estabrooks A, Jo T, Japkowicz N. A multiple resampling method for learning from imbalanced data sets. Computational Intell 2004 Feb;20(1):18-36. [doi: 10.1111/j.0824-7935.2004.t01-1-00228.x]
25. He H, Ma Y. Imbalanced Learning: Foundations, Algorithms, and Applications. Hoboken, NJ: Wiley; 2013:1-216.
26. Pawar R, Bone J, Ansermino M, Görges M. An algorithm for early detection of sepsis using traditional statistical regression modeling. Comput Cardiol 2019;46:1-4 [FREE Full text] [doi: 10.22489/cinc.2019.061]
27. Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, et al. Early prediction of sepsis from clinical data. Crit Care Med 2020;48(2):210-217. [doi: 10.1097/ccm.0000000000004145]
28. Buuren SV, 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. Kenward MG, Carpenter J. Multiple imputation: current perspectives. Stat Methods Med Res 2007 Jul 02;16(3):199-218. [doi: 10.1177/0962280206075304] [Medline: 17621468]

Abbreviations

AI: artificial intelligence
ANN: artificial neural network
AUPRC: area under the precision recall curve
AUROC: area under the receiver operating characteristic curve
FNR: false negative rate
FPR: false positive rate
ICU: intensive care unit
PICU: pediatric intensive care unit
PIM-2: pediatric index of mortality 2
PRISM-III: pediatric risk of mortality III
TPR: true positive rate
TRIPOD: Transparent Reporting if a multivariable prediction risk model for Individual Prognosis or Diagnosis
VPS: Virtual Pediatric Systems

Edited by C Lovis; submitted 02.09.20; peer-reviewed by M Aczon, S Fuglerud; comments to author 18.11.20; revised version received 06.04.21; accepted 10.07.21; published 31.08.21.

Please cite as:
Ghanad Poor N, West NC, Sreepada RS, Murthy S, Görges M
An Artificial Neural Network–Based Pediatric Mortality Risk Score: Development and Performance Evaluation Using Data From a Large North American Registry
JMIR Med Inform 2021;9(8):e24079
URL: https://medinform.jmir.org/2021/8/e24079
doi:10.2196/24079
PMID:34463636

©Niema Ghanad Poor, Nicholas C West, Rama Syamala Sreepada, Srinivas Murthy, Matthias Görges. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 31.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Predicting Antituberculosis Drug–Induced Liver Injury Using an Interpretable Machine Learning Method: Model Development and Validation Study

Tao Zhong[1*], BSc; Zian Zhuang[2,3,4*], BSc; Xiaoli Dong[5,6], PhD; Ka Hing Wong[5,6], PhD; Wing Tak Wong[5,6], PhD; Jian Wang[1], BSc; Daihai He[2,4], PhD; Shengyuan Liu[1], PhD

[1]Department of Tuberculosis Control, Shenzhen Nanshan Center for Chronic Disease Control, Shenzhen, China
[2]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China
[3]Department of Biostatistics, University of California, Los Angeles, CA, United States
[4]Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China
[5]Research Institute for Future Food, The Hong Kong Polytechnic University, Hong Kong, China
[6]Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hong Kong, China
[*]these authors contributed equally

Corresponding Author:
Shengyuan Liu, PhD
Department of Tuberculosis Control
Shenzhen Nanshan Center for Chronic Disease Control
Hua Ming Road No 7
Nanshan District
Shenzhen, 518000
China
Phone: 86 13543301395
Email: jfk@sznsmby.com

Related Article:

Correction of: https://medinform.jmir.org/2021/7/e29226

(JMIR Med Inform 2021;9(8):e32415) doi:10.2196/32415

In "Predicting Antituberculosis Drug–Induced Liver Injury Using an Interpretable Machine Learning Method: Model Development and Validation Study" (JMIR Med Inform 2021;9(7):e29226) two corrections were made.

1. In the originally published article, author *Daihai He* was listed as the corresponding author. The corresponding author has been changed to *Shengyuan Liu* and the corrected details are as follows:

> ***Corresponding Author:***
> *Shengyuan Liu, PhD*
> *Department of Tuberculosis Control*
> *Shenzhen Nanshan Center for Chronic Disease Control*
> *Nanshan District*
> *Hua Ming Road No 7*
> *Shenzhen, 518000*
> *China*
> *Phone: 86 13543301395*
> *Email: jfk@sznsmby.com*

2. For authors *Xiaoli Dong*, *Ka Hing Wong*, and *Wing Tak Wong*, the affiliation was originally listed as follows:

> *Xiaoli Dong[2], PhD; Ka Hing Wong[2], PhD; Wing Tak Wong[2]*
> *Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong, China*

This has been corrected to:

> *Xiaoli Dong[5,6], PhD; Ka Hing Wong[5,6], PhD; Wing Tak Wong[5,6]*
> *[5]Research Institute for Future Food, The Hong Kong Polytechnic University, Hong Kong, China*
> *[6]Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hong Kong, China*

The complete author information and affiliations in the corrected article are listed below.

> *Tao Zhong[1]*, BSc; Zian Zhuang[2,3,4]*, BSc; Xiaoli Dong[5,6], PhD; Ka Hing Wong[5,6], PhD; Wing Tak*

XSL•FO
RenderX

Wong[5,6], PhD; Jian Wang[1], BSc; Daihai He[2,4], PhD; Shengyuan Liu[1], PhD

[1]Department of Tuberculosis Control, Shenzhen Nanshan Center for Chronic Disease Control, Shenzhen, China

[2]Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong, China

[3]Department of Biostatistics, University of California, Los Angeles, CA, United States

[4]Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China

[5]Research Institute for Future Food, The Hong Kong Polytechnic University, Hong Kong, China

[6]Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hong Kong, China

** these authors contributed equally*

The correction will appear in the online version of the paper on the JMIR Publications website on August 13, 2021, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 27.07.21; this is a non–peer-reviewed article;accepted 06.08.21; published 13.08.21.

Please cite as:
Zhong T, Zhuang Z, Dong X, Wong KH, Wong WT, Wang J, He D, Liu S
Correction: Predicting Antituberculosis Drug–Induced Liver Injury Using an Interpretable Machine Learning Method: Model Development and Validation Study
JMIR Med Inform 2021;9(8):e32415
URL: https://medinform.jmir.org/2021/8/e32415
doi:10.2196/32415
PMID:34398802

©Tao Zhong, Zian Zhuang, Xiaoli Dong, Ka Hing Wong, Wing Tak Wong, Jian Wang, Daihai He, Shengyuan Liu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 13.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities

Muhammad Ayaz[1], MSc; Muhammad F Pasha[1], PhD; Mohammed Y Alzahrani[2], PhD; Rahmat Budiarto[3], PhD; Deris Stiawan[4], PhD

[1]Malaysia School of Information Technology, Monash University, Bandar Sunway, Malaysia
[2]Information Technology Department, College of Computer Science & Information Technology, Albaha University, Albaha, Saudi Arabia
[3]Informatics Department, Faculty of Science & Technology, Universitas Alazhar Indonesia, Jakarta, Indonesia
[4]Department of Computer Engineering, Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia

Corresponding Author:
Muhammad Ayaz, MSc
Malaysia School of Information Technology
Monash University
Jalan Lagoon Selatan
Bandar Sunway, 47500
Malaysia
Phone: 60 0355146224
Email: Muhammad.ayaz@monash.edu

Related Article:

Correction of: https://medinform.jmir.org/2021/7/e21929

(JMIR Med Inform 2021;9(8):e32869) doi:10.2196/32869

In "The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities" (JMIR Med Inform 2021;9(7):e21929) the authors noted one error.

The title of the originally published article contained an error in the abbreviation "FHIR." The title originally read as follows:

> *The Fast Health Interoperability Resources (FIHR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities*

In the corrected version, the title has been revised to:

> *The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities*

The correction will appear in the online version of the paper on the JMIR Publications website on August 17, 2021, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 12.08.21; this is a non–peer-reviewed article;accepted 12.08.21; published 17.08.21.

Please cite as:
Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D
Correction: The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities
JMIR Med Inform 2021;9(8):e32869
URL: https://medinform.jmir.org/2021/8/e32869
doi:10.2196/32869
PMID:34403353

XSL•FO
RenderX

©Muhammad Ayaz, Muhammad F Pasha, Mohammed Y Alzahrani, Rahmat Budiarto, Deris Stiawan. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 17.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Factors to Effective Telemedicine Visits During the COVID-19 Pandemic: Cohort Study

Kristin Nicole Gmunder[1], MS; Jose W Ruiz[2], MD; Dido Franceschi[3], MD; Maritza M Suarez[4], MD

[1]University of Miami Miller School of Medicine, Miami, FL, United States
[2]Department of Otolaryngology, University of Miami Miller School of Medicine, Miami, FL, United States
[3]Department of Surgery, University of Miami Miller School of Medicine, Miami, FL, United States
[4]Department of Medicine, University of Miami Miller School of Medicine, Miami, FL, United States

Corresponding Author:
Kristin Nicole Gmunder, MS
University of Miami Miller School of Medicine
1600 NW 10th Ave #1140
Miami, FL, 33136
United States
Phone: 1 908 635 9107
Email: kgmunder@med.miami.edu

Abstract

Background: With COVID-19 there was a rapid and abrupt rise in telemedicine implementation often without sufficient time for providers or patients to adapt. As telemedicine visits are likely to continue to play an important role in health care, it is crucial to strive for a better understanding of how to ensure completed telemedicine visits in our health system. Awareness of these barriers to effective telemedicine visits is necessary for a proactive approach to addressing issues.

Objective: The objective of this study was to identify variables that may affect telemedicine visit completion in order to determine actions that can be enacted across the entire health system to benefit all patients.

Methods: Data were collected from scheduled telemedicine visits (n=362,764) at the University of Miami Health System (UHealth) between March 1, 2020 and October 31, 2020. Descriptive statistics, mixed effects logistic regression, and random forest modeling were used to identify the most important patient-agnostic predictors of telemedicine completion.

Results: Using descriptive statistics, struggling telemedicine specialties, providers, and clinic locations were identified. Through mixed effects logistic regression (adjusting for clustering at the clinic site level), the most important predictors of completion included previsit phone call/SMS text message reminder status (confirmed vs not answered) (odds ratio [OR] 6.599, 95% CI 6.483-6.717), MyUHealthChart patient portal status (not activated vs activated) (OR 0.315, 95% CI 0.305-0.325), provider's specialty (primary care vs medical specialty) (OR 1.514, 95% CI 1.472-1.558), new to the UHealth system (yes vs no) (OR 1.285, 95% CI 1.201-1.374), and new to provider (yes vs no) (OR 0.875, 95% CI 0.859-0.891). Random forest modeling results mirrored those from logistic regression.

Conclusions: The highest association with a completed telemedicine visit was the previsit appointment confirmation by the patient via phone call/SMS text message. An active patient portal account was the second strongest variable associated with completion, which underscored the importance of patients having set up their portal account before the telemedicine visit. Provider's specialty was the third strongest patient-agnostic characteristic associated with telemedicine completion rate. Telemedicine will likely continue to have an integral role in health care, and these results should be used as an important guide to improvement efforts. As a first step toward increasing completion rates, health care systems should focus on improvement of patient portal usage and use of previsit reminders. Optimization and intervention are necessary for those that are struggling with implementing telemedicine. We advise setting up a standardized workflow for staff.

(JMIR Med Inform 2021;9(8):e27977) doi:10.2196/27977

KEYWORDS

telemedicine; COVID-19; patient portals; delivery of health care; telehealth; pandemic; digital health

XSL•FO
RenderX

Introduction

Background

With the rise of COVID-19 in the United States, there was a dramatic increase and widespread utilization of telemedicine—a technology that has existed for decades but represented a small fraction of care across US health systems. Telemedicine's impetus began with National Aeronautics and Space Administration (NASA) needing to monitor the vital signs of its astronauts during manned space flights [1]. In the 1960s and 1970s, the US government funded research programs to expand telemedicine to rural areas due to a provider shortage [1]. Additional government expenditures were put toward a NASA-sponsored pilot program termed Space Technology Applied to Rural Papago Advanced Health Care (STARPAHC) that monitored Papago Indians in Arizona [1]. This demonstrated the feasibility of using the technology to provide geographically distant health care. In more recent times, Kaiser Permanente has set up, "an integrated delivery system that implemented video-visit capability for all clinicians in 2014," allowing for use of this technology across their health system [2]. Their model demonstrated the usability of this technology to "extend established patient–physician relationships" [2]. Looking at telemedicine use beyond just the US borders, the Ontario Telemedicine Network has been one of the largest providers of telemedicine services in the world [3]. One of its aims was to increase access to underserved areas over large geographical distances, mirroring NASA's original goals to expand access to Papago Indians. However, overall, telemedicine has been used sporadically in the United States, without major widespread adoption. With the onset of COVID-19, the health care landscape changed dramatically with patients avoiding physicians' offices.

In order to provide quality care in an environment that allowed for social distancing and convenience, health care providers embraced the use of telemedicine. The quick scale-up of telemedicine required overcoming several barriers to acceptance and widespread usage. By expanding coverage and reimbursement, the Center for Medicare and Medicaid Services (CMS) addressed one of these issues when it announced on March 30, 2020, that it would begin covering telehealth at the same rates as in-person visits for a variety of services [4]. Other commercial insurance carriers quickly enacted similar policies; this improved reimbursement of telemedicine facilitated quick embracement of telemedicine by health care providers [4].

In addition to insurance changes, there were also Health Insurance Portability and Accountability Act (HIPAA) leniencies which allowed for more video application options to better facilitate rapid transitions to telemedicine. HIPAA enforcement was temporarily relaxed during the public health emergency (PHE), allowing providers to utilize video-calling apps such as FaceTime, Google Hangouts, and Skype, provided they were not public facing [4]. Specifically, the Office for Civil Rights (OCR) at the HHS stated they would not enforce a fine for violating HIPAA rules regarding the use of these non-public-facing audio/video applications during the COVID-19 PHE [5]. The OCR also listed vendors that claim to provide HIPAA-compliant communication including Zoom for health care [5].

Beyond the economic and HIPAA-related issues, there were further barriers to widespread implementation of telehealth by providers. Technical issues, organizational issues, and behavioral issues all played a role in reduced acceptance of telemedicine technology [6]. Many health care providers were not comfortable in acquiring and customizing this new technology workflow, nor were they sufficiently experienced in troubleshooting problems with it. Providers and support staff may not have had the time or inclination to develop the appropriate process for utilizing the technology, which typically requires organizational leadership and support. Finally, there was a challenge in terms of human behavior change. Some health care providers preferred continuing with historical procedures rather than changing their activities. While the financial and privacy-related issues were addressed, there remained these technical, organizational, and behavioral hurdles to full adoption of telemedicine by health care providers.

Regardless of the challenges in providers' acceptance, COVID-19 brought about a rapid and unforeseen rise in telemedicine implementation for health systems. This left insufficient time for providers or patients to adapt. A recent report found that "Nearly half (43.5%) of Medicare primary care visits were provided via telehealth in April, compared with less than one percent before the PHE in February (0.1%)" [7]. A similar dramatic increase in telemedicine usage was also experienced at our institution, the University of Miami Health System (UHealth). Rapid scale-up of telehealth at UHealth occurred during the early months of COVID-19, rising to a peak of 14,852 visits per week in May, compared with an average of 17 visits per week from January until early March 2020 (Figure 2).

Regarding previous literature addressing telemedicine completion rates, some studies have examined demographics associated with completing telemedicine visits. One such study found that only 46% of scheduled patients had completed their visit, with 54% canceling or not showing. Female, non-English–speaking, older, and poorer patients in this study group had lower odds ratios (ORs) associated with telemedicine completion [8]. An additional study found that 54.4% of patients completed telemedicine visits, with older patients, Asians, non-English–speaking patients, and Medicaid-insured patients having fewer completed visits [9]. Additionally, other studies have been performed only examining no-show rates of telemedicine visits, instead of overall completion rates. While this does not directly compare with overall completion rates, no-show rates are a subset of the "incomplete" visit group. One study that had begun before COVID was able to examine no-show rates pre-COVID compared with post-COVID. They found comparable rates for telemedicine visits (9.1% pre-COVID and 8.9% post-COVID). In comparison to in-person rates for this study group, in-person no-show rates were 13.6% (in 2018) and 14.4% (in 2019) [10]. Overall, there is limited research with large-scale data sets into completed telemedicine visit rates and factors associated with them. However, we do know that telemedicine users (from pre-COVID studies) have tended to be younger, female, and live in urban areas [11].

XSL•FO
RenderX

Additionally, patients with "technology access (patients living in a neighborhood with high rates of residential internet access[)] were more likely to choose a video visit than patients whose neighborhoods had low internet access" and patients with "in-person visit barriers (patients whose clinic had a paid parking structure[)] were more likely to choose a telemedicine visit than patients whose facility had free parking." [12].

When examining factors associated with telemedicine visit completion not necessarily related to patient demographics (ie, provider specialty or previsit reminder notifications), there is limited research investigating these strictly in relation to telemedicine. Research done pre-COVID found that visit reminders (whether automated or done by clinic staff) resulted in lower no-show rates for in-person visits [13]. Patient portal use has also been associated with improved appointment adherence and a reduction in no-show rates [14]. Provider specialties have seen differences in no-show rates throughout multiple studies conducted on different patient populations [15]. Looking at new patient appointments versus follow-up appointments, one study found a significant difference between the rate of no-shows for new patients (30%) compared with follow-up patients (21%) for in-person appointments scheduled within 30 days [16].

With all of this previous literature in mind, we hypothesized that, of course, there would be demographic drivers of differences in completion rates. However, we also conceptualized that visit reminders, patient portal use, provider specialty, and visit type (new patient vs follow-up) would likely play a role in not only no-show rates, but also overall completion rates (as no-shows would comprise part of the incomplete visit group). There is really a limited analysis of overall telemedicine completion rates in terms of characteristics that are not necessarily demographically linked. Thus, there is a need for large-scale studies that focus on aspects that could affect a wide variety of health systems that may serve different patient demographics.

Goal of This Study

As the demand for telemedicine is likely to continue in the future, it is crucial to gain a better understanding of how to ensure completed telemedicine visits in our health system. Identifying variables that may affect telemedicine completion rates is necessary for a proactive approach to addressing various issues. While there are demographically generated disparities among patients in access to telemedicine (ie, race, ethnicity, or age that may affect access), the focus of this analysis is to highlight those changes that are actionable (ie, patient portal activation status) and can be enacted across the entire health system, regardless of the demographics of the population served.

Methods

Telemedicine at UHealth

UHealth main campus (located in Miami-Dade County) includes a 560-bed hospital, outpatient clinics, Sylvester Comprehensive Cancer Center, and Bascom-Palmer Eye Hospital [17]. The main campus serves a wide population from all over South Florida, but Miami-Dade County, specifically, has a population of 2,716,940 and is almost entirely classified as urban. About 69% of the population in Miami-Dade County is Hispanic and 13% are non-Hispanic Whites [18]. Additionally, in terms of satellite clinics, there are over 30 outpatient centers in Miami-Dade, Broward, Palm Beach, and Collier counties [17]. The populations in these other counties are lower (Collier County only has 384,902 people) and are more diverse in their rural–urban classification [18]. Additionally, the demographics of the satellite clinics are different in those counties outside of Miami-Dade with a lower percentage of Hispanics (23%-31%), a higher percentage of non-Hispanic Whites (35%-62%), and higher socioeconomic status [18].

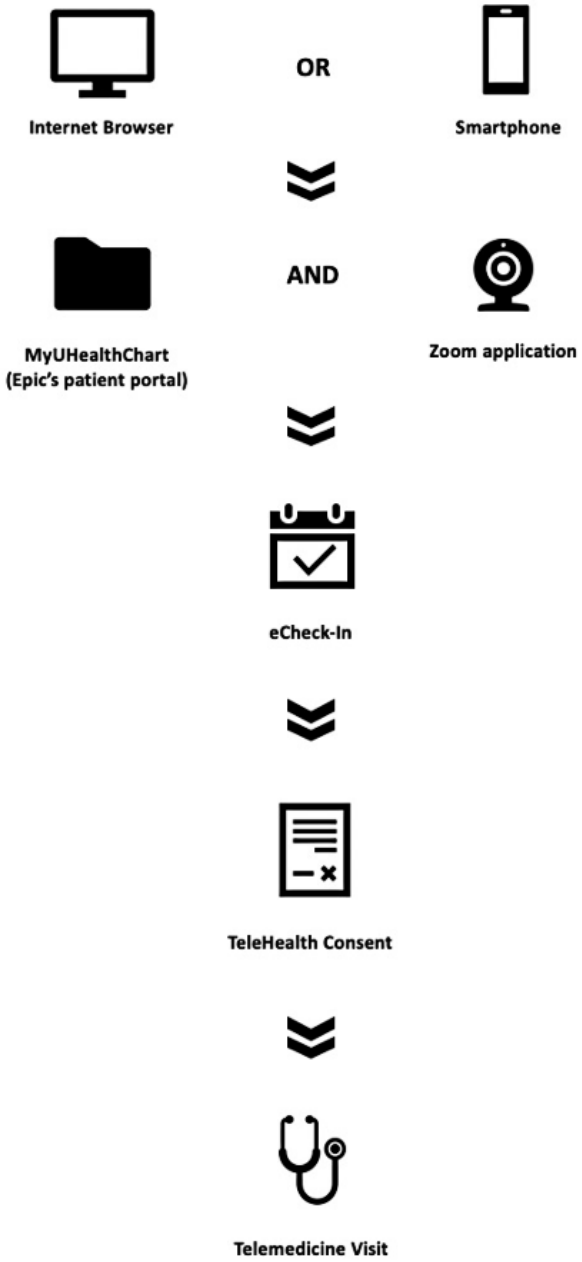
Within the UHealth system, Epic (Epic Systems Corporation) is used as the electronic medical record with the additional "MyUHealthChart" patient portal application. Within MyUHealthChart, patients are able to communicate (message) with their providers, view previous visit notes, examine tests results, schedule appointments, and upload Radiology images. These functions are in addition to the administrative purposes of viewing/paying bills. Specifically, to participate in a telemedicine visit at our institution, patients must go through this patient portal and perform several steps (Figure 1). Patients must first be signed up and registered for a MyUHealthChart account and log onto the patient portal via an internet browser or smartphone app. After logging into MyUHealthChart, patients must navigate to their visit and complete the eCheck-In. If it is the first visit, a consent for TeleHealth Services must be signed. Patients must download the Zoom application and click their appointment in MyUHealthChart to get to their telemedicine video visit via Zoom. In MyUHealthChart, there are videos and a guide to help patients navigate to their visit. Also, patients receive an automated appointment reminder before the visit by phone call or SMS text message based on their preferred communication method. Within MyUHealthChart, patients have access to a designated technical support number for telemedicine visit questions or troubleshooting before or during their visit. Also, the workflow is reviewed with patients on the phone with staff when scheduling the appointment and just prior to the scheduled visit. During scheduled Zoom appointments, providers are able to conduct a patient interview, but measurements (eg, blood pressure, electrocardiogram) are unable to be performed remotely. If a patient is unable to successfully access their telemedicine visit via the designated Zoom workflow within MyUHealthChart, the visit is completed via an alternative workflow such as Doximity or via a phone call (without video). Often times, this alternative workflow can occur in those that have not activated their MyUHealthChart.

In terms of implementation from the providers' perspective, many people, processes, and technologies were organized to rapidly scale-up and expand UHealth telemedicine services. Successful telemedicine implementation resulted from multiple factors such as some providers having previously provided telemedicine services, an IT group with experience in agile workflow for quick project turnaround, and buy-in from organizational leadership. Many providers' in-person clinics were closed by the pandemic, which allowed the associated clinic support staff to assist providers in their virtual clinics. There was not one mandated workflow, but instead there were

XSL•FO
RenderX

guidelines and best practices, along with constant multimodal communication on the quickly developing policies and processes. These factors and the interest of administration and clinical workers to do what was best for the patient drove UHealth to rapidly and successfully implement a long-term strategy of telemedicine services.

Figure 1. Telemedicine workflow for patients in the UHealth system. Patients must have access to an internet browser or smartphone to access MyUHealthChart and Zoom. Next, they must complete eCheck-In and TeleHealth consent prior to joining their telemedicine visit.



Clinical Data Collection

A clinical data request was made for all scheduled telemedicine visits (N=382,076) between January 1, 2020, and October 31, 2020. Deidentified patient-specific variables collected included age, race, ethnicity, sex, insurance, preferred language, and zip code (used to estimate income via an external data set [19]). Health system predictors collected were provider specialty, clinic location, name of provider, MyUHealthChart activation status, previsit phone/SMS text message confirmation status, new to the provider, and new to the UHealth system. All of the data were captured from the Epic system and transferred into the Clarity database, where it was pulled into exportable data sets. Data that were erroneous or had greater than 50% of the data points missing were excluded from the analysis (n=12,410). Unscheduled or "on-the-fly" telemedicine visits (n=6743) were also excluded. Deleted observations were analyzed to ensure there was no significant association (P>.05) between missing data and either of the completion status groups. The telemedicine visit was classified as completed if appointment status was either arrived or completed and the billing code was not null, erroneous, incomplete video, or patient left without being seen.

Statistical Analysis

The data set was analyzed using RStudio 1.2.1335 [20] with additional packages (*furniture* [21], *lme4* [22], *ROCR* [23], and

XSL•FO
RenderX

randomForest [24]), and visualizations were created in Tableau 2020.3.2 [25]. Data before March 1, 2020 (first officially reported COVID case in Florida) [26] were excluded from statistical tests (n=159). For descriptive statistics, continuous variables were analyzed with *t* tests and categorical variables with chi-square tests. A Bonferroni correction was utilized to adjust for multiple comparisons within descriptive statistics (α=0.05/14=.0036). Mixed effects logistic regression was used to model the completion status outcomes (α=.05) and to identify the most important system-wide hurdles to telemedicine completion. This method was used to adjust for clustering at the clinical site level, as clinical site, with 51 unique levels, was used as a random effect. The model initially included all collected patient demographic characteristics (age, race, ethnicity, sex, insurance, preferred language, estimated income, religion) that might have been possible confounders in addition to patient-agnostic variables (provider specialty, MyUHealthChart activation status, previsit phone call/SMS text message confirmation status, new to the provider, and new to the UHealth system). Using comparison of model fit statistics (Akaike information criterion and Bayesian information criterion), the model was optimized. Continuous variables were also scaled prior to modeling. Random forest was used as an additional method to examine the importance of predictors using the "importance" function to compare mean decrease Gini and mean decrease accuracy. To determine the predictive capabilities of both the logistic regression model and the random forest model, the data set was divided into a test and training set (with equal distribution of completion status between the 2 sets). Accuracy and area under the curve were assessed for both models. Data visualizations were made for individual specialties, clinics, and providers for internal use.

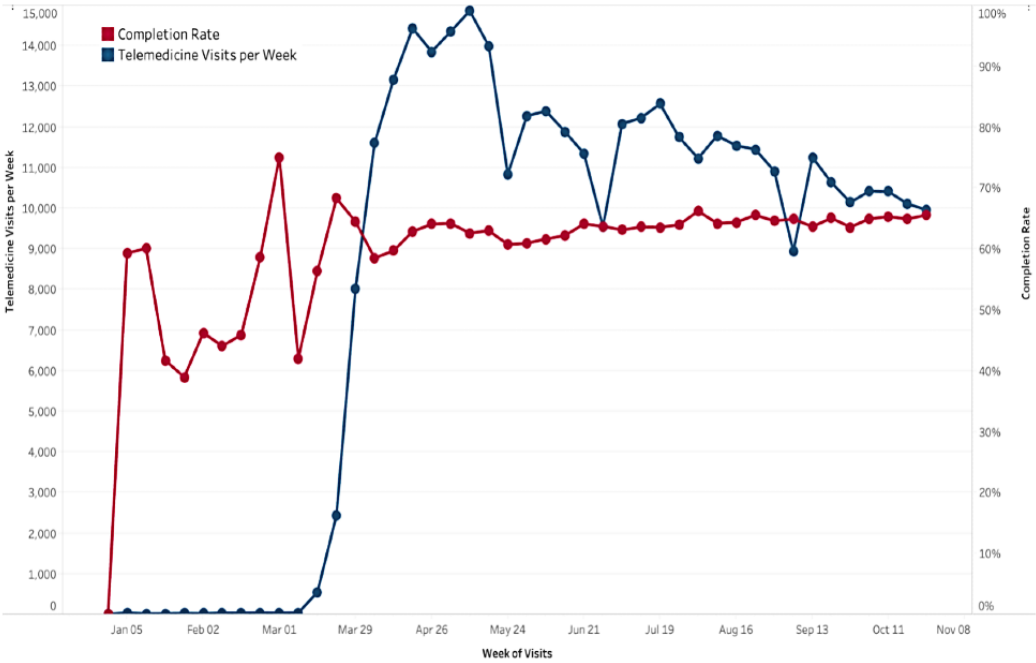
Results

UHealth Telemedicine Volume

At the UHealth system, telemedicine visits began to sharply rise at the end of March 2020, at the same time completion rate leveled off from high pre-COVID fluctuations (likely high variance due to small sample size pre-COVID; Figure 2). This upward trend in telemedicine visit volume corresponded with widespread implementation and organizational support of telemedicine across the UHealth system. Interestingly, as visit counts gradually trended downward over the summer and into the fall, completion rate held steady with a minor increase from the low to mid 60% range. Over the entire period from March 1 to October 31, 2020, a total of 362,764 visits were scheduled and 230,030 visits were completed.

In terms of overall visit volume over this period, pre-COVID there were 120,403 visits (34 virtual visits) in January 2020 and 116,902 visits (46 virtual) in February 2020. Corresponding to the aforementioned rise in telemedicine visits in March 2020, 4519 of the 77,414 overall visits were virtual (5.84%). While the telemedicine visit volume continued to trend upward over the next few months, in-person visits both decreased and fluctuated substantially. In April 2020, 68.10% of overall visits were virtual (36,541/53,659 [includes both scheduled and on-the-fly]), so 17,118 were in-person visits. In May 2020, 50.50% of overall visits were virtual (36,652/72,577) with 35,925 in-person visits. In June 2020, 32.56% of visits were virtual (33,981/104,376) with 70,395 in-person visits. Over the following few months, in-person visits continued to trend slowly upward (approximately 80,000 monthly) and virtual visits accounted on average for about 25% of all visits at this time.

Figure 2. Telemedicine visits and completion rates (by week) in the UHealth system (January 1, 2020 - October 31, 2020). This figure shows the abrupt increase in telemedicine visits in the last week of March corresponding to the COVID pandemic and change in reimbursement by the CMS. CMS: Center for Medicare and Medicaid Services.



XSL•FO
RenderX

Characteristics of Study Sample

The study sample mainly comprised females (217,221/362,764, 59.88%), who were White (265,451/362,764, 73.17%), Hispanic (186,268/362,764, 51.35%), having primary language as English (259,714/362,764, 71.59%), and had Commercial health insurance (209,750/362,764, 57.82%) with a mean age of 50.8 years (Table 1). Interestingly, 27.07% (98,194/362,764) of the population had Spanish selected as their preferred language.

Table 1. Demographic characteristics of the overall study sample and by visit completion status.

Demographics	Overall (n=362,764)	Complete (n=230,030)	Not complete (n=132,734)	*P* value
Sex, n (%)				<.001
Male	145,543 (40.12)	93,038 (63.92)	52,505 (36.08)	
Female	217,221 (59.88)	136,992 (63.07)	80,229 (36.93)	
Age (years), mean (SD)	50.8 (20.3)	50.5 (20.4)	51.3 (20.2)	<.001
Race, n (%)				<.001
White	265,451 (73.17)	169,549 (63.87)	95,902 (36.13)	
Black	45,790 (12.62)	28,464 (62.16)	17,326 (37.84)	
Asian	6027 (1.66)	3851 (63.90)	2176 (36.10)	
Other	3168 (0.87)	1821 (57.48)	1347 (42.52)	
Unknown	42,328 (11.67)	26,345 (62.24)	15,983 (37.76)	
Ethnicity, n (%)				<.001
Hispanic	186,268 (51.35)	115,910 (62.23)	70,358 (37.77)	
Non-Hispanic	153,114 (42.21)	99,619 (65.06)	53,495 (34.94)	
Unknown	23,382 (6.45)	14,501 (62.02)	8881 (37.98)	
Language, n (%)				<.001
English	259,714 (71.59)	168,523 (64.89)	91,191 (35.11)	
Spanish	98,194 (27.07)	58,732 (59.81)	39,462 (40.19)	
Other	3756 (1.04)	2136 (56.87)	1620 (43.13)	
Unknown	1100 (0.30)	639 (58.09)	461 (41.91)	
Insurance, n (%)				<.001
Commercial	209,750 (57.82)	134,986 (64.36)	74,764 (35.64)	
Medicare	98,737 (27.22)	62,562 (63.36)	36,175 (36.64)	
Medicaid	43,202 (11.91)	26,383 (61.07)	16,819 (38.93)	
Other	5366 (1.48)	3301 (61.52)	2065 (38.48)	
Uninsured	5709 (1.57)	2798 (49.01)	2911 (50.99)	
Weighted average income (thousands), mean (SD)	97.7 (146)	100.2 (151.5)	93.6 (135.4)	<.001

Additionally, 98.67% (357,922/362,764) of visits were not new to UHealth, with only 1.33% (4842/362,764) having this visit to be their first in the UHealth system (Table 2). Concerning the MyUHealthChart (patient portal) activation status, 93.34% (338,596/362,764) of patients had activated their account, with 6.58% (23,883/362,764) not having activated it, and only 0.08% (285/362,764) having declined to have a MyUHealthChart account. Most of the visits (279,159/362,764, 76.95%) were a follow-up visit with the given provider (meaning the patient had a prior encounter within 3 years with the given provider). Clinic locations were assigned to either the main campus (217,855/362,764, 60.06%) in downtown Miami, or one of the satellite clinics (144,909/362,764, 39.95%). When grouped into 4 categories, telemedicine visits were occurring most in medical specialties (225,326/362,764, 62.11%), followed by primary care (64,164/362,764, 17.69%), surgical specialties (53,190/362,764, 14.66%), and finally in other specialties (20,084/362,764 [5.54%]; eg, optometry, audiology, exercise physiology). The patient appointment automated phone call/SMS text message reminder resulted in 38.97% (141,369/362,764) confirmed, 60.16% (218,236/362,764) not confirmed, and 0.87% (3159/362,764) answered but did not confirm (phone call only) visits.

Looking at all variables, many patient demographic characteristics had significant differences between completed and not completed telemedicine visits (Table 1). However, the focus of this analysis was to identify patient-agnostic characteristics affecting telemedicine completion rate to guide actionable changes at the UHealth system and potentially across other health systems (Table 2). For new patients to UHealth, the visit completion rate (4842/2319, 47.89%) was significantly

XSL•FO
RenderX

lower than that of follow-up patients (227,711/357,922, 63.62%; *P*<.001). MyUHealthChart (patient portal) activation status also showed stark differences in completion, with activated patients completing 65.55% (221,933/338,596) of visits, while not activated or declined activation patients only completing 33.44% (7987/23,883) and 39% (110/285) of visits, respectively (*P*<.001). For patients new to a given provider, the completion rate (49,804/83,605, 59.57%) was lower than follow-up visit completion rates (180,226/279,159, 64.56%; *P*<.001). Telemedicine visits assigned to the main campus were completed slightly more often (138,994/217,855, 63.80%) compared with the satellite campuses (91,036/144,909, 62.82%; *P*<.001). In terms of completion rate based on the specialty of the provider, medical specialties had a much lower completion rate (137,195/225,326, 60.89%) than other groups, including primary care (42,388/64,164, 66.06%), surgical specialties (36,486/53,190, 68.60%), and other specialties (13,979/20,084, 69.60%; *P*<.001). Automated appointment confirmation by phone call/SMS text message was associated with a very high telemedicine completion rate (121,430/141,369, 85.90%), especially when compared with patients who answered but did not confirm or patients who did not confirm visits (63.66% [2011/3159] and 48.84% [106,589/218,236] completion rates, respectively; *P*<.001). Through more granular descriptive statistics, specific specialties, providers, and clinic locations were identified in order to provide targeted optimization.

Table 2. Patient-agnostic characteristics of overall sample and by telemedicine completion status.

Characteristic	Overall (n=362,764)	Complete (n=230,030)	Not complete (n=132,734)	*P* value
New to UHealth, n (%)				<.001
Yes	4842 (1.33)	2319 (47.89)	2523 (52.11)	
No	357,922 (98.67)	227,711 (63.62)	130,211 (36.38)	
MyUHealthChart status, n (%)				<.001
Activated	338,596 (93.34)	221,933 (65.55)	116,663 (34.45)	
Not activated	23,883 (6.58)	7987 (33.44)	15,896 (66.56)	
Patient declined	285 (0.08)	110 (38.60)	175 (61.40)	
New to provider, n (%)				<.001
Yes	83,605 (23.05)	49,804 (59.57)	33,801 (40.43)	
No	279,159 (76.95)	180,226 (64.56)	98,933 (35.44)	
Campus, n (%)				<.001
Main	217,855 (60.05)	138,994 (63.80)	78,861 (36.20)	
Satellite	144,909 (39.95)	91,036 (62.82)	53,873 (37.18)	
Specialty, n (%)				<.001
Primary care	64,164 (17.69)	42,388 (66.06)	21,776 (33.94)	
Medical specialty	225,326 (62.11)	137,195 (60.89)	88,131 (39.11)	
Surgical specialty	53,190 (14.66)	36,468 (68.56)	16,722 (31.44)	
Other	20,084 (5.54)	13,979 (69.60)	6105 (30.40)	
Phone reminder, n (%)				<.001
Confirmed	141,369 (38.97)	121,430 (85.90)	19,939 (14.10)	
Not confirmed	218,236 (60.16)	106,589 (48.84)	111,647 (51.16)	
Answered, not confirmed	3159 (0.9)	2011 (63.66)	1148 (36.34)	

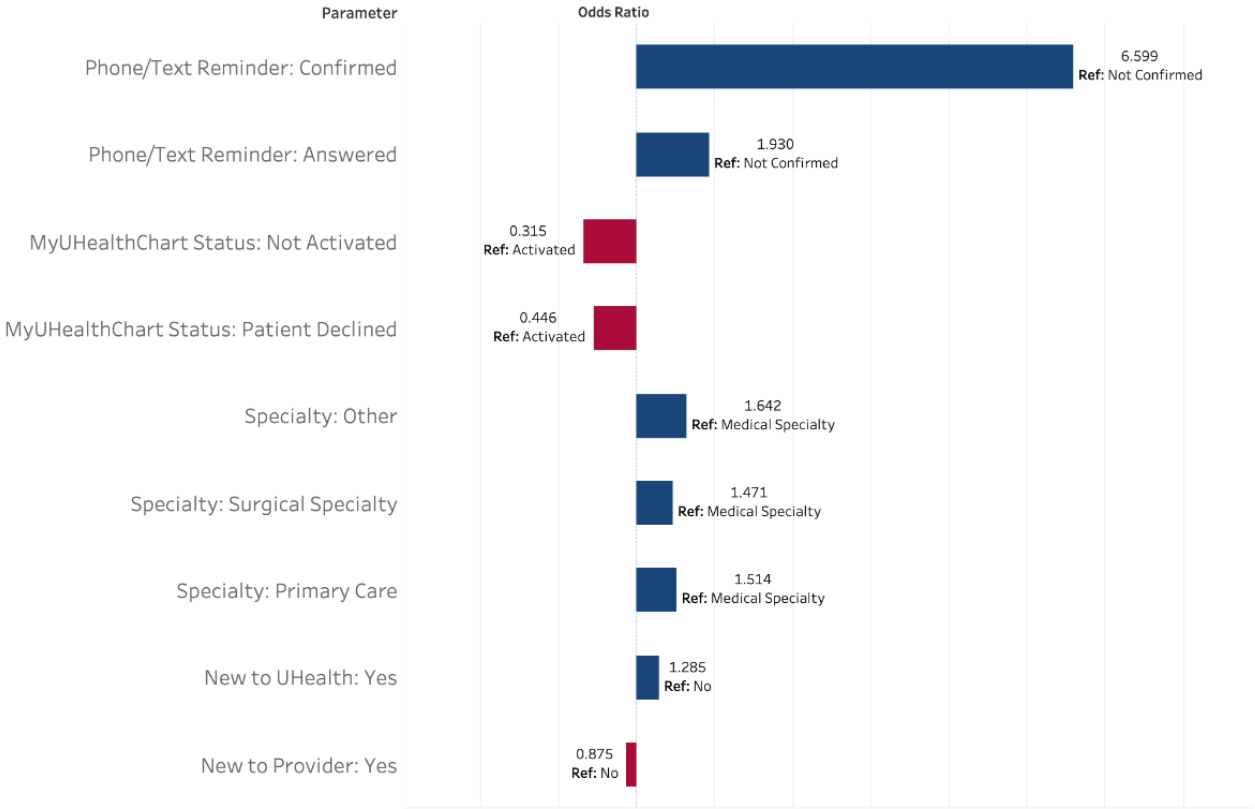
Modeling to Identify Important Patient-Agnostic Predictors

Through logistic regression (Figure 3), important patient-agnostic predictors (ie, excluding patient demographic factors) of completion included phone/SMS text message reminder status, MyUHealthChart portal status, provider's specialty, new to the UHealth system, and new to provider. People who confirmed their appointment were 6.6 times more likely to complete their visit compared with those that did not answer the phone or SMS text message (95% CI 6.483-6.717). Even those who only answered the phone call reminder but did not confirm the visit (by pressing the prompted button) were almost twice as likely to complete their visit than those who had not answered (OR 1.930, 95% CI 1.790-2.081). Also, the MyUHealthChart portal status "not activated" had a 68.5% decrease in odds of visit completion in comparison to the activated MyUHealthChart "reference" group (*P*<.001). The MyUHealthChart status of "patient declined" was also associated with a 55.4% decreased odds of completion compared with the MyUHealthChart "reference" group (95% CI 0.344-0.577). Provider's specialty also had a large effect on completion of telemedicine. The medical specialties group had the lowest completion and was used as the reference. Compared with

XSL•FO
RenderX

medical specialties, "other" specialties had a 64.2% increase in odds; surgical specialties had a 47.1% increase in odds, and primary care had a 51.4% increase in odds of telemedicine completion compared with the reference. Being a new patient to UHealth was actually associated with a 1.285 times higher odds of visit completion compared with an established patient (*P*<.001). This may seem counterintuitive as these patients would initially be unfamiliar with UHealth's specific telemedicine system, and descriptive statistics demonstrate that new patients fare worse than existing patients. In an unadjusted univariate analysis, the OR is less than one (0.526; *P*<.001), demonstrating that patients new to the health system have lower odds of completion. However, when used in the multivariable model and adjusting for clinical site-level clustering, the OR reverses as other potentially confounding variables are accounted for, showing the true direction of this data point. Conversely, being new to the provider (ie, not a follow-up visit) was associated with a 12.5% decrease in odds compared with being a follow-up for the provider (*P*<.001).

Figure 3. Mixed effects logistic regression model of visit completion status. These are the patient-agnostic variables (*P*<.001) that were included in the full model (which had the best fit statistics compared to reduced models). The full model also included: insurance, race, language, age, ethnicity, sex, religion, and weighted average income.



Random forest modeling was an additional means of verifying results from logistic regression. Using the "importance" function, the most relevant variables for predicting success in completing telemedicine visits were derived from the random forest model. These results mirrored those from the logistic model, with phone/SMS text message reminder status, MyUHealthChart status, and provider specialty being the most important in predicting telemedicine visit completion.

The predictive capabilities of both the logistic model and the random forest model were assessed. On the training data set, the logistic model had an accuracy of 69.1%, whereas on the test data set, it had an accuracy of 69.0%. This inconsequential difference in accuracy between the training and test sets indicates minimal overfitting of the model even with the large number of variables included. With regard to the random forest model, the accuracy on the training set was 71.9%, whereas on the test data set, the accuracy was 69.2%. Overall, the predictive usefulness of both of these models is quite limited given the low accuracy.

Patients "Not Activated" in Patient Portal

The subset of "not activated" MyUHealthChart patients (n=23,883) was identified to be important because it was strongly associated with not completing a telemedicine visit, as evidenced by only 33.44% (7987/23,883) completion and 68.1% decrease in odds of visit completion compared with activated patients. This "not activated" patient portal group was investigated further and found to be demographically distinct from the rest of the population (Table 3). There was a higher percentage of males, Black patients, Hispanics, Spanish speakers, and Medicare and Medicaid patients, and they were on average older (*P*<.001).

XSL•FO
RenderX

Table 3. Descriptive statistics of patients with the "not activated" MyUHealthChart status.

Characteristics	Not activated (n=23,883)	Other (n=338,881)	*P* value
Sex, n (%)			<.001
Male	11,196 (7.69)	134,347 (92.31)	
Female	12,687 (5.84)	204,534 (94.16)	
Age (years), mean (SD)	51.5 (24.6)	50.8 (20.0)	<.001
Race, n (%)			<.001
White	15,967 (6.02)	249,484 (93.98)	
Black	3633 (7.93)	42,157 (92.07)	
Asian	291 (4.83)	5736 (95.17)	
Other	419 (13.23)	2749 (86.77)	
Unknown	3573 (8.44)	38,755 (91.56)	
Ethnicity, n (%)			<.001
Hispanic	13,128 (7.05)	173,140 (92.95)	
Non-Hispanic	8522 (5.57)	144,592 (94.43)	
Unknown	2233 (9.55)	21,149 (90.45)	
Language, n (%)			<.001
English	13,647 (5.25)	246,067 (94.75)	
Spanish	9694 (9.87)	88,500 (90.13)	
Other	402 (10.70)	3354 (89.30)	
Unknown	140 (12.73)	960 (87.27)	
Insurance, n (%)			<.001
Commercial	9591 (4.57)	200,159 (95.43)	
Medicare	7911 (8.01)	90,826 (91.99)	
Medicaid	4699 (10.88)	38,503 (89.12)	
Other	581 (10.83)	4785 (89.17)	
Uninsured	1101 (19.29)	4608 (80.71)	
New to the UHealth system, n (%)			<.001
Yes	2816 (58.16)	2026 (41.84)	
No	21,067 (5.89)	336,855 (94.11)	

Discussion

Previsit Reminder

This analysis found that a patient who confirms his/her appointment via the automated phone or SMS text message is most strongly associated with a successful telemedicine visit completion. These results mirror what previous studies saw for in-person visits: patients who received automated reminders presented a significant difference in no-show rates compared with those that did not receive a reminder (17.3% vs 23.1%) [13]. However, it is important to note that, in this study, reminders done by clinic staff had an even lower no-show rate of 13.63% (445/3266, $P<.01$) (statistically significant at $\alpha=.05$) compared with both automated reminders and no reminders. While we were unable to directly evaluate staff reminders that occurred previsit, results from automated appointment reminders are elucidating. Perhaps, these reminders allowed for confirmation with the patient prior to the visit and may have served to identify and troubleshoot technical difficulties in accessing the telemedicine visit and to provide sufficient time to ask for assistance. Also, phone or SMS text message communication may have served as a reminder of the upcoming visit that patients would have otherwise forgotten. Regardless, phone/SMS text message confirmation status is an independent critical factor to predict a completed telemedicine visit.

Patient Portal "Activated"

The second most important variable to predict a completed telemedicine visit was having an active account for the MyUHealthChart patient portal. This underscores the importance of patients having previously activated their MyUHealthChart account prior to the visit. It is important to note that the patient portal is available in both English and Spanish. UHealth has also created multilingual telemedicine instructional videos and reference guides to best serve our diverse patient population.

XSL•FO
RenderX

However, there may be underlying disparities (beyond the already addressed language barrier) to patient portal activation among certain subsets of our patient population. The "not activated" subset of patients included more Black and Hispanic patients in comparison to the rest of the sample. This mirrors results found in a study on patient portal use among older adults, which found a significant decrease in use of the patient portal among Black and Hispanic patients, in comparison to non-Hispanic White patients [27]. In addition to issues patients may face within MyUHealthChart and the Zoom workflow, there are numerous other issues which may occur. For example, patients may have internet performance issues, out-of-date Zoom applications, popup blockers, slow processors, or microphone/camera/speaker problems. A technical support line is available to patients; however, this may require additional time, patience, and technical abilities from patients.

Provider Specialty

Provider's specialty also played a role in completion status, with the medical specialties group, including cardiology, gastroenterology, and pulmonology, having the lowest completion rates. The highest completion rates came from other specialties, surgical specialties, and then primary care. There may be specialty-specific considerations for telemedicine which could affect completion status. "Technical and medical requirements for telemedicine differ across medical specialties;" [6]; therefore, specialties may need a custom-designed workflow to be successful, such as hybrid visits, which include on-site testing and then telemedicine evaluation. There may also be other specialty-specific barriers such as willingness to change, leadership emphasis on telemedicine, or telemedicine support allocation. As a result of urgent and rapid implementation, specialty-specific implementation and optimization were limited. This illustrates the need to reevaluate outcomes after implementation to identify opportunities for improvement across a health system.

New to UHealth/New to Provider

Notably, new patients to the UHealth system were more likely to complete visits, which is opposite of the results seen in descriptive statistics, as additional confounders are controlled for via a multivariable model. Possibly, new patients had more time interacting with UHealth employees when scheduling their initial visit and therefore more assistance getting properly set up from a technical perspective. Concurrently, existing patients might receive relatively less previsit attention as it could be falsely assumed they had navigated the UHealth telemedicine system previously. Also, patients themselves might overestimate their familiarity with a telemedicine workflow, as they previously had an in-person visit. More research is needed to specifically examine patients new to a health system, as much of the literature focuses on new patients to providers.

New patients to a provider were less likely to complete visits compared with patients that had already established care with this provider, which is similar to results from previous studies on in-person no-show rates. One study found that, "New patients [to an academic otolaryngology department] had the highest rate of no-show [in-person] appointments" compared with other visit types (follow-up, procedure, postoperative) [28]. An additional study also found that there was a higher incidence of no-show rates (for in-person visits) among those that were new patients to a clinic (30.5%) compared with established patients (18.3%) (with $P<.0001$) [29]. Perhaps, these findings in relation to telemedicine visits could be due to the existing provider–patient relationship, which may be associated with this increase in follow-up visit completion. Established patients may be more likely to remember they have a visit and feel more accountable for attending their visit compared with new patients. Additionally, new patients may be more reluctant to seek care for a new medical issue during this PHE, which may lead to additional testing and exposure. A new patient to a provider might feel their condition requires an in-person visit and may avoid having a telemedicine visit.

Limitations

While this analysis reveals many insights from telemedicine implementation across our health system, there are some limitations to this study and data set. Patients who canceled or did not schedule a telemedicine visit are not accounted for in this study, as we only examined those who were willing to participate in and had scheduled a telemedicine visit. As far as phone/SMS text message confirmation status is concerned, there are patients that had opted out of receiving notifications and certain visit types or specialties that had opted out of sending notifications. Therefore, there is a level to this variable that is not represented in the data which could affect results. Also, because we were provided a deidentified data set, we lacked the ability to identify repeat visits and use this information to understand how repeat visits by the same patient affect completion rate. Some providers have noted that patients who were previously unsuccessful with video telemedicine visits (having needed to convert them to telephone visits) tend to continue having difficulty with subsequent videos visits. Also, this includes data from an academic health system and does not compare with other health systems. Finally, we lacked additional variables that could serve as a better predictor of completion status and could improve accuracy of the models. Anecdotally, having a registered nurse or medical assistant help patients in navigating the telemedicine workflow was most critical to success, as there can be notable time and effort required to assist patients. Further research is needed to identify additional variables that could be used for better prediction and also take into account repeat patients.

Future Research

Another area that was not examined was the views of providers and administrators on this technology, given that we were collecting mainly variables from the patients' perspective. Interestingly, Tanriverdi and Iacono [6] define 4 barriers to health care providers' acceptance of telemedicine. The technical barrier can be addressed by providing support for acquiring, developing, and customizing technology, as well as solving technical problems. The economic barrier requires the administration to develop business models that demonstrate the generation of revenue and provide a cost justification for the expense of telemedicine. This barrier also requires telemedicine reimbursement through insurance. From an organizational perspective, efforts must be exerted to create useful workflows

XSL•FO
RenderX

and to provide organizational support for regular usage of the technology. Finally, from a behavioral standpoint, to be successful, telemedicine requires champions who are skilled in change management.

Additionally, as Tanriverdi and Iacono [6] state, "Experiential learning to lower the four knowledge barriers and ratification of knowledge claims through scientific and pragmatic criteria were most effective in constructing the 'working' of a telemedicine application." [6]. It was expected that telemedicine visit completion rates would improve naturally over time with added provider/patient experience with this new technology. However, our analysis showed only minor improvements, indicating opportunities to progress. Likely, these completion rates could be increased by system-wide optimization compounded with reducing demographic disparities. The completion rates experienced across all disciplines may be attributable to the barriers cited by Tanriverdi and Iacono [6]. In particular, certain specialties experienced a lower completion rate potentially stemming from a lack of tailored workflows for their discipline (organizational barrier). Concerning the economic barrier, allocation of trained staff to guide patients before a telemedicine visit requires institutional finances. There are multiple barriers that can be addressed at the health system level to improve effective telemedicine overall, but more data and future studies are needed.

Conclusions

Telemedicine will continue to be a part of delivering health care in the future, which makes it extremely important to use these results and other analyses as a guide to continued improvement. Given the current findings, an emphasis on patient portal activation and patient confirmation of appointment are high-yield changes to increasing completion rates. This ensures that not only are patients reminded of their upcoming visit, but also given sufficient time to set up the required technology. We recommend implementing a standardized telemedicine checklist for patients and staff to improve workflow. In addition, patients new to a health system may be receiving more focused previsit attention in order to better onboard them. This could possibly lead to a relative neglect of existing patients within the health system that may not be familiar with telemedicine visit procedures which differ greatly from in-person visits. All patients new to telemedicine should receive effective guidance regardless of their previous usage of the particular health care system. Attention should be paid to those specialties, providers, and locations with lower completion rates compared with others. As telemedicine was implemented on a large scale across entire health systems, certain workflows or features may not be transferrable to particular providers. These users should receive greater technology acclimation intervention, as well as be consulted regarding telemedicine workflow changes that would be appropriate for them. While telemedicine should be tailored, there also needs to be a standardized workflow for clinic staff to guide patients through the system. With these changes, telemedicine completion rates can be improved on a wider-scale, paving the way for additional technology innovation in medicine for future years to come.

Acknowledgments

The authors thank Anantha Gangadhara who performed the clinical data request.

Authors' Contributions

KG performed the data cleaning, analysis, visualizations, interpretation, and majority of the manuscript writing. JR conceived of the project idea and was involved in all stages of the project including manuscript writing and editing. DF assisted with manuscript writing and editing. MS assisted with manuscript writing and editing.

Conflicts of Interest

None declared.

References

1. Zundel K. Telemedicine: history, applications, and impact on librarianship. Bull Med Libr Assoc 1996 Jan;84(1):71-79 [FREE Full text] [Medline: 8938332]
2. Reed ME, Parikh R, Huang J, Ballard DW, Barr I, Wargon C. Real-Time Patient-Provider Video Telemedicine Integrated with Clinical Care. N Engl J Med 2018 Oct 11;379(15):1478-1479. [doi: 10.1056/NEJMc1805746] [Medline: 30304654]
3. O'Gorman LD, Hogenbirk JC, Warry W. Clinical Telemedicine Utilization in Ontario over the Ontario Telemedicine Network. Telemed J E Health 2016 Jun;22(6):473-479 [FREE Full text] [doi: 10.1089/tmj.2015.0166] [Medline: 26544163]
4. Bajowala SS, Milosch J, Bansal C. Telemedicine Pays: Billing and Coding Update. Curr Allergy Asthma Rep 2020 Jul 27;20(10):60 [FREE Full text] [doi: 10.1007/s11882-020-00956-y] [Medline: 32715353]
5. Notification of Enforcement Discretion for Telehealth Remote Communications During the COVID-19 Nationwide Public Health Emergency. Washington, D.C: U.S. Department of Health & Human Services; 2020. URL: https://www.hhs.gov/hipaa/for-professionals/special-topics/emergency-preparedness/notification-enforcement-discretion-telehealth/index.html [accessed 2020-11-20]
6. Tanriverdi H, Iacono CS. Diffusion of telemedicine: a knowledge barrier perspective. Telemed J 1999;5(3):223-244. [doi: 10.1089/107830299311989] [Medline: 10908437]

XSL•FO
RenderX

7. Medicare Beneficiary Use Of Telehealth Visits: Early Data From The Start Of COVID-19 Pandemic. 2020 Jul 28. URL: https://aspe.hhs.gov/system/files/pdf/263866/HP_IssueBrief_MedicareTelehealth_final7.29.20.pdf [accessed 2020-11-20]
8. Eberly LA, Khatana SAM, Nathan AS, Snider C, Julien HM, Deleener ME, et al. Telemedicine Outpatient Cardiovascular Care During the COVID-19 Pandemic: Bridging or Opening the Digital Divide? Circulation 2020 Aug 04;142(5):510-512. [doi: 10.1161/CIRCULATIONAHA.120.048185] [Medline: 32510987]
9. Eberly LA, Kallan MJ, Julien HM, Haynes N, Khatana SAM, Nathan AS, et al. Patient Characteristics Associated With Telemedicine Access for Primary and Specialty Ambulatory Care During the COVID-19 Pandemic. JAMA Netw Open 2020 Dec 01;3(12):e2031640 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.31640] [Medline: 33372974]
10. Shur N, Atabaki SM, Kisling MS, Tabarani A, Williams C, Fraser JL, Rare Disease Institute. Rapid deployment of a telemedicine care model for genetics and metabolism during COVID-19. Am J Med Genet A 2021 Jan 14;185(1):68-72 [FREE Full text] [doi: 10.1002/ajmg.a.61911] [Medline: 33051968]
11. Barnett ML, Ray KN, Souza J, Mehrotra A. Trends in Telemedicine Use in a Large Commercially Insured Population, 2005-2017. JAMA 2018 Nov 27;320(20):2147-2149 [FREE Full text] [doi: 10.1001/jama.2018.12354] [Medline: 30480716]
12. Reed ME, Huang J, Graetz I, Lee C, Muelly E, Kennedy C, et al. Patient Characteristics Associated With Choosing a Telemedicine Visit vs Office Visit With the Same Primary Care Clinicians. JAMA Netw Open 2020 Jun 01;3(6):e205873 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.5873] [Medline: 32585018]
13. Parikh A, Gupta K, Wilson AC, Fields K, Cosgrove NM, Kostis JB. The effectiveness of outpatient appointment reminder systems in reducing no-show rates. Am J Med 2010 Jun;123(6):542-548. [doi: 10.1016/j.amjmed.2009.11.022] [Medline: 20569761]
14. Zhong X, Park J, Liang M, Shi F, Budd PR, Sprague JL, et al. Characteristics of Patients Using Different Patient Portal Functions and the Impact on Primary Care Service Utilization and Appointment Adherence: Retrospective Observational Study. J Med Internet Res 2020 Feb 25;22(2):e14410 [FREE Full text] [doi: 10.2196/14410] [Medline: 32130124]
15. Dantas LF, Fleck JL, Cyrino Oliveira FL, Hamacher S. No-shows in appointment scheduling - a systematic literature review. Health Policy 2018 Apr;122(4):412-421. [doi: 10.1016/j.healthpol.2018.02.002] [Medline: 29482948]
16. Drewek R, Mirea L, Adelson PD. Lead Time to Appointment and No-Show Rates for New and Follow-up Patients in an Ambulatory Clinic. Health Care Manag 2017;36(1):4-9. [doi: 10.1097/hcm.0000000000000148]
17. UHealth. University of Miami. URL: https://welcome.miami.edu/uhealth/index.html [accessed 2021-05-05]
18. U.S. Census Bureau. QuickFacts. Miami-Dade County, Collier County, Palm Beach County, Broward County, Florida. URL: https://www.census.gov/quickfacts/fact/table/palmbeachcountyflorida,browardcountyflorida,miamidadecountyflorida,colliercountyflorida/PST045219 [accessed 2021-05-05]
19. Statistics of Income (SOI) Division. SOI Tax Stats - Individual Income Tax Statistics. URL: https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi [accessed 2020-11-20]
20. R Foundation for Statistical Computing. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing URL: https://www.R-project.org/ [accessed 2020-08-01]
21. Barrett T, Brignone E. Furniture for Quantitative Scientists. The R Journal 2017;9(2):142-148. [doi: 10.32614/rj-2017-037]
22. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using. J. Stat. Soft 2015;67(1):1-48. [doi: 10.18637/jss.v067.i01]
23. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics 2005 Oct 15;21(20):3940-3941 [FREE Full text] [doi: 10.1093/bioinformatics/bti623] [Medline: 16096348]
24. Liaw A, Wiener M. Classification and regression by randomForest. R News 2002 Dec;2/3(3):22 [FREE Full text]
25. Tableau Desktop [Computer Software]. Version 2020.3.2. Seattle, WA: Salesforce; 2003.
26. Rivkees SA, Roberson S. The Florida Department of Health STEPS Public Health Approach: The COVID-19 Response Plan and Outcomes Through May 31, 2020. Public Health Rep 2020 Aug 06;135(5):560-564. [doi: 10.1177/0033354920946785] [Medline: 32758023]
27. Gordon NP, Hornbrook MC. Differences in Access to and Preferences for Using Patient Portals and Other eHealth Technologies Based on Race, Ethnicity, and Age: A Database and Survey Study of Seniors in a Large Health Plan. J Med Internet Res 2016 Mar 04;18(3):e50 [FREE Full text] [doi: 10.2196/jmir.5105] [Medline: 26944212]
28. Fiorillo CE, Hughes AL, I-Chen C, Westgate PM, Gal TJ, Bush ML, et al. Factors associated with patient no-show rates in an academic otolaryngology practice. Laryngoscope 2018 Mar 16;128(3):626-631 [FREE Full text] [doi: 10.1002/lary.26816] [Medline: 28815608]
29. Cheung DL, Sahrmann J, Nzewuihe A, Espiritu JR. No-show rates to a sleep clinic: drivers and determinants. J Clin Sleep Med 2020 Sep 15;16(9):1517-1521. [doi: 10.5664/jcsm.8578] [Medline: 32933644]

Abbreviations

CMS: Center for Medicare and Medicaid Services
HIPAA: Health Insurance Portability and Accountability Act
NASA: National Aeronautics and Space Administration

XSL•FO
RenderX

OCR: Office for Civil Rights
PHE: public health emergency
STARPAHC: Space Technology Applied to Rural Papago Advanced Health Care
UHealth: University of Miami Health System

Edited by C Lovis; submitted 15.02.21; peer-reviewed by E van der Velde, B Smith, F Fatehi; comments to author 14.03.21; revised version received 05.06.21; accepted 10.07.21; published 27.08.21.

Please cite as:
Gmunder KN, Ruiz JW, Franceschi D, Suarez MM
Factors to Effective Telemedicine Visits During the COVID-19 Pandemic: Cohort Study
JMIR Med Inform 2021;9(8):e27977
URL: https://medinform.jmir.org/2021/8/e27977
doi:10.2196/27977
PMID:34254936

©Kristin Nicole Gmunder, Jose W Ruiz, Dido Franceschi, Maritza M Suarez. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 27.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Improving Human Happiness Analysis Based on Transfer Learning: Algorithm Development and Validation

Lele Yu[1], BSc; Shaowu Zhang[1], PhD; Yijia Zhang[1], PhD; Hongfei Lin[1], PhD

College of Computer Science and Technology, Dalian University of Technology, Dalian, China

Corresponding Author:
Yijia Zhang, PhD
College of Computer Science and Technology
Dalian University of Technology
No 2 Linggong Road
Dalian, 116023
China
Phone: 86 411 84708704
Email: zhangyijia1979@gmail.com

Abstract

Background: Happiness refers to the joyful and pleasant emotions that humans produce subjectively. It is the positive part of emotions, and it affects the quality of human life. Therefore, understanding human happiness is a meaningful task in sentiment analysis.

Objective: We mainly discuss 2 facets (Agency/Sociality) of happiness in this paper. Through analysis and research on happiness, we can expand on new concepts that define happiness and enrich our understanding of emotions.

Methods: This paper treated each happy moment as a sequence of short sentences, then proposed a short happiness detection model based on transfer learning to analyze the Agency and Sociality aspects of happiness. First, we utilized the unlabeled training set to retrain the pretraining language model Bidirectional Encoder Representations from Transformers (BERT) and got a semantically enhanced language model happyBERT in the target domain. Then, we got several single text classification models by fine-tuning BERT and happyBERT. Finally, an improved voting strategy was proposed to integrate multiple single models, and "pseudo data" were introduced to retrain the combined models.

Results: The proposed approach was evaluated on the public dataset happyDB. Experimental results showed that our approach significantly outperforms the baselines. When predicting the Agency aspect of happiness, our approach achieved an accuracy of 0.8653 and an F1 score of 0.9126. When predicting Sociality, our approach achieved an accuracy of 0.9367 and an F1 score of 0.9491.

Conclusions: By evaluating the dataset, the comparison results demonstrated the effectiveness of our approach for happiness analysis. Experimental results confirmed that our method achieved state-of-the-art performance and transfer learning effectively improved happiness analysis.

(JMIR Med Inform 2021;9(8):e28292) doi:10.2196/28292

KEYWORDS

happiness analysis; sentiment analysis; transfer learning; text classification

Introduction

As the pressure of social life increases, people's mental health has also received extensive attention. Taking depression as an example, the World Health Organization reported that more than 350 million people suffer from depression, and the growth in the rate of patients with depression over the past 10 years is about 18%. From these data, psychological illness has an essential impact on human health and has become the leading cause of health problems. Therefore, sentiment analysis has become a valuable research hotspot. Happiness is a positive part of the sentiment, and research on happiness also has the prospect of practical application and the value of sentiment analysis.

The current research on happiness mainly comes from the CL-Aff Shared Task 2019: in Pursuit of Happiness [1]. This shared task has published 2 tasks. The first task is a semisupervised classification task: predict thematic labels

XSL•FO
RenderX

(Agency and Sociality) on unseen data, based on small labeled and large unlabeled training data. The second task is to suggest interesting ways to automatically characterize the happy moments in terms of affect, emotion, participants, and content. Our focus is on the first task, and we challenge the current understanding of emotion through a task that models the experiential, contextual, and agentic attributes of happy moments. This paper mainly explores 2 aspects of happiness, namely Agency and Sociality. Agency mainly focuses on whether happy moments are dominated by people, while Sociality focuses more on whether happy moments involve other people. As shown in Figure 1, from the sentence "The day I got my degree in industrial engineering," we can see that this happy moment comes from the author's degree and the author controls this behavior. Therefore, the Agency label for this happy moment is set to "YES"; at the same time, this happy moment does not involve other people, so the Sociality label of this happy moment corresponds to "NO." It can be seen from this example that our proposed method should focus on different aspects of sentences. Therefore, we used inconsistent text classification models to predict the Agency and Sociality of happiness.

Figure 1. Examples of happy moments along two binary dimensions: Agency and Sociality.

Agency: Is the author in control? YES/NO

Examples (Answer is YES):

– "The day I got my degree in industrial engineering"

– "I went to office hour of one of my professors, and I realized that he was the most caring professor/mentor ever."

Examples (Answer is NO):

– "My son woke me up to a fantastic breakfast of eggs, his special hamburger patty and pancakes."

– "The weather has been warm and gorgeous for the first time in a long time and I'm loving it."

Social: Does this moment involve other people other than the author? YES/NO

Examples (Answer is YES):

– "I went to office hour of one of my professors, and I realized that he was the most caring professor/mentor ever."

– "My son woke me up to a fantastic breakfast of eggs, his special hamburger patty and pancakes."

Examples (Answer is NO):

– "The day I got my degree in industrial engineering"

– "The weather has been warm and gorgeous for the first time in a long time and I'm loving it."

Happiness analysis is an essential part of sentiment analysis, which aims to classify the Agency and Sociality of a happy moment and be regarded as a typical text classification task. Traditional text classification methods are mainly based on machine learning methods, such as feature engineering. For feature engineering, the most commonly used feature is the bag-of-words feature. In addition, some more complex features have been proposed, such as n-grams [2] and entities in ontologies [3]. These methods have achieved good results in text classification tasks, but they require much manual intervention and consume a lot of time and energy. Recently, deep learning technology has gradually replaced traditional machine learning technology as the mainstream method for text classification [4]. For example, Mikolov et al [5] proposed the neural network–based language models Continuous Bag of Words (CBOW) and Skip-gram as well as distributed word vectors. Kim [6] proposed a multiscale, parallel, single-layer convolutional neural network (CNN) combined with pretrained word vectors to achieve sentence-level text classification. Hochreiter and Schmidhuber [7] proposed long short-term memory (LSTM) for text classification to solve the problem of gradient disappearance and gradient explosion in the original recurrent neural network (RNN) during training. Vaswani et al [8] proposed a transformer mechanism in which the encoder and decoder are formed by stacking the basic feedforward neural network and attention mechanism. The aforementioned methods play an important role in text classification tasks in a field, but there are some limitations in short text classification tasks for detecting happiness. The main reasons are that the size of the dataset is small, the text length of the dataset is short, the context of sentences is not close, and the number of emotional words contained in the text of the dataset is too small. Therefore, we proposed a method based on transfer learning and deep learning to solve these problems.

With the emergence of more machine learning application scenarios, the existing better-performing supervised learning requires a large amount of labeled data. However, labeling data is a tedious and costly task, so transfer learning has received increasing attention. Transfer learning has significant influence in the field of computer vision. Most models applied in the computer vision field use existing models for fine-tuning and rarely train from scratch. Pretrained models are obtained on big data such as ImageNet and MS-COCO [9-11]. The transfer learning currently applied to natural language processing (NLP)

XSL•FO
RenderX

is mainly aimed at the first layer of the model. By fine-tuning the pretrained word embedding, it can be considered a simple transfer learning technique, but it has great value in practical applications and can be applied to various deep learning models. Based on transfer learning, we used model fine-tuning to complete the task of short text classification about happiness. To improve model performance and training efficiency, we used the triangle learning rate [12] and made full use of the hidden layer state information of the model. At the same time, transfer learning has also been widely applied to NLP. Embeddings from Language Models (ELMo) [13] appeared as a dynamic word vector in 2018, expressing different words in different contexts. Devlin et al [14] and others proposed a pretraining language model called Bidirectional Encoder Representations from Transformers (BERT) in 2018, which adopted a general pretraining model for more extensive and more profound network training.

This paper treated the happiness analysis task as a short text classification task and implemented transfer learning based on BERT. Considering the effectiveness of the pretrained model, we used model-tuned transfer learning technology to complete the task of happiness analysis. The main contributions of this paper are as follows. First, we got a semantic enhancement model happyBERT in the target domain by retraining BERT. The experimental results confirmed that domain-specific BERT outperforms general domain BERT on the HappyDB dataset [15]. Second, by fine-tuning the classification model, we mainly compared the influence of [CLS] tokens in different hidden layers of the model and the influence of other tokens in the last hidden layer on the experimental results and the further combination of the model and the deep learning neural network. The experiment proved that the fine-tuned model improved experimental results. We merged the fine-tuned model. Then, we proposed an improved voting fusion strategy to fuse the fine-tuning model, which could get the best model fusion combination, and introduced the “pseudo data” to retrain the model combination. Third, the experimental results showed that our proposed model achieved state-of-the-art performance in the task of happiness analysis.

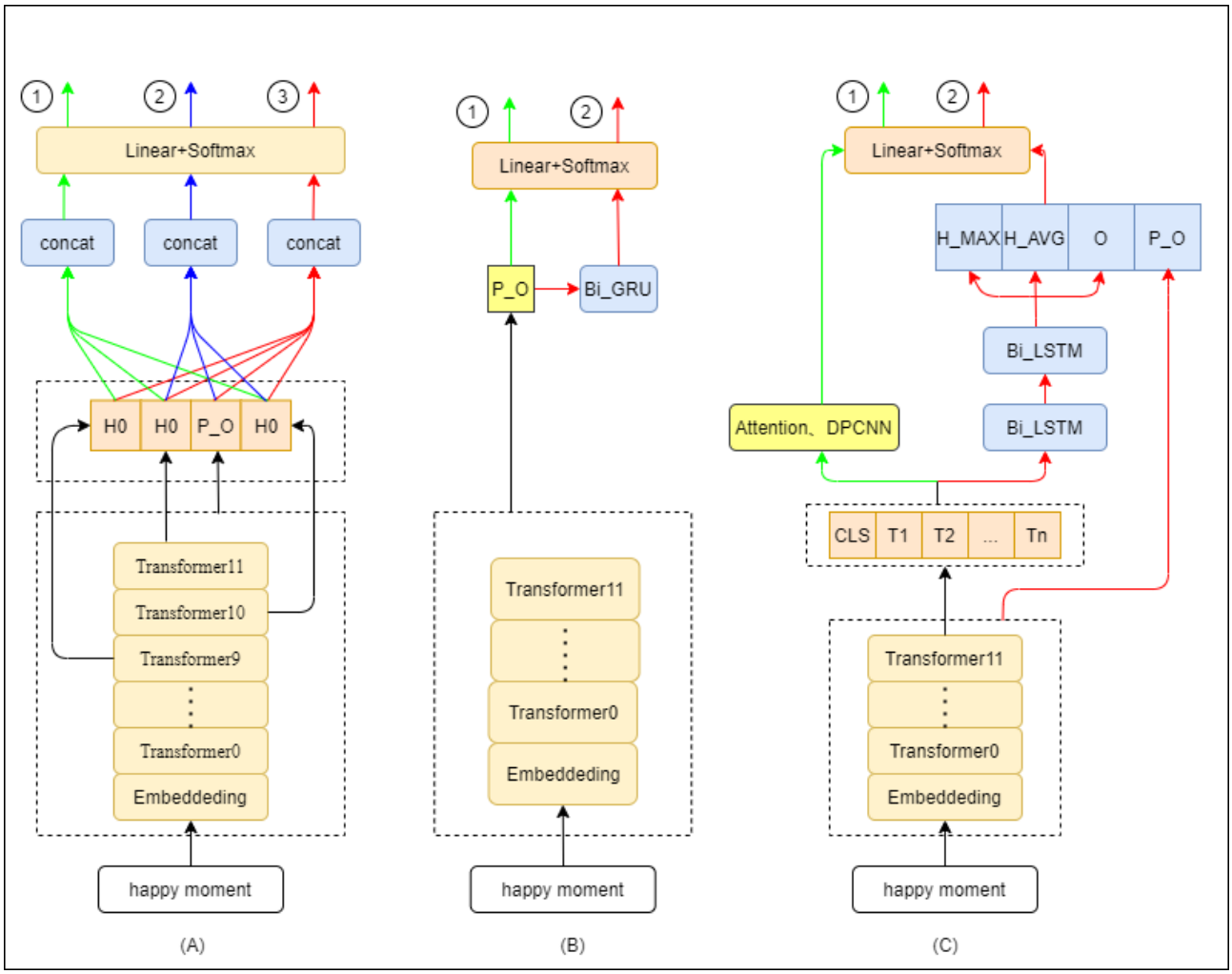
Methods

Architecture

Our proposed model architectures (Figure 2) take as input preprocessed data (data splicing, data cleaning), which is input into the pretraining language model at a word level, and output “YES” or “NO” over a discrete label space. Unlike the general methods, we focus on the [CLS] token of the last layer of the language model and focus on the other tokens in the last layer of the language model and the output of other layers. We spliced these outputs with neural network models and got the classification results through the softmax layer. The pooler_output represents the hidden state of the first token of the sequence further processed by linear layer and Tanh activation function in the last layer of BERT or happyBERT. Based on the BERT model and happyBERT model, we made the following improvements. We extracted the first state output of the hidden layer in the model (Figure 2A). Then, we concatenated the first status output of the last 3 layers and passed a fully connected layer to achieve classification, as shown in 1. We concatenated the pooler_output and the first status output of the last 2 layers, then passed a fully connected layer to achieve classification, as shown in 2. Finally, we concatenated the pooler_output and the first status output of the last 3 layers, then passed a fully connected layer to achieve classification, as shown in 3.

XSL•FO
RenderX

Figure 2. Introduction to the model structure used in the experiment: (A) extract the first state output of the hidden layer inside the model, (B) extract the model pooler_output, and (C) utilize all the state information of the last hidden layer of the model. BiLSTM: bidirectional long short-term memory; DPCNN: deep pyramid convolutional neural networks.



As shown in Figure 2B, we extracted the model pooler_output and directly used the pooler_output of the original model for classification, which is also the common method of the original model for classification, as shown in 1. Then, we used the pooler_output of the original model as the input of the upper BiGRU [16] and then classified as shown in 2.

As shown in Figure 2C, we utilized all the state information of the last hidden layer of the model. All the last hidden layer state information can be used as input and then connected to other network models, such as self-attention and deep pyramid convolutional neural networks (DPCNN) [17]. Then, we classified it, as shown in 1. The status information can be connected to deeper network models, such as bidirectional LSTM (BiLSTM) and bidirectional gated recurrent unit (BiGRU) [16]. We extracted the higher-dimensional features of the text through a deeper network model and then aggregated the BiGRU output and hidden layer state features by extracting the hidden layer state, average pooling, and max pooling, finally concatenating the pooler_output of the BERT model for classification, as shown in 2.

The research was mainly divided into 3 stages: The first stage was fine-tuning the pretrained language model BERT, the second stage was to transform the upper structure of the language model obtained in the first stage to obtain a text classification model and then fine-tune the classification model, and the third stage was to ensemble the classification model obtained in the second stage, so we could get the best model combination, and then introduce “pseudo data” to retrain the best combination models to improve the overall classification results.

Language Model

Observing the overall architecture of the model, there are many deep learning models used in this architecture. The following sections mainly introduce the language models.

BERT

We chose the pretraining language model BERT in this study. Proposed by the Google AI research institute in October 2018, BERT is a pretraining model that can achieve excellent machine reading comprehension, text classification, and other NLP tasks. This study adopted the base version of BERT, which is named BERT_base. BERT_base has less parameter information compared with BERT_large. On the BERT_based, the number of Transformer blocks is 12, the hidden layer size is 768, the number of self-attention heads is 12, and the total number of parameters for the pretrained model is 110,000,000.

XSL•FO
RenderX

happyBERT

The general field dataset used by Google to train the BERT model is very diverse, but the data in the relative happiness field have different distributions. Since the HappyDB dataset [15] contains a large amount of unlabeled data, we retrained BERT on the unlabeled corpus and updated the weights of the original BERT. Then, the resulting new pretraining model was called happyBERT. To adapt the pretrained language model to the happiness analysis task, we fine-tuned the model using the tilted triangular learning rate to quickly converge to the appropriate region of the parameter space at the beginning of training and optimize its parameters.

BiLSTM

LSTM is an improved RNN model based on RNN, which is widely used in many NLP tasks. The LSTM model overcomes the vanishing gradient problem by introducing a gating mechanism. Therefore, it is suitable to capture the long-term dependency feature. The LSTM unit consists of 3 components: the input gate i_t, the forget gate f_t, and the output gate o_t. At the time step t, the LSTM unit utilizes the input word x_t, the previously hidden state $h_{(t-1)}$ and the previous cell state $c_{(t-1)}$ to calculate the currently hidden state h_t and cell state c_t. The equations are as follows:

$$f_t = \sigma(W_f x_t + U_f h_{(t-1)} + b_f) \quad \textbf{(1)}$$

$$o_t = \sigma(W_o x_t + U_o h_{(t-1)} + b_o) \quad \textbf{(2)}$$

$$g_t = \sigma(W_g x_t + U_g h_{(t-1)} + b_g) \quad \textbf{(3)}$$

$$i_t = \sigma(W_i x_t + U_i h_{(t-1)} + b_i) \quad \textbf{(4)}$$

$$c_t = f_t \odot c_{(t-1)} + i_t \odot g_t \quad \textbf{(5)}$$

$$h_t = o_t \odot \tanh(c_t) \quad \textbf{(6)}$$

where W, U, b are the weight and bias parameters and $\odot$ denotes element-wise multiplication. This study uses the BiLSTM model that can simultaneously capture the forward and backward context features. The BiLSTM model combines a forward LSTM and a backward LSTM.

BiGRU

GRU can be regarded as a variant of LSTM. GRU replaces the forget gate and the input gate in LSTM with the update gate z_t. Combining the cell state and the hidden state h_t, calculating the new information at the current moment is different from that with LSTM. The following figures show the process of GRU updating h_t:

$$r_t = \sigma(W_r x_t + U_r h_{(t-1)} + b_r) \quad \textbf{(7)}$$

$$z_t = \sigma(W_z x_t + U_z h_{(t-1)} + b_z) \quad \textbf{(8)}$$

$$h_t = \tanh(W x_t + r_t U h_{(t-1)} + b) \quad \textbf{(9)}$$

$$h_t = (1 - z_t) + z_t h_{(t-1)}) \quad \textbf{(10)}$$

where W, U, b are the weight and bias parameters. The BiGRU model combines a forward GRU and backward GRU.

Self-Attention

Attention was first proposed in 2017, and self-attention is one of the mechanisms. Different from general Attention, self-attention is the Attention of the sentence itself. To calculate self-attention, we need to declare the 3 vectors Q, K, and V. These vectors are obtained by dot multiplication of the word embedding vector H and the training matrix W created in the training process, including $Q = HW^Q$, $K = HW^K$, and $V = HW^V$. The formula for calculating Attention is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

where Q, K, and $V \in \mathbb{R}^{n \times d}$ represent the 3 matrices of query, key, and value, respectively, and d represents the dimension of K.

DPCNN

The DPCNN [17] model was first proposed in 2017. The model belongs to a low-complexity, word-level, deep CNN text classification architecture. By continuously deepening the network, it can solve the problem that the traditional CNN model cannot obtain the long-distance dependence of the text through convolution, so it can effectively represent the long-distance dependence of the text. With the deepening of the deep learning network, the related computational complexity also increases, bringing severe challenges to practical application. The DPCNN model is based on the deepening of word-level CNN to obtain the global representation of the text. The best accuracy can be obtained by increasing the network depth without increasing computational cost by much.

Classification Model

For the happiness analysis, we first retrained BERT to get the happyBERT model. Second, we made many attempts on the model output and used 4 different deep learning models to achieve classification. The deep learning models include DPCNN, BiLSTM, BiGRU, and self-attention; the model classifiers formed by splicing them with the aforementioned BERT and happyBERT models are as follows: bert_last3embeddingcls, happybert_last3embeddingcls, bert_last2embeddingcls, happybert_last2embeddingcls, bert_last3embedding,happybert_last3embedding, bert_base, happybert_base, bert+attention, happybert+attention, bert+gru, happybert+gru, bert+grulstm, happybert+grulstm, bert+dpcnn, happybert+dpcnn. In these, the "+" means that the output of the last transformer layer of the pretraining model is input to the corresponding layer of the classification model, "_" means that the output of the pretraining model is adjusted, and the last3embedding represents the 1 in Figure 2A. The last2embeddingcls and last3embeddingcls represent 2 and 3 in Figure 2A. The base represents the pooler_output of the pretraining model. To get the result, the input of a fully connected layer is classified directly.

Model Ensemble and "Pseudo Data"

We thought about improving the single model in general tasks at first, but when the single model encountered a bottleneck, we utilized a model ensemble to improve the experimental results further. There are many methods for a model ensemble;

XSL•FO
RenderX

we used a voting mechanism to improve the performance of the entire classification system.

Through the analysis of happy moments via BertViz [18], different models pay extra attention to happy moments. Therefore, different model combinations have different voting results on Agency and Society. When predicting Agency, the best model combination was happybert_last3embedding, happybert_base, bert+grulstm, bert+attention, bert_base, and voting between these 5 models; the best results can be obtained on the validation set. We used the voting results of the obtained 5 models on the test set as the final classification result. Accuracy reached 0.8574, and the F1 score reached 0.9000. Furthermore, when predicting Sociality, the best model combination was happybert+attention, happybert_last3embeddingcls, happybert+grulstm, bert+dpcnn, happybert+dpcnn, happybert+gru, bert_base, happybert_base, and voting between these 8 models. The results can achieve the best performance on the validation set, and then the voting results of the 8 models on the test set were used as the final classification result. The accuracy reached 0.9280, and the F1 score reached 0.9360. This paper used the best_com_voting model to represent the model combination that achieves the best results on the validation set.

Since the HappyDB dataset has many unlabeled training sets, it is worth paying attention to accurately using this part of the data in the experiment. In this study, we used the unlabeled training set as the test set of the single model in the aforementioned optimal model combination, and each unlabeled training set obtained the prediction results; we added these training set data as "pseudo data" into the original labeled training set and then retrained the models in the optimal model combination. Finally, these newly obtained models were used to obtain the prediction results on the test set through a voting strategy. We used the best_com_pse model to represent these newly obtained model combinations. When predicting the Agency aspect of happiness, we achieved an accuracy of 0.8653 and an F1 score of 0.9126. When predicting Sociality, we achieved an accuracy of 0.9367 and an F1 score of 0.9491.

Results

Dataset and Task Description

The happiness analysis task based on transfer learning originates from the CL-Aff Happiness Shared Task 1. According to the predefined happy moment given by the official, it returns "YES" or "NO" in the Agency and Sociality dimensions. The HappyDB dataset used in this paper is from the CL-Aff Happiness Shared Task, which includes a labeled training set, unlabeled training set, and test set. The statistics for the number of datasets are shown in Table 1.

Table 1. Statistics of the HappyDB dataset.

Dataset	Agency		Sociality		Total, n
	Positive, n	Negative, n	Positive, n	Negative, n	
Labeled training set	7796	2764	5625	4935	10,560
Unlabeled training set	-[a]	-[a]	-[a]	-[a]	72,324
Test set	12,156	5059	9798	7417	17,215

[a]Not applicable.

Assessment Criteria

We evaluated the performance of the happiness analysis task by using the F1 score and accuracy, as follows:

where T_p represents true positive, F_p represents false positive, T_n represents true negative, and F_n represents false negative.

Experiment Settings

Hyperparameter Settings

The model codes used in this task were modified and implemented based on the open-source project transformers of the HunggingFace team [19]. The pretraining language model used was the BERT pretraining model provided by the Google team. To save memory, a single GPU batch size during fine-tune was set to 4; gradient accumulation steps were set to 4. Hence, every time 1 sample was input, the gradient was accumulated 4 times, and then backpropagation was performed to update the parameters to sacrifice a certain training speed. The hyperparameter settings used in the experiment are shown in Table 2. The dropout rate of the model was set to 0.1, and the learning rate was set to 1e-5. Since the HappyDB dataset belongs to the short text dataset, the sequence length was set to 56. In addition, the number of training steps and some parameters of DPCNN and LSTM were set.

XSL•FO
RenderX

Table 2. Hyperparameter settings.

Parameter	Value	Parameter	Value
Dropout rate	0.1	Filter num (DPCNN[a])	256
Learning rate	1e-5	Filter size (DPCNN)	3
Max sequence length	56	Block size (DPCNN)	2
Optimizer	AdamW	Hidden size (LSTM[b])	128
Training steps	30,000	Bidirectional (LSTM)	True

[a]DPCNN: deep pyramid convolutional neural network.

[b]LSTM: long short-term memory.

Loss Function

Since the happiness task involves 2 subtasks, which are Agency and Sociality classifications of the Happy moment, these 2 subtasks contained 2 categories (Agency: "YES" and "NO"; Sociality: "YES" and "NO"). These 2 subtask sample categories were relatively balanced and easy to distinguish. We used the standard cross-entropy loss function as the loss function of the happiness task:

where N is the number of samples and F is the dimension of the output feature, which is equal to the number of classes. And, p is the true value, and q is the predicted value after softmax.

Our Methods and Analysis

We finally implemented 16 neural network models for happiness detection. For each model, we adopted a 5-fold cross-validation of stratified sampling. Stratified sampling ensured that the proportion of samples in each category in each fold dataset remained unchanged. The model with the highest F1 score on the validation set was selected to predict the test set, and the probability average was used for the final 5-fold fusion. Then, we used voting to do the final model fusion of these models and selected the best model combination. Finally, we introduced "pseudo data" to retrain the single model in the best combination model so that a new single model could be obtained, and then, these new models could be fused by a voting strategy. The classification results for Agency and Sociality are shown in Table 3 and Table 4.

As we can see from Table 3 and Table 4, when predicting Sociality, the happybert+dpcnn model achieved the best result of the 12 single models, with an F1 score of 0.9350; thus, it can be proved that after the language model, a splicer neural network model can improve the classification results on specific tasks. Fine-tuning the model can improve the classification results. When predicting Agency, the happybert_last3embeddingcls model achieved the best results; the F1 score was 0.8987. Different pretraining models and different deep learning neural network models can be spliced to obtain different experimental results. The knowledge characteristics learned from the HappyDB dataset [15] for every single model were different. The integrated models can complement each other to improve the performance of the entire classification system. In addition, adding "pseudo data" to the training set can expand the scale of the dataset, thus effectively improving the performance of the classification system. For predicting Agency, the F1 score we finally submitted was 0.9126, and the accuracy was 0.8653; the F1 score was 1.57% higher, and the accuracy was 1.1% higher than bert_base. For predicting Sociality, the F1 score was 0.9421, the accuracy was 0.9367; the F1 score was 1.62% higher, and the accuracy was 1.18% higher than bert_base, proving the effectiveness of our model.

XSL•FO
RenderX

Table 3. Experimental results for Agency and Sociality.

Models	Agency		Sociality	
	Accuracy	F1	Accuracy	F1
bert_base	0.8543	0.8969	0.9249	0.9332
happybert_base	0.8545	0.8959	0.9264	0.9347
bert+attention	0.8515	0.8955	0.9247	0.9330
happybert+attention	0.8516	0.8943	0.9244	0.9324
bert+grulstm	0.8531	0.8983	0.9203	0.9289
happybert+grulstm	0.8491	0.8968	0.9197	0.9289
bert_last2embeddingcls	0.8512	0.8980	0.9157	0.9291
happybert_last2embeddingcls	0.8530	0.8982	0.9197	0.9289
bert_last3embedding	0.8516	0.8955	0.9159	0.9278
happybert_last3embedding	0.8528	0.8986	0.9189	0.9305
bert+gru	0.8497	0.8964	0.9255	0.9335
happybert+gru	0.8532	0.8969	0.9260	0.9340
bert+dpcnn	0.8514	0.8948	0.9253	0.9332
happybert+dpcnn	0.8567	0.8958	0.9268	0.9350
bert_last3embeddingcls	0.8522	0.8978	0.9200	0.9285
happybert_last3embeddingcls	0.8536	0.8987	0.9180	0.9272
all_voting	0.8554	0.8997	0.9268	0.9349
best_com_voting	0.8574	0.9000	0.9280	0.9360
best_com_pse	0.8653	0.9126	0.9367	0.9491

Table 4. Results of the ablation experiments for Agency and Sociality.

Models	Agency		Sociality	
	Accuracy	F1	Accuracy	F1
bert	0.8489	0.8902	0.9154	0.9301
bert_fine	0.8543	0.8969	0.9249	0.9332
bert_best_com	0.8551	0.8996	0.9268	0.9347
bert_com_pse	0.8623	0.9086	0.9293	0.9417

Ablation Study

In order to verify the effectiveness of fine-tuning strategies, model fusion strategies, and the introduction of "pseudo data," we set up ablation experiments for comparison. The results are shown in Table 4, where bert_fine means fine-tuning the pretraining language model BERT.Compared with bert without fine-tuning, when predicting Agency, fine-tuning the language model can improve accuracy by 0.54% and the F1 score by 0.67%. When predicting Sociality, fine-tuning the language model can improve the accuracy by 0.95% and the F1 score by 0.31%, which fully proves the effectiveness of the fine-tuning model. The bert_best_com model represents the best model voting combination based on the BERT model. Compared with bert_fine, the bert_best_com model can improve the accuracy by 0.08% and the F1 score by 0.15% when predicting Agency and can increase the accuracy by 0.19% and the F1 score by 0.15% when predicting Sociality, which fully proves the effectiveness of model fusion. bert_com_pse represents the model combination obtained by introducing "pseudo data" based on the bert_best_com model. When bert_com_pse predicts Agency, it can increase the accuracy by 0.72% and the F1 score by 0.90%. When predicting Sociality, it can increase the accuracy by 0.25% and the F1 score by 0.70%, which fully proves the effectiveness of introducing "pseudo data."

Compared Experiments and Analysis

We used the following classification models to conduct comparative experiments on the HappyDB dataset to verify the effectiveness of the proposed model. For IoH-RCNN, we utilized a recurrent convolutional neural network (RCNN) and combined words with their context to get a more precise word embedding. For SAWD-LSTM, we employed an inductive transfer learning technique, pretrained an AWD-LSTM neural net on the WikiText103 corpus, and then introduced an extra step to adapt the model to happy moments. For XGBoosted

XSL•FO
RenderX

Forest and CNN, we used different feature sets to train their model, including syntactic features, emotional features, and survey features. Then, we used semisupervised learning and experimented with XGBoosted Forest and CNN models.

The results of the comparative experiment are shown in Table 5. It can be seen that our proposed method achieves the best results on the HappyDB dataset, verifying the effectiveness of transfer learning on the task of happiness analysis.

Table 5. Experimental results of the existing methods.

Models	Agency		Sociality	
	Accuracy	F1	Accuracy	F1
IoH-RCNN[a]	0.83	0.89	0.91	0.92
SAWD-LSTM[b]	0.84	0.89	0.92	0.93
XGBoosted Forest and CNN	0.83	0.88	0.89	0.90
best_com_pse (our model)	0.86	0.91	0.93	0.94

[a]RCNN: recurrent convolutional neural network.

[b]LSTM: long short-term memory.

Error Analysis

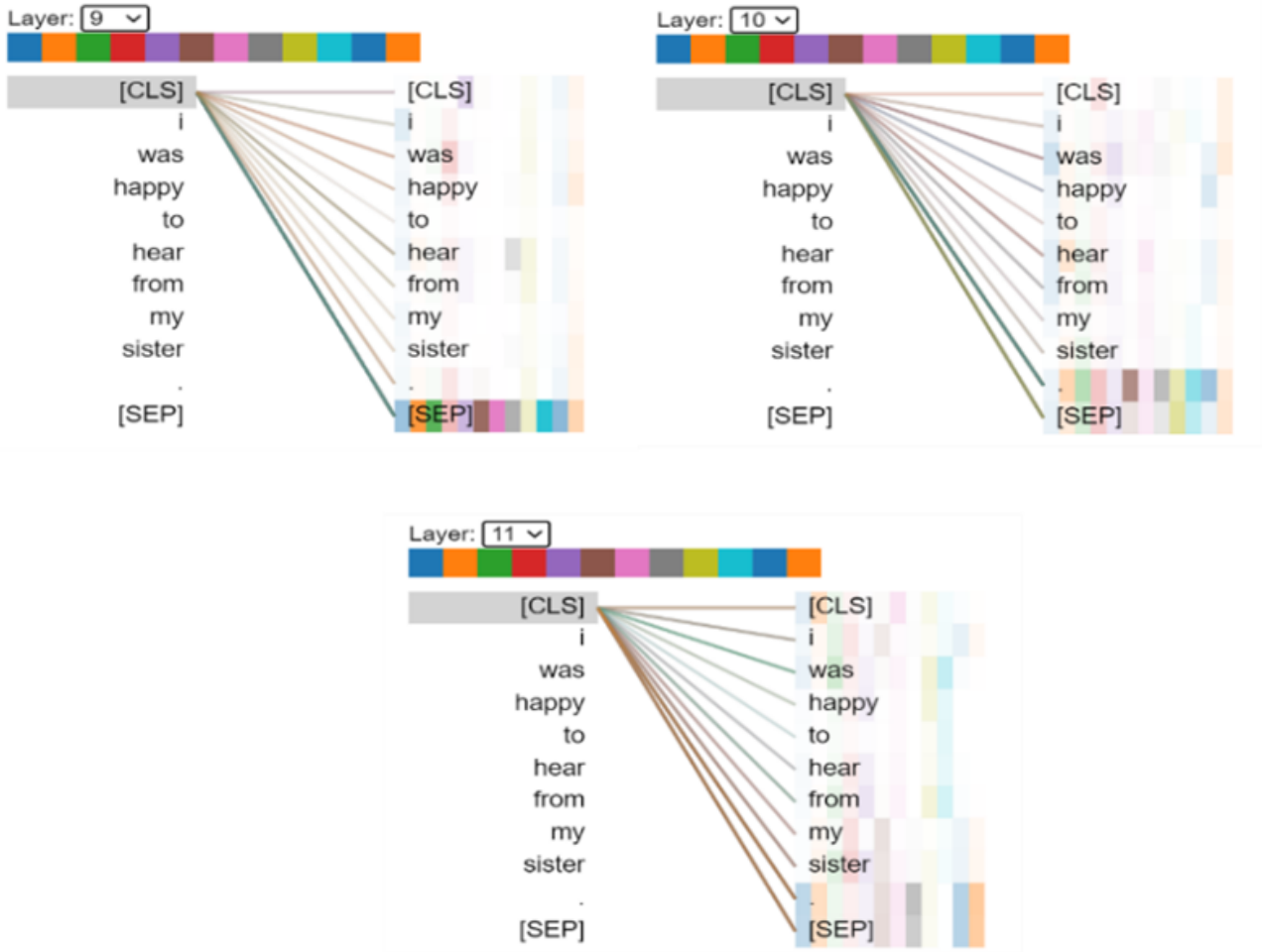
To understand our model better, we performed error analyses on the output of our final results. We observed that in some of the cases (eg, “When I got my first paycheck”), the bert_base model predicted Sociality “YES” but the happybert_base model predicted Sociality “NO”; in fact, when Sociality is “NO,” the happy_bert model learned more on the Sociality classification. When predicting “I was happy to hear from my sister,” the bert_base model predicted Agency “NO,” but the bert_last3embedding model predicted agency “YES”; in fact, when the Agency is “NO,” the bert_last3embedding model performed better on Agency classification. In the future, we will consider preferable preprocessing and postprocessing techniques to solve these problems.

Visualization of Attention Maps in BERT

Visualization can help us understand how BERT forms representations of text to understand languages. Figure 3 reveals the last 3 layers’ attention induced by a sample input text. We can see that the [CLS] of the last 3 layers of BERT had inconsistent attention to the same word, which is consistent with our proposed model concept. Our model combined the output of multiple Transformer layers of BERT to form the final output. Such attention information helped predict Agency and improved our model performance.

XSL•FO
RenderX

Figure 3. Visualization of different layer attention in an example sentence via BertViz [19].



Discussion

This paper proposed happyBERT. The happyBERT model is obtained by retraining BERT using an unlabeled training corpus in the HappyDB dataset. The purpose of retraining is to update the BERT parameters. Compared with BERT, happyBERT is more domain-relevant so that it can show better results on happiness analysis tasks, and the experimental results can better support this.

The contributions of different layers of BERT and different tokens of the same layer to the task were inconsistent. In the experimental section, we discussed the impact of the token in the BERT's last 3-layer Transformer on the experiment. Based on this thinking, we proposed single models based on BERT and happyBERT. The classification results of every single model on Agency and Sociality are given. In subsequent experiments, we also introduced an improved model fusion strategy and "pseudo labels." These strategies also improved the performance of the classification model to a certain extent.

Limitations

The happiness analysis is a novel task. So far, HappyDB is the only public dataset in this field. Moreover, only about 10,000 of the data in HappyDB are labeled. One of the limitations is that our method was only evaluated on HappyDB. In future work, we plan to annotate a larger dataset for happiness analysis.

Another limitation of our study is that we only evaluated the effectiveness of the BERT model. In recent studies, the latest pretrained models, such as Roberta [20] and GPT [21], have successfully applied NLP tasks. In a future study, we will validate these latest pretrained models on the happiness analysis task.

Conclusion

We proposed a happiness detection model based on transfer learning. Our approach utilized an unlabeled training set for training a semantically enhanced language model in the target domain and fine-tune the language model. Model fusion was applied to improve the performance of the entire happiness detection system. In addition, "pseudo data" were also introduced, which can further improve the classification performance. The experimental results suggest that our method achieves state-of-the-art performance, fully demonstrating the effectiveness of our method.

XSL•FO
RenderX

Acknowledgments

The work was supported by grants from the National Natural Science Foundation of China (No. 62072070). We would like to thank the Natural Science Foundation of China. We also would like to thank all the anonymous reviewers for their valuable suggestions and constructive comments.

Authors' Contributions

LY completed the experiment and the results from the analysis. SZ participated in the data preprocessing. YZ led the project and participated in the manuscript revision. HL provided theoretical guidance.

Conflicts of Interest

None declared.

References

1. CL-AFF Shared Task: in Pursuit of Happiness. AffCon2019. URL: https://sites.google.com/view/affcon2019/cl-aff-shared-task [accessed 2021-07-10]
2. Wang S, Manning CD. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. 2012 Presented at: 50th Annual Meeting of the Association for Computational Linguistics; July 8-14, 2012; Jeju, Republic of Korea.
3. Chenthamarakshan V, Melville P, Sindhwani V, Lawrence RD. Concept labeling: Building text classifiers with minimal supervision. 2011 Presented at: Twenty-Second International Joint Conference on Artificial Intelligence; July 16–22, 2011; Barcelona, Spain.
4. Tang D, Qin B, Liu T. Deep learning for sentiment analysis: successful approaches and future challenges. WIREs Data Mining Knowl Discov 2015 Oct 23;5(6):292-303. [doi: 10.1002/widm.1171]
5. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their CompositionalityJ. 2013 Presented at: 26th International Conference on Neural Information Processing Systems; December 5-10, 2013; Lake Tahoe, NV.
6. Kim Y. Convolutional neural networks for sentence classification. 2014 Presented at: Conference on Empirical Methods in Natural Language Processing; October 25–29, 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1181]
7. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
8. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 2018 Presented at: 31st Conference on Neural Information Processing Systems; December 5-7, 2017; Long Beach, CA.
9. Razavian AS, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: an astounding baseline for recognition. 2014 Presented at: 27th IEEE Conference on Computer Vision and Pattern Recognition; June 24-27, 2014; Columbus, OH. [doi: 10.1109/cvprw.2014.131]
10. Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. IEEE Trans. Pattern Anal. Mach. Intell 2017 Apr 1;39(4):640-651. [doi: 10.1109/tpami.2016.2572683]
11. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 Presented at: 29th IEEE Conference on Computer Vision and Pattern Recognition; June 27-30, 2016; Las Vegas, NV. [doi: 10.1109/cvpr.2016.90]
12. Howard J, Ruder S. Universal language model fine-tuning for text classification. 2018 Presented at: 56th Annual Meeting of the Association for Computational Linguistics; July 15-20, 2018; Melbourne, Australia. [doi: 10.18653/v1/p18-1031]
13. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Luke K, et al. Deep contextualized word representations. 2018 Presented at: 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 1-6, 2018; New Orleans, LA.
14. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019); June 2-7, 2019; Minneapolis, MN.
15. Jaidka K, Chhaya N, Mumick S, Killingsworth M, Halevy A, Ungar L. Beyond positive emotion: Deconstructing happy moments based on writing prompts. 2020 Presented at: Fourteenth International AAAI Conference on Web and Social Media; June 8–11, 2020; Virtual.
16. Cho K, Merrienboer BV, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014 Presented at: Conference on Empirical Methods in Natural Language Processing (EMNLP); October 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1179]
17. Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization. 2017 Presented at: 55th Annual Meeting of the Association for Computational Linguistics; July 2017; Vancouver, Canada. [doi: 10.18653/v1/p17-1052]
18. Vig J. A multiscale visualization of attention in the transformer model. 2019 Presented at: 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations; July 2019; Florence, Italy. [doi: 10.18653/v1/p19-3007]

XSL•FO
RenderX

19. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. 2020 Presented at: Conference on Empirical Methods in Natural Language Processing: System Demonstrations; November 8-12, 2020; Dominican Republic. [doi: 10.18653/v1/2020.emnlp-demos.6]
20. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Cornell University. 2019. URL: https://arxiv.org/abs/1907.11692 [accessed 2021-07-10]
21. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. 2020 Presented at: Thirty-third Conference on Neural Information Processing Systems; December 8-14, 2019; Vancouver, Canada.

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BiGRU: bidirectional gated recurrent unit
BiLSTM: bidirectional long short-term memory
CBOW: Continuous Bag of Words
CNN: convolutional neural network
DPCNN: deep pyramid convolutional neural networks
ELMo: Embeddings from Language Models
GRU: gated recurrent unit
LSTM: long short-term memory
NLP: natural language processing;
RCNN: recurrent convolutional neural network
RNN: recurrent neural network

Edited by T Hao; submitted 27.02.21; peer-reviewed by S Zhu, J Kim; comments to author 19.04.21; revised version received 04.06.21; accepted 07.06.21; published 06.08.21.

Please cite as:
Yu L, Zhang S, Zhang Y, Lin H
Improving Human Happiness Analysis Based on Transfer Learning: Algorithm Development and Validation
JMIR Med Inform 2021;9(8):e28292
URL: https://medinform.jmir.org/2021/8/e28292
doi:10.2196/28292
PMID:34383680

©Lele Yu, Shaowu Zhang, Yijia Zhang, Hongfei Lin. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 06.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Matching Biomedical Ontologies: Construction of Matching Clues and Systematic Evaluation of Different Combinations of Matchers

Peng Wang[1,2], PhD; Yunyan Hu[1], BEng; Shaochen Bai[2], BEng; Shiyi Zou[3], BEng

[1]School of Computer Science and Engineering, Southeast University, Nanjing, China

[2]School of Artificial Intelligence, Southeast University, Nanjing, China

[3]Southeast University - Monash University Joint Graduate School, Suzhou, China

Corresponding Author:
Peng Wang, PhD
School of Computer Science and Engineering
Southeast University
Si Pai Lou 2
Nanjing
China
Phone: 86 2552090977
Email: pwang@seu.edu.cn

Abstract

Background: Ontology matching seeks to find semantic correspondences between ontologies. With an increasing number of biomedical ontologies being developed independently, matching these ontologies to solve the interoperability problem has become a critical task in biomedical applications. However, some challenges remain. First, extracting and constructing matching clues from biomedical ontologies is a nontrivial problem. Second, it is unknown whether there are dominant matchers while matching biomedical ontologies. Finally, ontology matching also suffers from computational complexity owing to the large-scale sizes of biomedical ontologies.

Objective: To investigate the effectiveness of matching clues and composite match approaches, this paper presents a spectrum of matchers with different combination strategies and empirically studies their influence on matching biomedical ontologies. Besides, extended reduction anchors are introduced to effectively decrease the time complexity while matching large biomedical ontologies.

Methods: In this paper, atomic and composite matching clues are first constructed in 4 dimensions: terminology, structure, external knowledge, and representation learning. Then, a spectrum of matchers based on a flexible combination of atomic clues are designed and utilized to comprehensively study the effectiveness. Besides, we carry out a systematic comparative evaluation of different combinations of matchers. Finally, extended reduction anchor is proposed to significantly alleviate the time complexity for matching large-scale biomedical ontologies.

Results: Experimental results show that considering distinguishable matching clues in biomedical ontologies leads to a substantial improvement in all available information. Besides, incorporating different types of matchers with reliability results in a marked improvement, which is comparative to the state-of-the-art methods. The dominant matchers achieve F1 measures of 0.9271, 0.8218, and 0.5 on Anatomy, FMA-NCI (Foundation Model of Anatomy-National Cancer Institute), and FMA-SNOMED data sets, respectively. Extended reduction anchor is able to solve the scalability problem of matching large biomedical ontologies. It achieves a significant reduction in time complexity with little loss of F1 measure at the same time, with a 0.21% decrease on the Anatomy data set and 0.84% decrease on the FMA-NCI data set, but with a 2.65% increase on the FMA-SNOMED data set.

Conclusions: This paper systematically analyzes and compares the effectiveness of different matching clues, matchers, and combination strategies. Multiple empirical studies demonstrate that distinguishing clues have significant implications for matching biomedical ontologies. In contrast to the matchers with single clue, those combining multiple clues exhibit more stable and accurate performance. In addition, our results provide evidence that the approach based on extended reduction anchors performs well for large ontology matching tasks, demonstrating an effective solution for the problem.

(JMIR Med Inform 2021;9(8):e28212) doi:10.2196/28212

XSL•FO
RenderX

KEYWORDS

biomedical ontology; ontology matching; matching clues; reduction anchors

Introduction

Background

In recent years, various biomedical ontologies, such as National Cancer Institute (NCI) Thesaurus [1], Foundation Model of Anatomy (FMA) [2], Systemized Nomenclature of Medicine (SNOMED-Clinical Terms [SNOMED-CT]) [3], have been widely used in various fields, such as for medical data formats standardization [4], medical or clinical knowledge representation and integration [5], and medical decision making [6]. With the continuous evolution of biomedical data, biomedical terminology is characterized by complexity and ambiguity, which further complicates intelligent biomedical applications. Furthermore, emerging biomedical ontologies are built independently, with various ways of defining same biomedical components, resulting in heterogeneous problems. To implement the interoperability across biomedical ontologies, the establishment of meaningful connections between heterogeneous biomedical concepts is critically important [7]. Ontology matching is a solution to such semantic heterogeneity problem by determining the correspondences between concepts in different biomedical ontologies.

Because constructing alignments manually is time-consuming and labor-intensive, especially for large ontologies with thousands of concepts, some matching methods have been proposed to automatically generate ontology mappings [8]. These methods can be divided into 3 categories: terminological, structural, and external. Terminological methods are string based and designed to match names or name descriptions of ontology elements. Structural methods exploiting various types of ontology information, such as elements names, comments, and structural hierarchies, are proposed to compensate for the morphological differences between identical elements [8-14]. External methods obtain semantic mappings between syntactically dissimilar ontologies using auxiliary sources, such as taxonomies, dictionaries, and thesauri [15-18]. With the advancement of deep learning, there also exist some studies (eg, DeepAlignment [19], SCBOW + DAE(O) [20]) that try to discover alignments with representation learning based on deep learning. In the biomedical domain, some ontology matching methods based on deep learning have demonstrated the potential to facilitate the interoperability between ontologies [20-22].

Meanwhile, among the various matching techniques, to the best of our knowledge, there are surprisingly few systematic studies about the extraction and combination of matching clues and methods. As achieving satisfactory ontology alignments with a single technique is difficult, a composite approach is more efficient where different criteria or properties are considered within a single dimension. A composite approach, by contrast, that incorporates the results of some individual matchers may be simple or hybrid. This allows for high flexibility, as there is the potential for selecting the match algorithms to be executed based on the biomedical matching tasks. Moreover, there are different possibilities for combining the individual matching results. This paper attempts to empirically investigate and analyze the effectiveness of matching clues and the hybrid matching approaches.

Additionally, the inherent heterogeneity and large scale of biomedical ontologies have made discovering alignments a computationally intensive task. The divide and conquer approach [23,24] and ontology modularization [25] techniques have been proposed to decompose a large matching problem into some smaller submatching tasks. It does, however, have 2 limitations. First, most existing ontology partitioning approaches are unable to control the size of modules [23]. Consequently, many unproportionate modules (either too small or too large), which are inappropriate for matching, may be generated. Second, partitioning ontologies into modules may lead to the loss of valuable semantic information regarding the boundary elements. As a consequence, the quality of ontology matching may be impacted. Therefore, we extend *Reduction Anchors* [26], our previous method for dealing with large-scale ontology matching, to improve the performance of matching large-scale biomedical ontologies. Extended positive reduction anchors utilize the concept hierarchy to predict the ignorable similarity calculations, while the negative reduction anchors obtain the ignorable similarity calculations based on the locality of matching. The proposed method has 2 advantages over previous studies. First, it does not need to partition ontologies while maintaining the high performance as the divide and conquer approaches. Second, it is indeed a general large ontology matching framework, in which most existing matching techniques could be used.

Our main contributions in this paper are as follows:

- We provide several kinds of individual matchers with the utilization of different matching atomic clues. In order to investigate the effect of different clues in different dimensions, various combination strategies are studied to match biomedical ontologies.
- We represent multiple matchers in 4 dimensions: terminology, structure, external knowledge, and representation learning. To systematically examine and compare the effectiveness of different hybrid matchers, we design various matching strategies and combine the individual matchers for biomedical ontology matching tasks.
- We propose the extended reduction anchors-based approach for matching large-scale biomedical ontologies. It not only solves the scalability problem, but also achieves good performance with a significant reduction of execution time. Our approach achieves F1 measures of 0.925, 0.820, and 0.523 on Anatomy, FMA-NCI, and FMA-SNOMED, respectively, and reduces the matching time by nearly one-tenth. The high coverage (minimal information loss) achieved, combined with the reduction of the search space and the decreasing computation times, indicates that the extended reduction anchors are efficient.

Related Work

In recent years, ontology matching has become a popular research field. Euzenat and Shvaiko [8] present a comprehensive

XSL•FO
RenderX

overview of matching approaches and categorize techniques as terminological, structural, external, and representation learning dimensions [8]. We will focus on discussing related work on ontology matching of the biomedical domain.

Biomedical Ontology Matching

According to the features used in ontology matching, matching approaches can be classified into 4 categories: terminology-based approach, structure-based approach, external knowledge–based approach, and representation learning–based approach.

Terminology-Based Approach

In the biomedical domain, discovering alignments relying on dictionaries and similarities of terms and labels is a typical ontology matching approach, which is still widely used [8]. In some matching systems such as ASMOV [15], SAMBO [27], Falcon [28], and AgreementMakerLight [16], the terminological matcher is exploited as a basic matching method. However, the terminology-based approach often provides good precision but a low recall because it is difficult to deal with variations in the form of terms or labels (eg, equivalence between *hindlimb bone* and *bone of the lower extremity*).

Structure-Based Approach

According to the intuition that elements of 2 distinct ontologies are similar when their adjacent elements are similar, structure-based matchers utilize property attributes and taxonomy hierarchy structure [29]. CroMatcher [30] focuses on the aggregation of distinct matchers in structural level: super-element matcher, subelement matcher, domain matcher, and range matcher. Similarity flooding [29] presents a structural algorithm based on fixpoint computation and propagation of similarities along with the property relationships between elements that are usable across different scenarios, including biomedical applications. Falcon-AO [28] uses a linguistic matcher combined with a technique that represents the structure of the ontologies to be matched as a bipartite graph. Besides, the similarities between domain elements and between statements in ontologies are computed by recursively propagating similarities in the bipartite graphs. FCA-Map [31] constructs relation-based formal context to describe the biomedical elements in taxonomic, partonomic, and disjoint relationships with the anchors, and then uses the context to validate the initial lexical mappings. LogMap [17] combines the structural indexation to represent the extended class hierarchy. Contexts for the same anchor are expanded by using the class hierarchies of the input biomedical ontologies to discover new mappings.

External Knowledge–Based Approach

Matching strategies based on external knowledge provide additional lexical or structural information, allowing for the obtaining of new alignments. Biomedical ontology matching systems explore potential resources or auxiliary knowledge, such as upper-level ontology, WordNet [32], UMLS [33], and BioPortal [34], to find synonyms, spelling variants, and annotations for the concepts to be matched. Systems such as LogMap-Bio [35] and AgreementMakerLight [16] exploit a set of ontologies as background knowledge to generate equivalent mappings. In addition to the anchoring mappings related to the same background ontology, Annane et al [36] utilize alignments produced by matching intermediate ontology between each other. Faria et al [37] present a novel approach based on building the specific mapping graph as background knowledge and take into account the limitation of the selection and the combination of heterogeneous existing mappings stored in a biomedical repository. It allows getting high-quality alignments between biomedical ontologies without using complex lexical and structural measures.

Representation Learning–Based Approach

Representation learning is so far rare in ontology matching, particularly in biomedical ontologies. There are a few approaches exploring unsupervised representation learning techniques to capture the interactions among element's descriptions within biomedical ontologies. Zhang et al [38] investigated the use of representation learning for ontology matching and presented a hybrid method to incorporate word embeddings into the computation of semantic similarities among elements. Wang et al [39] proposed a neural architecture for biomedical ontology matching called OntoEmma [39]. It encodes a variety of descriptions, and derives large amounts of labeled data from biomedical thesaurus for training the model. Considering the problem of distinguishing semantic similarity and descriptive association on rare phrases, Kolyvakis et al [20] proposed a representation learning method: SCBOW+DAE(O) [20]. This approach is a representation framework based on terminological embeddings, in which the refinement of pretrained word vectors is introduced and learned by the domain knowledge encoded in ontologies and semantic lexicons. However, there still exist the limitations of the sparsity problem of structural relations and heavy dependence on pretraining. MultiOM [22] models the matching process by embedding techniques from multiple views and then optimizes the vector of concepts through a novel proposed negative sampling skill designed for structural relations in biomedical ontology.

Generally, multiple kinds of ontological clues are available, but matching biomedical ontologies based on a single category is constrained to achieve ideal performance. Consequently, most current matching systems, such as [15-17], focus on the hybrid and composite combination of various clues and matchers. Most composite methods, however, are confined to the customized combination of different matching clues and algorithms. By contrast, we attempt to study and evaluate multiple individual matchers with different combinations of matching clues and methods using different strategies. In addition, a systematic comparison of different matching clues and their integrations based on well-defined description clues does not exist so far.

Large-Scale Biomedical Ontology Matching

Many matching systems cannot work well when dealing with large matching problems. These systems perform an all-against-all comparison between concepts of the input ontologies, which requires quadratic complexity n^2 of similarity computing. To avoid the Cartesian product of the concept pairs of the source and the target ontologies, reduction of search space is indispensable.

XSL•FO
RenderX

Ontology modularization [40-42] aims to extract modules from a large and complex ontology, which is self-contained and logically consistent and can speed up the reasoning process and optimize memory utilization. Modular ontology is a popular way to partition large ontologies. However, existing modular ontology methods focus on the correctness and completeness of logics but cannot control the size of modules [23,27,43,44], that is, they would generate too large or too small modules. Algergawy et al [45] developed a seeding-based partitioning approach (OAPT) and introduced an information theoretic model selection method. It makes use of Bayesian information criterion (BIC) to determine the optimal number of modules that should be generated. However, the size of partitioned module remains uncertain.

Malasco [46] and Falcon-AO [28] are based on the divide and conquer approach that partitions a large ontology into a set of small clusters or blocks. Malasco employs 3 ontology partitioning algorithms: naive algorithm based on Resource Description Framework (RDF) sentences, structure-based algorithm [47], and ontology modularity based on ε-connection [40] for matching. Falcon-AO utilizes structural clustering to initially partition the ontologies into relatively small and disjoint blocks.

Although the modularization and divide and conquer approaches are effective to reduce the execution time, they still suffer from the contradiction between semantic completeness and information loss. After partitioning, ontology elements near boundaries of modules may lose some essential semantics, lowering the quality of alignments [26]. To overcome this problem, we introduce 2 kinds of reduction anchors to mitigate the impact of boundary loss, and simultaneously are able to reduce the number of entity pairs for which the similarity should be calculated during ontology matching.

Methods

Problem Formulation

An ontology is composed of triples like <*s*, *p*, *o*>, where *s*, *p*, and *o* stand for the subject, predicate, and object, respectively. There are 3 kinds of ontology resources: uniform resource identifier (URI) resources, literals, and blank nodes. In a triple, the subject can be URIs resources or blank nodes but not literals, and the predicate must be URI resources.

Ontology

Let O be the RDFS (RDF Schema) or OWL (Ontology Web Language) ontology represented by a set of RDF triples T. The RDF triple $t(t \cdot T)$ denotes a statement in the form of <*subject, predicate, object*>. Any node in an RDF triple may be a URI with an optional local name, a literal, or a blank node. An ontology can be represented as $O = (C, R, I)$, where C, R, and I denote sets of atomic concepts, relations (also named properties), and individuals, respectively. For simplicity, the set of concepts and properties is indicated by E.

We follow the work in [8] and give a formal definition for the ontology matching problem.

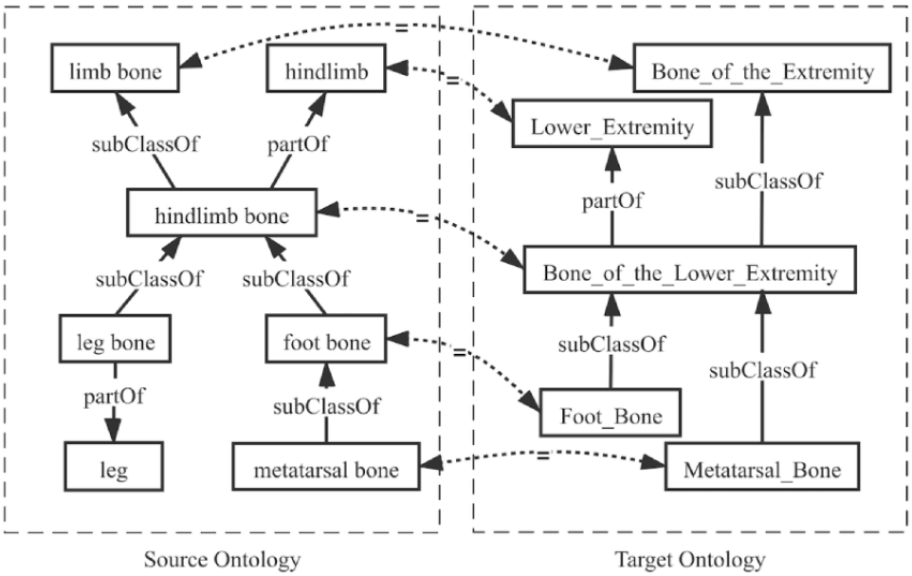
Ontology Matching

The matching between 2 ontologies O_1 and O_2 is $M = \{m_k | m_k = <e_i, e_j, r, s>\}$, where M is an alignment; m_k denotes a correspondence with a tuple $<e_i, e_j, r, s>$; e_i and e_j represent the expressions which are composed of elements from O_1 and O_2, respectively; r is the semantic relation between e_i and e_j; r could be equivalence (=), generic/specific (/), disjoint (⊥), and overlap ( ), etc.; and s is the confidence about an alignment and typically in the [0,1] range. Therefore, an alignment M is a set of correspondences m_k.

Figure 1 shows an example of alignments between a mouse anatomy ontology and the NCI Thesaurus. <*hindlimb bone, Bone_of_Lower_Extremity,=,0.7*> and <*limb bone, Bone_of_the_Extremity,=, 0.8*> are equivalent correspondences. In this paper, we only focus on identifying one-to-one equivalence correspondences between 2 concepts belonging to different ontologies.

XSL•FO
RenderX

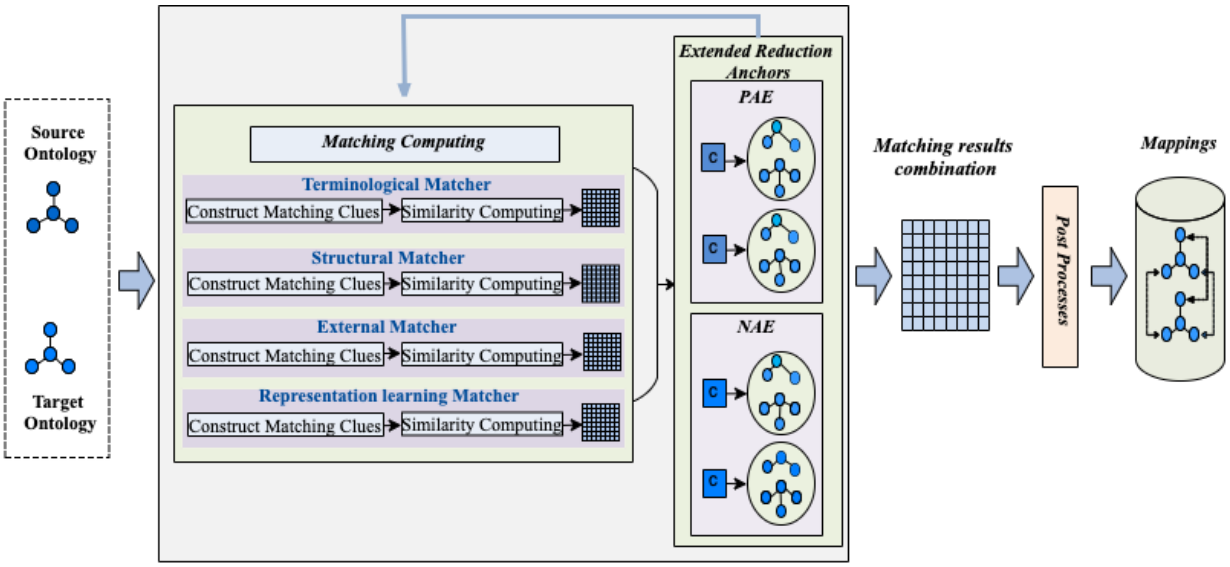
Figure 1. An example of biomedical ontology matching.



Biomedical Ontology Matching Framework

Figure 2 depicts an overview of our biomedical ontology matching framework, which includes 3 steps: (1) constructing matching clues in different dimensions: terminology, structure, external knowledge, representation learning, and building different matchers based on the extracted clues to calculate the similarities between elements; (2) constructing and updating the extended reduction anchor set iteratively through the similarity results of each matching computation and skipping the ignorable computations based on the anchors set; and (3) combining similarity matrices of different matchers assigned with different weights to obtain the alignments. For each element in input ontologies, we first create matching clues in the form of virtual documents based on the re-defined dimensions, and then single matchers are built based on the extracted clues in each dimension. Then, the similarity matrix is measured by the similarity between corresponding documents of elements. According to the similarity of each pair of elements, extended reduction anchors sets are updated and optimized continuously, which are helpful to skip meaningless similarity computations and minimize time complexity as well as search space. After obtaining similarity matrices, predefined weights are assigned to each single matcher and the matching results are combined based on re-defined superiority. Finally, the alignments are obtained through filtering processing with a given threshold.

Figure 2. Overview of biomedical ontology matching. PAE: extended positive anchors; NAE: extended negative anchors.



Matching Clues

Generally, biomedical ontology generalizes and summarizes the categories of elements in the biomedical domain. Beyond names, ontology is concerned with the principled definition of biological classes and the relations between them. Apart from the knowledge contained in the ontologies, some external resources and semantic models can be exploited to enrich the element, potentially improving ontology matching efficiency. In this section, we describe the atomic clues available for ontology matching followed by the composite clues.

XSL•FO
RenderX

Atomic Clues

Overview

The atomic clues in ontology matching are given in Table 1, and are divided into 4 types: terminological clues, structural clues, external clues, and representation learning clues. For the nodes declared in an OWL/RDF ontology, we construct virtual documents to define their clues.

Table 1. Atomic clues for ontology matching.

Clues and their sources	Description
Terminological	
Local name	Words in the local name of *e*
Label	Words in the rdfs:label of *e*
Comment	Words in the rdfs:comment of *e*
Synonym	Words in the synonym statements such as {*owl: sameAs*} and {*rdfs: seeAlso*}
Structural	
Property	Property attributes of concepts: property name, domain, range, and constraints
Hierarchy	Hierarchical context of concepts or properties, containing ancestors, descendants, siblings, and disjoint elements
External	
General dictionary	Retrieval of alternative labels and synonyms from general dictionaries such as WordNet, BabelNet
Lexicon	Cross-searching synonyms as well as cross-references from specific-domain thesauri
Representation learning	
General model	The embeddings of elements via general pretrained language models such as Word2Vec and BERT[a]
Specialized model	The embeddings via domain-specific pre-trained models such as BioBERT

[a]BERT: Bidirectional Encoder Representations from Transformers.

Terminological Clues

The terminological clues are generally the direct and representative information that distinguishes between elements. As shown in Figure 1, the concepts *foot bone* and *metatarsal bone* in source ontology are equivalent to *Foot_Bone* and *Metatarsal_Bone* in target ontology, respectively. It illustrates that terms of elements are important clues for ontology matching. The terminological clues include the words in local names, comments, labels, and synonyms in triples with the predicates: rdfs:seeAlso, owl:sameAs, owl:hasExactSynonym, and owl:hasRelatedSynonym.

Structural Clues

While lacking sufficient and consistent linguistic information about the elements, ontology structure is a piece of useful information for finding alignments. In Figure 1, for *hindlimb bone* of source ontology and *Bone_of_Lower_Extremity* of target ontology, it is difficult to discover the mapping through the terminological representations. But they have similar neighbors, *foot bone* and *Foot_Bone*, based on which we can infer that the 2 concepts would be similar. For the structural clues, they could be divided into property clues and hierarchy clues.

The property clues contain the properties attributes of concepts. The properties are represented with name, domain, and range. Some constraints might be associated with these properties, for instance, the notion of functional property.

The hierarchy clues are the context of the corresponding elements, which are reflected by ancestors, descendants, siblings, and disjoint nodes. The direct children reflect its basic structure, while the leaves reflect its semantic context. (1) The ancestor context of a node n_i could be the descriptions of upper nodes that directly link to n_i or parent nodes within a given hierarchical depth. For a blank node, likewise, we obtain the ancestor context through recursively forward traversing until the occurrence of the nonblank nodes. (2) The descendant context of a node n_i could be the set of basic and extensional descriptions of its nearest subelements or leave nodes of the subtrees of the node. Because the context of a blank node is an empty set, we could recursively obtain the context from the set of leaf nodes of subtrees rooted at node n_i. (3) The sibling context of a node n_i is defined as a set of linguistic descriptions of nodes in the same hierarchy with n_i, and these nodes share the same parent with node n_i. (4) The disjoint context of n_i is defined as the collection of linguistic descriptions of nodes that are disjoint with n_i.

External Clues

To compensate for the lack of structure and lexical information, some auxiliary knowledge and external representations are used to extract further alignments.

We retrieve alternative labels for elements to be mapped from general dictionaries (ie, WordNet). In addition, considering the specialization of biomedical ontology matching, domain-specific ontologies such as UBERON [48] and UMLS [33] are employed as auxiliary information. These ontologies are exploited to extract the cross-references and alternative synonyms, which are available to identify additional anchors.

XSL•FO
RenderX

Representation Learning Clues

Apart from the features of terminological, structural, and external resources, representation learning also has the potential to bring more semantics to biomedical ontology matching.

Word embedding can represent the implicit semantics behind elements. The general pretrained models, for example, Bidirectional Encoder Representations from Transformers (BERT) [49], trained on the large text corpus could be used to encode the element and then compute the alignments. Domain-specific language representation models are more preferable to obtain the embeddings of elements within biomedical ontologies. BioBERT [50], a domain-specific language representation model pretrained on large-scale biomedical corpora, might perform better in capturing the semantic of biomedical classes than the general models. Furthermore, fine-tuning BioBERT with synonym marginalization algorithm presented in [51] is expected to improve the quality of element representation.

Composite Clues

The composite clues are the combinations of atomic clues with different weights. The composite clues of the element e are constructed as follows:

$$\text{Clue}(e)= \alpha_1 * \text{Term}(e) + \alpha_2 * Struc(e) + \alpha_3 * Ext(e) + \alpha_4 * Rps(e)$$

where $\alpha_1, \alpha_2, \alpha_3$, and α_4 are weights in [0,1], and $Term(e)$, $Struct(e)$, $Ext(e)$, and $Rps(e)$ denote the terminological clues, structural clues, external clues, and representation learning clues of e, respectively.

Matching Process

In this section, we describe in detail the overall biomedical ontology matching process.

Name Matcher

Element names represent an important information for accessing similarities. However, in some ontologies, the local names of elements are represented in the form of ID, such as *NCI_C12269* in NCI Thesaurus and *MA_0000216* in MA ontology, which is meaningless. Consequently, we first simply obtain the mapping results through comparing the label sets of the pairs of elements. The normalized edit distance similarity metric is applied to compute linguistic similarities between label sets:

$$\text{SIM}_{\text{name}} = \text{DNE}(s,t) = [\text{DE}(s,t)]/[\text{DE}(s,t) + \text{SE}(s,t)]$$

$$\text{SE}(s,t) = [(|s| + |t| - \text{DE}(s,t))]/2$$

where DE(s,t) denotes the edit distance between string s and t, and SE(s,t) denotes the edit similarity between s and t. The normalized edit distance similarity is denoted as DNE(s,t), and the function $|x|$ denotes the length of x. After that, mapping results are generated through a given threshold filtering and similarity ranking.

Terminology Matcher

Biomedical ontologies are characterized by terminological components in the form of names and various types of synonyms along with comments. We combine the labels with corresponding extensional clues to get the similarity between 2 elements. We chose term frequency-inverse document frequency (TF-IDF) to measure the similarity between the terminological documents of element e_s from source ontology and element e_t from target ontology:

$$\text{SIM}(x, y) = \text{TF*IDF}$$

$$\text{TF} = w/W$$

$$\text{IDF} = 1/2*(1+\log_2[\text{N}/n])$$

$$\text{SIM}_{\text{term}}(e_s,e_t) = \text{SIM}(Term[e_s], Term[e_t])$$

For each description document, w denotes the refined word occurrence; and W denotes the total refined occurrence among all the words in a specific document. And finally, we select the matching pairs with maximum similarity values.

Structure Matcher

The structural clues, including hierarchies (subclass, superclass, sibling class, disjoint class) and property attributes (domain, range), allow more possible candidate mappings to be discovered. The structural similarity of the element relies on the similarity of context descriptions. The extracted context set of each node may contain the terminological clues of direct neighbors, of adjacent nodes within local graph, that is, extracted semantic subgraph [52], or adjacent nodes in global graph. The similarity value of structural clues between each pair of nodes is initially measured by TF-IDF similarity between the structural documents:

$$\text{SIM}_{struc}(e_s,e_t)=\text{SIM}(Struct[e_s],Struct[e_t])$$

External Matcher

Two kinds of external resources are used to further promote the matching. From the general dictionary, synonyms are retrieved by the name of element. Meanwhile, from the domain-specific repository, equivalent classes are obtained based on terms of elements, and then discovering the synonyms of the discovered classes. In addition, cross-references are extracted based on the property *DbXref*. The extracted synonyms are viewed as the extensional comments of corresponding elements, and cross-references are used as the reference alignments. TF-IDF is used to determine the degree of each pair of collections of thesaurus information, respectively:

$$\text{SIM}_{ext}(e_s,e_t)=\text{SIM}(Ext[e_s],Ext[e_t])$$

Representation Learning Matcher

For the chosen pretrained language models, either BERT or BioBERT, and refined models, the input is a mention–synonym pair, and the outputs of the last hidden layer are concatenated to represent the mention. After getting the embeddings of ontological terms, we then use the cosine distance over the pairs of embedding representations as the similarity score:

$$\text{SIM}_{rps}(e_s,e_t)=\text{SIM}(Rps[e_s],Rps[e_t])$$

Hybrid Matcher

To obtain more accurate similarity values, the hybrid matchers are constructed through a fixed combination of simple matchers and other hybrid matchers. A straightforward strategy is summing up the values of all match results and getting the averages to denote the similarity between each pair of elements.

XSL•FO
RenderX

However, it may introduce lots of wrong mappings due to the neglect of differences between matchers. Therefore, we combine the similarity matrices by assigning varying weights to reflect their importance.

Matching Large Biomedical Ontologies

P-Anchors and N-Anchors

To deal with the scalability problem of large biomedical ontology matching, this paper extends the matching method based on reduction anchors with related nodes. The reduction anchors–based approach [26] is desirable for matching large ontologies. It utilizes positive reduction anchor (P-Anchors), based on the coherence of structural hierarchy of ontology alignment, and negative reduction anchor (N-Anchors), based on the locality characteristic of matching, to reduce undesirable comparisons. However, because matching based on P-Anchors is highly dependent on the hierarchical depth of ontology while usually the average depth of biomedical ontologies is typically limited, the matching cannot achieve the ideal high performance. Therefore, we extend reduction anchors with the related nodes to refine the matching. In this section, we first introduce the definition of our improved extended reduction anchors and then present the matching method based on extended reduction anchors.

Reduction Anchors

Extended Positive Reduction Anchor (P-AnchorE)

Given a concept a_i in ontology O_1 with equivalent concept set $a_{i1}, a_{i2}, \ldots, a_{im}$, let the similarities between a_i and concepts $b_1, b_2, \ldots, b_n$ in ontology O_1 be $S_{i1}, S_{i2}, \ldots, S_{in}$, respectively. If S_{ij} is larger than the predefined threshold *ptValue*, the concept pair (a_i, b_j) is a positive reduction anchor, and all positive reduction anchors about a_i are denoted by $\mathrm{PA}(a_i) = \{b_j | S_{ij} > \mathrm{ptValue}\}$. Then, the extended positive reduction anchor of a_i is [illegible].

Extended Negative Reduction Anchor (N-AnchorE)

Given a concept a_i in ontology O_1 with equivalent concept set $a_{i1}, a_{i2}, \ldots, a_{im}$, let the similarity values between a_i and concepts $b_1, b_2, \ldots, b_n$ in ontology O_2 be $S_{i1}, S_{i2}, \ldots, S_{in}$, respectively. If S_{ij} is smaller than the predefined threshold *ntValue*, the concept pair (a_i, b_j) is a negative reduction anchor, and all negative reduction anchors about a_i are denoted by $\mathrm{NA}(a_i) = \{b_j | S_{ij} < ntValue\}$. Then, the extended negative reduction anchors of a_i are as follows:

[illegible]

LOM-PE: Large Ontology Matching Algorithm Based on P-AnchorsE

Let $\mathrm{PSE}(a_i)$, the extended positive reduction set of a_i, be all the ignorable similarity calculations predicted by $\mathrm{PAE}(a_i)$. If $|\mathrm{PAE}(a_i)| > 0$, we select the top-*k* P-AnchorsE with maximum similarities. Let $\mathrm{PS}(a_i)$ be the initial positive reduction set about a P-AnchorE (a_i, b_j), which is calculated as follows:

[illegible]

Meanwhile, the reduction computation can be propagated to the concepts that are highly similar to $\mathrm{sub}(a_i)$. Therefore, $\mathrm{sub}(a_i)$ can be extended as follows:

[illegible]

Plus, $\mathrm{sup}(a_i)$, $\mathrm{sub}(b_j)$, and $\mathrm{sup}(b_j)$ can be calculated analogously. Then the extended reduction set of $\mathrm{PSE}(a_i|b_j)$ is:

[illegible]

If $\mathrm{PSE}(a_i) = \{b_1, b_2, \ldots, b_k\}$, the corresponding extended reduction set can be calculated as follows:

[illegible]

where *lub*() and *glb*() are the functions to obtain the least upper bound and greatest lower bound, respectively. The formula above indicates that smaller *top-k* will generate larger $\mathrm{PSE}(a_i)$. In our implementation, *top-k* is assigned a value from 1 to 4. The total positive reduction set during matching is:

[illegible]

Multimedia Appendix 1 presents a large ontology matching algorithm based on P-AnchorsE (LOM-PE). Here, LOMPE-Algorithm() is the main function, ComputerSim() matches elements on the hierarchy path recursively, and GetPAnchorsE() obtains top-*k* P-AnchorsE.

The time complexity of the LOM-PE algorithm is analyzed as follows: Given 2 matched ontologies, if all concepts are on a hierarchy path, the matching process can generate $n(n–2)$ size valid positive reduction set, and it just needs $2n$ similarity calculations, that is, the algorithm has the best time complexity $O(2n)$. However, such an ideal case almost does not exist in the real world. Suppose there are *m* hierarchy paths, then the average depth of the ontology is [illegible]. Consequently, we can derive the time complexity of Multimedia Appendix 1 as follows:

[illegible]

LOM-NE: Large Ontology Matching Algorithm Based on N-AnchorsE

The set of all ignorable similarity calculations predicted by N-AnchorsE is called the extended negative reduction set. Let $Nb(a_i) = \{a_x | d(a_x, a_i) <= nScale\}$ be the neighbors with *nScale* distance to a_i. Therefore, the initial negative reduction set generated by a_i is:

[illegible]

According to the formula, $\mathrm{NAE}(a_i)$ will be propagated to neighbors of a_i. And there are 3 constraints being introduced to reduce the risk of low credible negative reduction set: (1) all N-AnchorsE must be obtained in similarity calculating; (2) all N-AnchorsE of a_i can only be propagated to the neighbors in

XSL•FO
RenderX

the semantic subgraph of a_i; and (3) all N-AnchorsE of a_i can be propagated only if the description document of a_i contains more than t items.

Similar to LOM-PE, the reduction computation can also be propagated to the concepts that are highly similar with $Nb(a_i)$. Therefore, the extended neighbors set is as follows:

Then, the final extended negative reduction set can be denoted as:

where the extended set NSE() should also comply with above 3 constraints.

Multimedia Appendix 2 presents a large ontology matching algorithm based on N-AnchorsE (LOM-NE). All concepts are sorted by their degrees (line 2). If a similarity s is smaller than *ntValue* and satisfies 3 constraints (line 9), an N-AnchorE (line 10) is used to get the extended negative reduction set (line 12). After refining the extended negative reduction set, we obtain the valid extended negative reduction set (line 13). The time complexity of the algorithm is $O([1-w\lambda]n^2)$, where w is the average degree and λ is determined by *ntValue* and constraints. The bigger w and λ are, the higher performance the algorithm has.

LOM-Hybrid: Hybrid Large Ontology Matching Algorithm

We use a hybrid algorithm, called LOM-Hybrid, to combine the LOM-PE and LOM-NE algorithms to obtain as large a valid reduction set as possible. It can be a benefit for the LOM-PE algorithm if the LOM-NE algorithm that calculates the elements with large degree is implemented first: (1) Because the average depth of a real ontology is often small, while LOM-PE relies on the depth of concept hierarchy, the LOM-PE might not have an ideal performance. (2) The elements with large degree, most of which are located in the middle of hierarchy, would be calculated first by LOM-NE, and it can benefit the LOM-PE. Therefore, the LOM-Hybrid algorithm is mainly based on the framework of the LOM-NE algorithm, in which the LOM-PE algorithm is embedded. LOM-Hybrid can generate the valid positive reduction set and negative reduction set. Theoretically, the time complexity of LOM-Hybrid is between the complexity of LOM-NE and the complexity of LOM-PE. Indeed, it is very close to LOM-NE. However, in the real-world cases, the actual time complexity is indeed close to that of LOM-NE.

Results

Overview

We performed a comprehensive evaluation of the match processing strategies on real-word ontologies. The main goal is to investigate the impact of different combination strategies, that is, selection and aggregation of clues, on match quality, and to compare the effectiveness of different matchers, that is, single matcher and different combinations of individual matchers. We used Java to implement our approaches and conduct the experiments on a computer with an Intel Xeon 4110 CPU and 64-GB memory.

Data Set

Our experiments are conducted on 4 ontologies that appear in the Ontology Alignment Evaluation Initiative (OAEI). Two of them (the Adult Mouse Anatomy Ontology and the Foundational Model of Anatomy) are pure anatomical ontologies, while the other 2 (SNOMED-CT and NCI Thesaurus) are broader biomedical ontologies.

Adult Mouse Anatomy is a structured dictionary that provides standardized nomenclature for anatomical terms in the postnatal mouse and organizes anatomical structures for the postnatal mouse spatially and functionally [53].

Foundational Model of Anatomy (FMA) is an evolving computer-based knowledge source for biomedical informatics. The FMA is a domain ontology of the concepts and relationships that pertain to the structural organization of the human body [2].

NCI Thesaurus (NCI) provides reference terminology for many NCIs and other systems. It covers vocabulary for clinical care, translational and basic research, public information, and administrative activities [1].

SNOMED-CT is a systematically organized computer-processable collection of medical terms providing codes, terms, synonyms, and definitions used in clinical documentation and reporting [3].

The detailed statistics of each ontology matching task are presented in Table 2. For Anatomy, there are 2737 concepts in source ontology and 3298 concepts in target ontology, simultaneously including many labels and synonyms but only the *PART_OF* property with both ontologies. FMA-NCI task selects a small part of FMA and NCI ontology, with 3696 concepts from FMA and 6488 concepts from NCI, and FMA-SNOMED also selects a fragment of these ontologies with tens of thousands of concepts, 10,157 concepts in the source ontology FMA and 13,412 concepts in the target ontology SNOMED. For FMA-NCI and FMA-SNOMED, there exists no synonym within the ontologies but some properties to define the relations between entities, 24 properties for FMA, 63 properties for NCI, and 18 properties for SNOMED. For each concept, there are almost several aliases (labels) that are important for the alignments of heterogeneous ontologies. The evaluation of tasks is summarized through the MELT (Matching EvaLuation Toolkit) framework supported in OAEI. Actually, the alignments of tasks FMA-NCI and FMA-SNOMED are conducted on a small fragment of the aforementioned ontologies.

XSL•FO
RenderX

Table 2. Summary statistics of the biomedical ontology matching tasks.

Task and ontology	#Concepts	#Labels	#Synonyms	#Properties	#Triples
Anatomy					
MA	2737	3084	344	2	15,958
NCI[a]	3298	9403	5246	1	35,354
FMA-NCI					
FMA[b]	3696	9142	0	24	16,919
NCI	6488	17,109	0	63	64,857
FMA-SNOMED					
FMA	10,157	26,989	0	24	47,730
SNOMED	13,412	13,431	0	18	110,029

[a]NCI: National Cancer Institute.

[b]FMA: Foundation Model of Anatomy.

Measures

In order to measure the performance of the matching system, we selected precision, recall, and F-measure adapted for ontology matching evaluation.

We compare the mapping *M*, which consists of all those correspondences generated by our system, against reference mapping *R* to compute precision *p*, recall *r*, and F1-measure *F*. The standard measures for evaluating mappings are denoted as follows:

$$p(M,R)=(|M\cap R|)/M$$

$$r(M,R)=(|M\cap R|)/R$$

$$F(M,R)=[2\cdot p(M,R)\cdot r(M,R)]/[p(M,R) + r(M,R)]$$

Experiment Settings

We define several hybrid matchers in different combinations of atomic clues and matching dimensions. The details of designed matchers are listed in Table 3. For the clues in this table, *syn* means the sets of synonyms of concepts, *prop* is the abbreviations of property, *dh* denotes direct hierarchy utilizing the nearest neighbors, *lh* is the local hierarchy using structural clues within corresponding semantic subgraphs, and *gh* represents the global hierarchy that uses global structure based on transitive rules. In addition, *WN* denotes the general dictionary WordNet selected in our notion, and U_{dic} denotes the domain-specific dictionaries, UBERON and UMLS, in our experiments. For the representation learning clues, we choose BERT as the general model; and BioBERT and fine-tuned BioBERT denoted as fBio as the specialized representation models.

There are 3 terminological matchers, 5 structural matchers, 2 external matchers, and 3 representation learning matchers. The comment is absent in the data sets, so we eliminate it in atomic clues. In the ontologies of the Anatomy task, the local names of elements are in the form of ID, and for the largebio tasks, the names are wholly contained in labels. Thus, we mainly focus on the comparison of labels instead of names. Owing to the crucial role of terminological clues in ontology matching, the terminological matcher is constructed as the basis of matchers in the other dimensions.

XSL•FO
RenderX

Table 3. Relations between clues and matchers.

Matcher	Clue											
	Name	Label	syn	prop	dh	lh	gh	WN	U_{dic}	BERT[a]	BioBERT	fBio
Terminological												
M_1	✓											
M_2		✓										
M_3		✓	✓									
Structural												
M_4		✓	✓	✓								
M_5		✓	✓		✓							
M_6		✓	✓	✓	✓							
M_7		✓	✓			✓						
M_8		✓	✓				✓					
External												
M_9		✓	✓					✓				
M_{10}		✓	✓						✓			
Representation learning												
M_{11}		✓	✓									
M_{12}		✓	✓								✓	
M_{13}		✓	✓									✓

[a]BERT: Bidirectional Encoder Representations from Transformers.

Research Questions

We attempted to investigate the following research questions to understand the influence of different aggregations of ontology clues, to compare the effectiveness of different matcher combinations, and to verify the practical usefulness of extended reduction anchors.

Research Question 1 (Influence of the Combination Strategies of Clues): How Do the Different Combination Strategies of the Clues Perform in Ontology Matching?

The purpose of research question 1 is to investigate how a combination strategy influences the matching performance. There are generally several kinds of available clues in a single dimension. For instance, in the point of structure, there are intra structure and extra structure. However, some part of them may have a negative effect on the matching results. The study of research question 1 could help discover the influence of key clues during matching.

Research Question 2 (Effectiveness of Matcher Combinations): How Effective Are the Combinations of Matchers Implemented in Ontology Matching?

The purpose of research question 2 is to investigate whether utilizing the different aggregations of matchers could promote the matching effect, and to explore which combinations could be useful to match ontology. The study of research question 2 aims to learn how the integration of matchers influences the matching results.

Research Question 3 (Scalability of Reduction Anchors): What Is the Performance of Extended Reduction Anchors While Matching Large Biomedical Ontologies?

The purpose of research question 3 is to verify the effectiveness of our reduction anchors. Because of the large scale of biomedical ontologies, the matching is time-consuming and acquires amounts of space. As a result, there is a strong need to eliminate meaningless computations to reduce time and space complexity. The study of research question 3 is dedicated to demonstrate the effectiveness of resolving the scalability problem brought by the reduction anchor–based approach.

Results of Terminology-Based Matcher

In the initial phase, we examine the matcher with the direct linguistic description of concepts: name, labels, synonyms, and comments.

Table 4 shows the matching results by utilizing terminological clues in the ontologies. For the task of Anatomy, there is no matching to be obtained for the name being in ID format. It can be observed that labels act as a strong distinguishing feature for matching biomedical ontologies, and relying on the string similarity of labels can achieve a fundamental precision and F1 measure, particularly for Anatomy and FMA-NCI. Besides, integrating internal synonyms to get the name variants can further improve the performance of system. Although it slightly

XSL•FO
RenderX

decreases the precision of alignments, it can increase the recall and F1 measure distinctly. For the tasks FMA-NCI and FMA-SNOMED, there is no change in the metrics owing to the absence of synonyms in the input ontologies. In addition, we can find that the terminological matchers result in a substantial difference in the matching performance of different tasks. The terminological matcher achieves precision of 0.9658 and 0.9001 on Anatomy and FMA-NCI, respectively. But the precision on FMA-SNOMED is 0.3549. The Anatomy ontologies mainly focus on the anatomical terms of adult mouse anatomy and human anatomy whose terminological names are highly similar. By contrast, for the tasks of FMA-NCI and FMA-SNOMED, the ontologies cover a variety of topics and name the objects in different criteria, which cause the difference between their phenotypic representations.

Table 4. Results of terminology-based matcher.[a]

Method	Anatomy			FMA[b]-NCI[c]			FMA-SNOMED		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
M_1	0	0	—	86.30	53.74	66.23	25.22	28.14	26.60
M_2	**96.58**	68.87	80.40	**90.01**	**70.90**	**79.32**	**35.49**	**36.59**	**36.03**
M_3	92.28	**74.14**	**82.22**	**90.01**	**70.90**	**79.32**	**35.49**	**36.59**	**36.03**

[a]Values in bold indicate best experimental results.

[b]FMA: Foundation Model of Anatomy.

[c]NCI: National Cancer Institute.

Results of Structure-Based Matcher

There are 5 structure-based matchers which consider different structural clues of ontologies. First, we measured the effect of property and hierarchy, respectively, while only the nearest neighbors are considered in the hierarchy. Then, we evaluated the results using a combination of both. In addition, the transitive closure is taken into account to get the global structure with predefined decay coefficients. Furthermore, we assess the results of the structural matcher which has the local structure within the constructed semantic subgraph for each concept.

Table 5 shows the matching results of different combination strategies of structural atomic clues. Specifically, we use G to denote the F1 measure difference between structure-based matchers and the best terminology–based matcher. Overall, G demonstrates that the structure has a positive impact on matching performance. In the Anatomy task, the property information slightly degrades the mapping results, the reason is that there are only 2 properties, *UNDEFINED_is_a* and *UNDEFINED_part_of*, which bring valueless information. It is evident that the hierarchical structure has a positive effect, which improves the F1 score by about 4%. When we choose the entire hierarchies extended with transitive rules, the recall and F1 measure fall a little compared with the direct structure that combines the clues of direct nodes, for the reason that it would bring about some redundancy and much more noise by integrating too many hierarchical clues. Furthermore, the recall achieves an impactful enhancement while utilizing extracted semantic subgraphs to capture the real meaning of concepts, but is accompanied by a decline in precision compared with some other matchers. In the task of FMA-NCI, the local structural clues can result in about 2% improvement. However, there is hardly an obvious change when we combine other different hierarchical information. Because of the relatively obvious morphological difference between different ontologies, utilizing the local structure to discover more semantically related entities could result in better performance compared with direct structural clues. Because of the large size of FMA-SNOMED, which takes too much time to recursively retrieve global structure and construct the semantic subgraphs, the results of M_7 and M_8 are ignored in our experiments. However, utilizing all clues cannot guarantee the improvement of mappings. For instance, exploiting the property has a little positive effect than extending clues with only the direct linking nodes.

XSL•FO
RenderX

Table 5. Results of structure-based matcher.[a]

Method	Anatomy				FMA[b]-NCI[c]				FMA-SNOMED			
	P (%)	R (%)	F1 (%)	G (%)	P (%)	R (%)	F1 (%)	G (%)	P (%)	R (%)	F1 (%)	G (%)
M_4	90.66	74.14	81.56	–0.66	88.36	72.82	79.85	+0.53	37.21	41.59	39.27	+3.24
M_5	91.82	81.40	**86.30**	**+4.08**	88.96	72.48	79.88	+0.56	37.25	41.68	39.34	+3.31
M_6	**92.23**	80.61	86.03	+3.81	**89.62**	72.28	80.03	+0.71	**37.29**	**41.88**	**39.46**	**+3.45**
M_7	88.04	**84.56**	86.27	+4.05	88.25	**74.93**	**81.05**	**+1.73**	—[d]	—	—	—
M_8	91.93	80.41	85.78	+3.56	89.36	72.49	80.03	+0.71	—	—	—	—

[a]Values in bold indicate best experimental results.

[b]FMA: Foundation Model of Anatomy.

[c]NCI: National Cancer Institute.

[d]Not available.

Results of External-Based Matcher

This section studies the performance of external-based matchers utilizing general dictionaries (WordNet) and external domain-specific knowledge (UBERON and UMLS). The experimental results of the 2 methods M_9 and M_{10} are presented in Table 6.

We obtain the precedent sense of names through WordNet to enrich the synonyms of ontology concepts. Because WordNet is difficult to get relevant synonyms unless the sense of the term is known a priori and that compound terms are strongly covered, it brings a negative influence on all the 3 tasks. Then, UBERON and UMLS, which are related to biomedical science, are selected to further enrich ontology descriptions. On the one hand, we acquire all the correlative synonyms and cross-search references about the input ontologies. On the other, a reverse synonym lexicon is constructed, which is initiated by the idea that there may be a lack of description $Syn(b) = a$ while $Syn(a) = b$ exists. It can be observed that the specialized lexicon produces an effective influence compared with the common repositories. For Anatomy, specialized lexicon brings an increase of 8.3%, while the common lexicon causes a 1.33% decrease in F1 score. However, the domain-specific lexical brings about a much less positive effect on mapping results of FMA-NCI with an increase of 1.33%, and an increase of 8.3% and 13.49% for Anatomy and FMA-SNOMED, respectively. The synonyms and cross-search references are extracted from the auxiliary knowledge with the terminological names directly. However, there are a few available auxiliary clues for the mapping of FMA and NCI, which accounts for the little rise of the FMA-NCI task.

Table 6. Results of external-based matcher.[a]

Method	Anatomy				FMA[b]-NCI[c]				FMA-SNOMED			
	P (%)	R (%)	F1 (%)	G (%)	P (%)	R (%)	F1 (%)	G (%)	P (%)	R (%)	F1 (%)	G (%)
M_9	90.05	73.42	80.89	–1.33	79.50	**74.01**	76.66	–2.66	32.38	36.48	34.31	–1.72
M_{10}	**92.67**	**88.46**	**90.52**	**+8.3**	**89.65**	73.47	**80.75**	**+1.33**	**46.78**	**52.60**	**49.52**	**+13.49**

[a]Values in bold indicate best experimental results.

[b]FMA: Foundation Model of Anatomy.

[c]NCI: National Cancer Institute.

Results of Representation Learning–Based Matcher

We select 3 pretrained language models to study the influence of word embedding. Table 7 shows the comparison of different word embedding techniques applied to biomedical ontology matching. Surprisingly, although BERT is trained by amounts of general corpora, it still has resulted in a slight decline in the results. It seems that capturing implicit semantics in a specific domain could be more difficult than we imagined. BioBERT is a domain-specific language representation model pretrained on large-scale biomedical corpora. It shows a slightly better performance than BERT, indicating that incorporating domain-specific language representation can be valuable to ontology matching. The fine-tuned BioBERT is trained on the correspondence between concepts and synonyms within UBERON which is much more relevant to the tasks. The results show that the fine-tuned BioBERT is able to produce much more positive influence on capturing the semantics in ontologies. In fact, the cost of representation learning technique outweighs the benefit that it can bring to our mapping tracks. Especially for the FMA-NCI, it results in a negative effect on the mapping performance for the reason that the training corpora is less relevant to FMA-NCI. In addition, there are restrictive synonyms within the ontologies of FMA-NCI and FMA-SNOMED, which is insufficient for training of representative learning matchers.

XSL•FO
RenderX

Table 7. Results of representation learning–based matcher.[a]

Method	Anatomy				FMA[b]-NCI[c]				FMA-SNOMED			
	P (%)	R (%)	F1 (%)	G (%)	P (%)	R (%)	F1 (%)	G (%)	P (%)	R (%)	F1 (%)	G (%)
M_{11}	89.55	74.68	81.44	–0.78	87.94	71.47	78.85	–0.47	33.54	37.66	35.47	–0.56
M_{12}	90.21	**76.32**	82.69	+0.47	**88.19**	71.93	79.23	–0.09	**35.04**	37.08	36.03	+0.00
M_{13}	**93.31**	74.39	**83.02**	**+0.8 0**	86.84	**72.97**	**79.30**	**–0.02**	34.22	**38.63**	**36.29**	**+0.26**

[a]Values in bold indicate best experimental results.

[b]FMA: Foundation Model of Anatomy.

[c]NCI: National Cancer Institute.

Results of Matcher Combinations

After evaluating the single matchers in different dimensions, we conducted a series of experiments to examine the effectiveness of different combination strategies. For each selection strategy, we choose the optimal parameter range, in which the best match result is to be expected.

Table 8 shows the performance of hybrid matchers combined with different matchers. There are 7 hybrid matchers used in our experiments, where *Term*, *Struc*, *Ext*, and *Rps* represent the terminological matcher, structural matcher, external matcher, and representation learning matcher, respectively. Different combinations have different influences on different tracks, and integrating some matchers may thus exert negative effect on matching. We can observe that the combination of all 4 matchers, that is, *Term* + *Struc* + *Ext* + *Rps*, can achieve the best performance for Anatomy, while incorporating *Term*, *Struc*, and *Ext* together leads to the best results for FMA-NCI and FMA-SNOMED. As the representation learning matcher is trained with the synonym marginalization algorithm while there are less synonym clues within the ontologies of FMA-NCI and FMA-SNOMED compared with Anatomy, the *Rps* may be less helpful than the other 3 matchers for FMA-NCI and FMA-SNOMED.

Table 8. Results of hybrid matcher.[a]

Method	Anatomy			FMA[b]-NCI[c]			FMA-SNOMED		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Term	92.28	74.14	82.22	90.01	70.90	79.32	35.49	36.59	36.03
Term + Struc	92.23	80.61	86.03	88.25	74.93	81.05	37.29	41.88	39.46
Term + Ext	92.67	88.46	90.52	89.65	73.47	80.75	46.78	*52.60*	49.52
Term + Rps	93.31	74.39	83.02	87.94	71.47	78.85	33.54	37.66	35.47
Term + Struc + Ext	93.78	90.44	92.08	*90.64*	*75.17*	*82.18*	*47.97*	52.21	*50.00*
Term + Ext + Rps	91.56	88.92	90.22	88.90	72.95	80.14	43.76	47.83	45.70
Term + Struc + Ext + Rps	*94.95*	*90.57*	*92.71*	90.04	74.47	81.52	47.52	52.01	49.66

[a]Italicized values indicate best experimental results.

[b]FMA: Foundation Model of Anatomy.

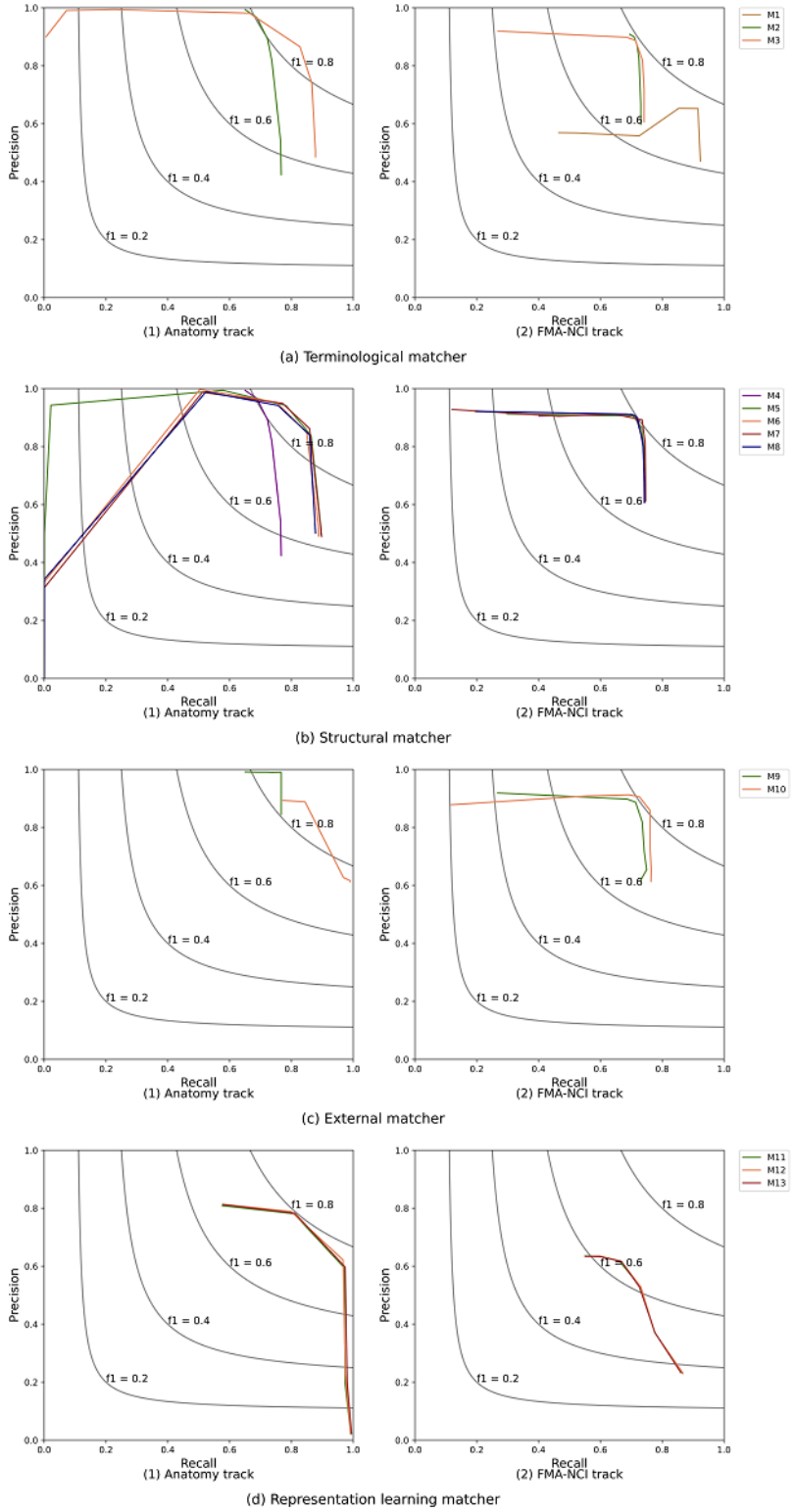
[c]NCI: National Cancer Institute.

Performance Evaluation of Matchers

Here we present the metrics of precision and recall along with F1 measure of each matcher and analyze the correlation between them. In general, the higher the precision, the lower is the recall. However, there are also some exceptions when the similarity threshold is high enough. In Figure 3, the comparisons in the precision–recall space over 4 aspects of Anatomy and FMA-NCI track are depicted. From Figure 3, the terminological matcher that combines labels and synonyms could achieve the best results. For the structural matchers, most of them perform similarly and there is no obvious difference among them. For the FMA-NCI task, there is slightly a little difference in matching performance between most matchers. As for the external matchers, M_{10} utilizing domain-specific lexical has a significant difference with M_9. However, for the matchers that are using the method based on representation learning, the precision–recall curve is obviously similar. This is because similar names may refer to different objects, while names in different morphologies may refer to the same object. Therefore, exploiting only the semantic representations remains hard to capture the true meaning for these elements of specific domain.

XSL•FO
RenderX

Figure 3. Comparison in PR space.



Effectiveness of Extended Reduction Anchor

To examine the validity of the extended reduction anchors, we conduct comparison and evaluation through integrating extended reduction anchors with the best performing matchers. The results are indicated in Table 9.

We can learn from the table that our improved reduction anchors could benefit the time complexity during matching. According to the results shown in Table 9, the reduction anchors are more comparative while the ontology size is much larger. For Anatomy, reduction anchors–based approach does not demonstrate distinct advantages with the middle-size ontologies. However, for the track of FMA-NCI, especially for FMA-SNOMED, the runtime has been notably reduced compared with the other matchers. Besides, reduction anchors barely bring metric loss while simultaneously promoting the efficiency of matching. While integrating the reduction anchors together, it is effective to skip large numbers of ignorable

XSL•FO
RenderX

similarity computations and is efficient to reduce the time complexity.

Table 10 presents the comparison of LOM-RAE with LOM-RA, which demonstrates the superiority of our extended reduction anchors compared with previous reduction anchors. From Table 10, we can observe that RAE is effective to skip much more similarity computations than RA. Because the similarity threshold is set properly high and the reduction set is obtained through strictly abiding by the defined constraints, it is evident there is almost little or no loss in performance with considerable matching comparisons being omitted.

In addition, to examine the practicability of the extended reduction anchors, we compare our matching approach based on RAE with some other systems participating in OAEI, for example, AML [16], LogMap [35], Wiktionary [54], and ALOD2Vec [55]. The execution times of these systems are shown in Table 11. In contrast to the matching systems utilizing modulization and clustering, such as AML and LogMap, the reduction set is generated dynamically based on the similarity calculations of entity pairs, which require much more time during matching. Nevertheless, it can be observed that our proposed approach can still achieve promising performance among these matching systems, which demonstrates that RAE is practicable and effective for large-scale ontology matching scenario.

Table 9. Effectiveness of extended reduction anchors.[a,b]

Task and matcher	P (%)	R (%)	F1 (%)	Time (minutes)
Anatomy				
Term	92.28	74.14	82.22	1.5
Term + Struc	91.82	81.40	86.30	3.0
Term + Ext	92.67	88.46	90.52	4.9
Term + Rps	93.31	74.39	83.02	2.8
Term + Struc + Ext	93.78	90.44	92.08	7.1
Term + Struc + Ext + Rps	**94.95**	**90.57**	**92.71**	8.9
Term + Struc + Ext + RAE	94.74	90.36	92.50	**0.7**
FMA[c]-NCI[d]				
Term	90.01	70.90	79.32	11.4
Term + Struc	89.36	73.87	80.88	47.9
Term + Ext	92.67	**88.46**	**90.52**	18.6
Term + Rps	**93.31**	74.39	83.02	12.8
Term + Struc + Ext	89.65	76.92	82.80	56.8
Term + Struc + Ext + Rps	90.04	76.71	82.84	65.9
Term + Struc + Ext + RAE	89.65	75.56	82.00	**2.8**
FMA-SNOMED				
Term	35.49	36.59	36.03	64.4
Term + Struc	37.21	41.59	39.27	85.6
Term + Ext	46.78	52.60	49.52	113.4
Term + Rps	**93.31**	**74.39**	**83.02**	117.1
Term + Struc + Ext	47.97	52.21	50.00	161.0
Term + Struc + Ext + Rps	47.52	52.01	49.66	227.5
Term + Struc + Ext + RAE	53.41	51.25	52.31	**12.5**

[a]The best performing matcher is italicized.

[b]Values in bold indicate best experimental results.

[c]FMA: Foundation Model of Anatomy.

[d]NCI: National Cancer Institute.

XSL•FO
RenderX

Table 10. Effectiveness of extended reduction anchors.[a]

Task and matcher	P (%)	R (%)	F1 (%)	Time (minutes)
Anatomy				
LOM-RA	*95.41*	*90.64*	*92.96*	2.2
LOM-RAE	94.28	90.36	92.50	*0.7*
FMA[b]-NCI[c]				
LOM-RA	*90.01*	*75.61*	*82.18*	10.4
LOM-RAE	89.65	75.56	82.00	*2.8*
FMA-SNOMED				
LOM-RA	*53.77*	50.89	52.29	42.7
LOM-RAE	53.41	*51.25*	*52.31*	*12.5*

[a]Italicized values indicate best experimental results.
[b]FMA: Foundation Model of Anatomy.
[c]NCI: National Cancer Institute.

Table 11. Execution time (minutes) of systems.[a]

Matching system	Anatomy	FMA[b]-NCI[c]	FMA-SNOMED
AML	0.48	0.63	1.68
LogMap	0.01	0.03	0.87
Wikitionary	1.08	4.3	11.62
ALOD2Vec	3.93	2.97	—
LOM-RAE	*0.7*	*2.8*	*12.5*

[a]Italicized values indicate best experimental results.
[b]FMA: Foundation Model of Anatomy.
[c]NCI: National Cancer Institute.

Performance of Extended Reduction Anchors

There is a need to analyze the influence of key parameters of our proposed reduction anchors. Here we use a new metric called benefit rate (G) to measure how much an LOM algorithm can improve the performance: $G = N/(n_1 * n_2)$, where N is the size of the total reduction set; and n_1 and n_2 represent the number of concepts in 2 ontologies. The larger the value of G, fewer the times of similarity calculations required and the higher the efficiency of the algorithm.

The LOM-NE algorithm has 4 important parameters: *ntValue*, *nScale*, SDD constraint, and SSG constraint. We evaluate these parameters on the Anatomy data set. Figure 4 shows the relation between *ntValue* and F1 measure on different *nScales*. Figure 5 shows the relation between benefit rate and *ntValue* under different *nScales*. We observe that (1) *ntValue* has a certain effect on matching quality and efficiency, that is, different *ntValues* will lead to some fluctuation of matching quality. Meanwhile, LOM-NE also causes a higher benefit rate with an increase of *ntValue*. (2) *nScale* also affects matching quality and efficiency. As *nScale* increases, matching quality will decrease, but the benefit rate will increase to a certain extent. Results also show that *ntValue* = 0.15 and *nScale* = 3 will lead to a good matching quality and benefit rate. Figures 6 and 7 show the influences of SDD and SSG on matching quality and benefit rate. The W/A constraint represents the results without any constraint. We can see that (1) under 3 constraints, the matching quality will increase, but the benefit rate will decrease; (2) the SDD constraint has a higher influence on matching quality and benefit rate.

XSL•FO
RenderX

Figure 4. ntValue-nScale-matching quality.

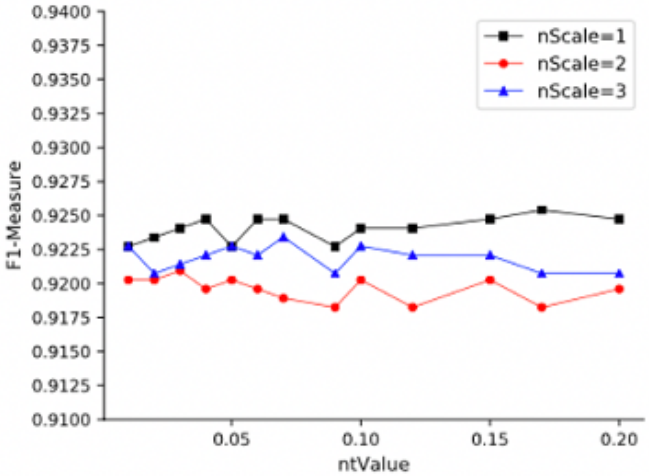


Figure 5. ntValue-nScale-benefit rate.

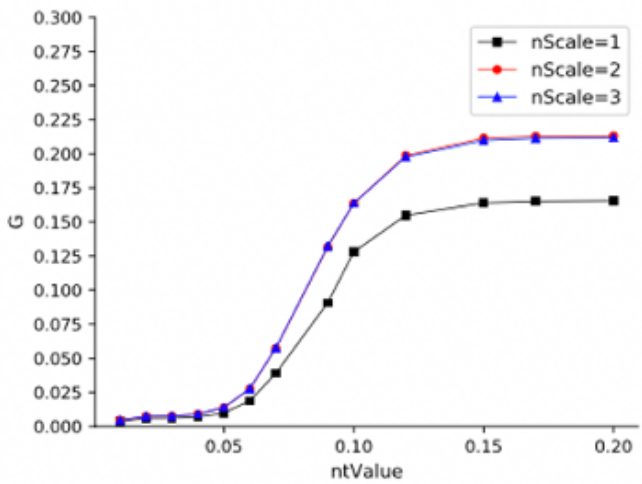
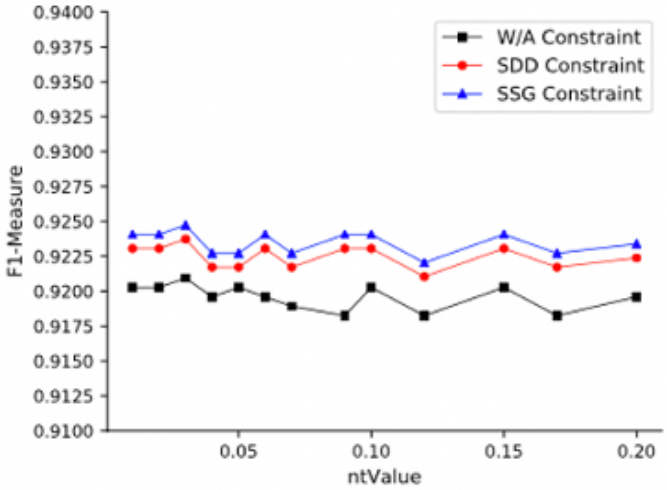
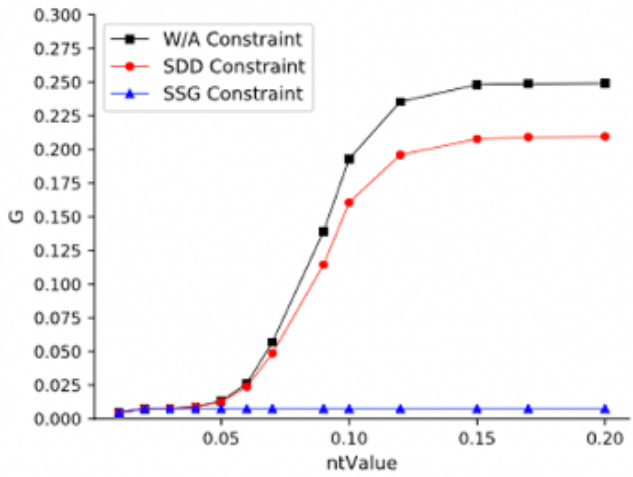


Figure 6. SDD-SSG-matching quality.



XSL•FO
RenderX

Figure 7. SDD-SSG-benefit rate.



Discussion

Principal Findings

In this section, we discuss our experimental results according to the research questions. First, we will analyze the influence of the matchers in a single dimension. Second, we will report on how information imposes an effect on the final performance compared with the distinguishing clues of concepts. Finally, we will illustrate the practical benefits of adopting reduction anchors in large biomedical ontology matching.

How Do the Different Combination Strategies of the Clues Perform in Ontology Matching?

After analyzing the results of matchers in 4 predefined dimensions separately, it is obvious that there are some parts playing an inessential role in the matching process. According to the results presented in Table 5, it can be observed that property degrades the performance in the Anatomy track, and using the transitive rule in hierarchy to gain more structural presentations would also input noise to mappings. For all these 3 tracks, the structural clues play a significant role in the matching process and bring about a positive improvement. While using external knowledge as auxiliary resources, the general dictionary, such as WordNet, has resulted in a less positive impact than the biomedical lexical, such as UBERON, as illustrated in Table 6. WordNet is a dictionary that works in the general domain, and may be deficient in synonymy for biomedical concepts or generate erroneous synonymy.

When one integrates the semantic embedding method into biomedical ontology matching, results from Table 7 suggest that despite the ability of the former to capture the underlying potential semantic, it can also worsen the results. Besides, the BERT model used in the common domain is incapable of catching the semantic in biomedicine (Table 7). The fine-tuned BioBERT model trained on data sets related to the test suite is much more competent than BioBERT. Therefore, mining and combining the key clues from ontologies and auxiliary sources are more important compared with utilizing the whole sources.

How Effective Are the Combinations of Matchers Implemented in Ontology Matching?

According to the results illustrated in Table 8, the combination strategies of matcher have different levels of impact on the tracks. It is evident that although large amounts of information could be mined from ontologies, some may result in scarce improvement and bring about an increase in time complexity at the same time. It can be observed that incorporating the structural matcher and the external matcher could have momentous benefits to biomedical ontology matching. The representation learning matcher is able to boost the performance of Anatomy and FMA-BCI to some extent, but it causes a decline in the performance of FMA-SNOMED. As a result, combining all matchers is not helpful to promote mappings. Thus, for different tasks, matchers may exert different effects, either positive or negative.

What Is the Performance of the Proposed Reduction Anchors While Matching Large Biomedical Ontologies?

The results listed in Tables 9-11 demonstrate that reduction anchors are effective in reducing the running time during large ontology matching, with the superiority becoming more comparative when the volume of ontology is much larger. Reduction sets leverage the hierarchy concept to skip subsequent matching between subconcepts of one concept and super-concepts of the other concept, which also include the extended highly related concept nodes. By contrast, if 2 concepts have low similarity, based on the locality phenomenon of matching, it can skip subsequent matching between 1 concept and the neighbors of the other concept as well as the concepts with high similarity. When the ontology is large, the structure of ontology graph becomes complicated which would possess deeper hierarchical levels. Therefore, extended reduction anchors are able to and practicable to skip more unnecessary computations.

Conclusions

In this paper, we presented an empirical study of biomedical ontology matching based on a number of experiments performed on terminology-based, structure-based, external

XSL•FO
RenderX

knowledge–based, representation learning–based measures in detail. Biomedical ontology matching relying on the terminological description of elements, combined with a structural, external knowledge, and embedding similarity approach, is effective for the matching of ontologies to some extent. According to our results, composite matchers are very effective. Despite the imprecision of single matchers, their combinations are impressive in improving the mapping quality, and bring about more accurate and stable similarity for biomedical ontologies. Structural and external clues are proved to produce better match results and support good precision as they could best compensate the shortcomings of single terminological matchers.

We can also find that the knowledge information has either a neutral or a negative impact on the F measure (as shown in Tables 4-7), which suggests that this result is an artifact. It is obvious that utilizing all the clues cannot always achieve best performance in ontology matching. The hierarchical interpretations play an important role in the matching task of Anatomy, whereas in FMA-NCI, they exert little influence, which relies rather much on terminologies. Using the WordNet dictionary to retrieve alternative labels for concepts only has a weak effect compared with domain-specific resources UBERON and UMLS. Furthermore, representation learning techniques are relatively effective to improve recall. However, it still needs to be further deepened and enhanced. Based on the results produced in the experiments, we can learn that utilizing credible and distinguishable clues can effectively boost the ontology matching process as compared with matching 2 ontologies with all.

Moreover, we propose a new, efficient, large ontology matching method based on extended reduction anchors. The proposed approach is generic and could be applied to different fields. RAE is applied to predict the ignorable similarity calculations in ontology matching. Our experimental results also overwhelmingly demonstrate that the proposed method presents significant and encouraging improvement, especially in runtime efficiency.

Acknowledgments

The work was supported by the National Key R&D Program of China (2018YFD1100302), National Natural Science Foundation of China (No.61972082), and All-Army Common Information System Equipment Pre-Research Project (No. 31511110310, No. 31514020501, No. 31514020503).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Algorithm 1.
[PDF File (Adobe PDF File), 394 KB - medinform_v9i8e28212_app1.pdf]

Multimedia Appendix 2
Algorithm 2.
[PDF File (Adobe PDF File), 385 KB - medinform_v9i8e28212_app2.pdf]

References

1. Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's Thésaurus and Ontology. Journal of Web Semantics 2003 Dec;1(1):75-80. [doi: 10.1016/j.websem.2003.07.007]
2. Rosse C, Mejino JL. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform 2003 Dec;36(6):478-500 [FREE Full text] [doi: 10.1016/j.jbi.2003.11.007] [Medline: 14759820]
3. Schulz S, Cornet R, Spackman K. Consolidating SNOMED CT's ontological commitment. Applied Ontology 2011 Jan;6(1):1-11. [doi: 10.3233/ao-2011-0084]
4. Cimino JJ, Zhu X. The practical impact of ontologies on biomedical informatics. Yearb Med Inform 2006:124-135. [Medline: 17051306]
5. Isern D, Sánchez D, Moreno A. Ontology-driven execution of clinical guidelines. Comput Methods Programs Biomed 2012 Aug;107(2):122-139. [doi: 10.1016/j.cmpb.2011.06.006] [Medline: 21752487]
6. De Potter P, Cools H, Depraetere K, Mels G, Debevere P, De Roo J, et al. Semantic patient information aggregation and medicinal decision support. Comput Methods Programs Biomed 2012 Nov;108(2):724-735. [doi: 10.1016/j.cmpb.2012.04.002] [Medline: 22640816]
7. Xue X. A compact firefly algorithm for matching biomedical ontologies. Knowl Inf Syst 2020 Feb 08;62(7):2855-2871. [doi: 10.1007/s10115-020-01443-6]
8. Euzenat J, Shvaiko P. Ontology Matching. Heidelberg, Germany: Springer; 2007.
9. Bergamaschi S, Castano S, Vincini M, Beneventano D. Semantic integration of heterogeneous information sources. Data & Knowledge Engineering 2001 Mar;36(3):215-249. [doi: 10.1016/s0169-023x(00)00047-1]

XSL•FO
RenderX

10. Embley DW, Jackman D, Xu L. Attribute match discovery in information integration: exploiting multiple facets of metadata. J Braz Comp Soc 2002 Nov;8(2):32-43. [doi: 10.1590/s0104-65002002000200004]
11. Madhavan J, Bernstein PA, Rahm E. Generic schema matching with cupid. In: VLDB '01: Proceedings of the 27th International Conference on Very Large Data Bases. New York, NY: Association for Computing Machinery; 2001 Sep Presented at: VLDB '01: 27th International Conference on Very Large Data Bases; September 11, 2001; Roma, Italy p. 49-58. [doi: 10.5555/645927.672191]
12. Giunchiglia F, Shvaiko P. Semantic matching. The Knowledge Engineering Review 2004 May 13;18(3):265-280. [doi: 10.1017/s0269888904000074]
13. Giunchiglia F, Shvaiko P, Yatskevich M. S-Match: An algorithm and an implementation of semantic matching. 2004 May Presented at: Proceedings of the 1st European Semantic Web Symposium (ESWS); May 14, 2004; Crete, Greece p. 61-75.
14. Giunchiglia F, Shvaiko P, Yatskevich M. Semantic schema matching. 2005 Presented at: OTM Confederated International Conferences; October 31-November 4, 2005; Berlin, Heidelberg p. 347-365. [doi: 10.1007/11575771_23]
15. Jean-Mary YR, Shironoshita EP, Kabuka MR. Ontology matching with semantic verification. Journal of Web Semantics 2009 Sep;7(3):235-251. [doi: 10.1016/j.websem.2009.04.001]
16. Faria D, Pesquita C, Santos E. The agreementmakerlight ontology matching system. Heidelberg, Germany: Springer; 2013 Presented at: Proceedings of the 12th OTM Confederated International Conferences; September 13, 2013; Graz, Austria. [doi: 10.1007/978-3-642-41030-7_38]
17. Jiménez-Ruiz E, Grau BC. Logmap: Logic-based and scalable ontology matching. 2011 Presented at: Proceedings of the 10th International Semantic Web Conference; 2011; Bonn, Germany p. 23-27. [doi: 10.1007/978-3-642-25073-6_18]
18. Gracia J, Lopez V, d'Aquin M. Solving semantic ambiguity to improve semantic web based ontology matching. 2007 Nov 11 Presented at: Proceedings of the 2nd International Workshop on Ontology Matching (OM-2007) Collocated with the 6th International Semantic Web Conference (ISWC-2007) and the 2nd Asian Semantic Web Conference (ASWC-2007); November 11, 2007; Busan, Korea URL: http://ceur-ws.org/Vol-304/ [doi: 10.1007/978-3-540-76298-0_19]
19. Kolyvakis P, Kalousis A, Kiritsis D. Deepalignment: Unsupervised ontology matching with refined word vectors. Stroudsburg, PA: Association for Computational Linguistics; 2018 Jun Presented at: Proceedings of 16th Conference of the North American Chapter of the Association for Computational Linguistics; June 1-6, 2018; New Orleans, LA p. 01-06. [doi: 10.18653/v1/n18-1072]
20. Kolyvakis P, Kalousis A, Smith B, Kiritsis D. Biomedical ontology alignment: an approach based on representation learning. J Biomed Semantics 2018 Aug 15;9(1):21 [FREE Full text] [doi: 10.1186/s13326-018-0187-8] [Medline: 30111369]
21. Ferré A, Deléger L, Zweigenbaum P, Nédellec C. Combining rule-based and embedding-based approaches to normalize textual entities with an ontology. Paris, France: European Language Resources Association; 2018 May Presented at: Proceedings of the 11th International Conference on Language Resources and Evaluation; May 7-12, 2018; Miyazaki, Japan.
22. Li WZ, Duan XX, Wang M, Zhang XP, Qi GL. Multi-view embedding for biomedical ontology matching. 2019 Oct 26 Presented at: Proceedings of the 14th International Workshop on Ontology Matching collocated with the 18th International Semantic Web Conference; October 26, 2019; Auckland, New Zealand.
23. Hu W, Qu Y, Cheng G. Matching large ontologies: A divide-and-conquer approach. Data & Knowledge Engineering 2008 Oct;67(1):140-160. [doi: 10.1016/j.datak.2008.06.003]
24. Algergawy A, Massmann S, Rahm E. A clustering-based approach for large-scale ontology matching. Heidelberg, Germany: Springer; 2011 Sep Presented at: Proceedings of 15th East European Conference on Advances in Databases and Information Systems; September 20-23, 2011; Vienna, Austria. [doi: 10.1007/978-3-642-23737-9_30]
25. Pathak J, Johnson TM, Chute CG. Survey of modular ontology techniques and their applications in the biomedical domain. Integrated Computer-aided Engineering 2009 Jun 22;16(3):225-242. [doi: 10.3233/ica-2009-0315]
26. Wang P, Zhou Y, Xu B. Matching large ontologies based on reduction anchors. Palo Alto, CA: AAAI Press; 2011 Jul Presented at: Proceedings of 22nd International Joint Conference on Artificial Intelligence; July 2011; Barcelona, Catalonia, Spain p. 16-22.
27. Lambrix P, Tan H. SAMBO—A system for aligning and merging biomedical ontologies. Journal of Web Semantics 2006 Sep;4(3):196-206. [doi: 10.1016/j.websem.2006.05.003]
28. Hu W, Qu Y. Falcon-AO: A practical ontology matching system. Journal of Web Semantics 2008 Sep;6(3):237-239. [doi: 10.1016/j.websem.2008.02.006]
29. Melnik S, Garcia-Molina H, Rahm E. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. New York, NY: IEEE; 2002 Jun Presented at: Proceedings of the 18th International Conference on Data Engineering; February 26-March 1, 2002; San Jose, CA. [doi: 10.1109/icde.2002.994702]
30. Gulić M, Vrdoljak B, Banek M. CroMatcher: An ontology matching system based on automated weighted aggregation and iterative final alignment. Journal of Web Semantics 2016 Dec;41:50-71. [doi: 10.1016/j.websem.2016.09.001]
31. Zhao M, Zhang S, Li W, Chen G. Matching biomedical ontologies based on formal concept analysis. J Biomed Semantics 2018 Mar 19;9(1):11 [FREE Full text] [doi: 10.1186/s13326-018-0178-9] [Medline: 29554977]
32. Miller GA. WordNet. Commun ACM 1995 Nov;38(11):39-41. [doi: 10.1145/219717.219748]

XSL•FO
RenderX

33. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
34. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 2009 Jul;37(Web Server issue):W170-W173 [FREE Full text] [doi: 10.1093/nar/gkp440] [Medline: 19483092]
35. Jiménez-Ruiz E, Grau B, Cross V. LogMap family participation in the OAEI 2017. 2017 Oct 21 Presented at: Proceedings of the 12th International Workshop on Ontology Matching collocated with the 16th International Semantic Web Conference; October 21, 2017; Vienna, Austria.
36. Annane A, Bellahsene Z, Azouaou F, Jonquet C. Selection and combination of heterogeneous mappings to enhance biomedical ontology matching. Cham, Switzerland: Springer; 2016 Nov Presented at: Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management; 2016 Nov 19-23; Bologna, Italy. [doi: 10.1007/978-3-319-49004-5_2]
37. Faria D, Pesquita C, Santos E, Cruz IF, Couto FM. Automatic background knowledge selection for matching biomedical ontologies. PLoS One 2014 Nov 7;9(11):e111226 [FREE Full text] [doi: 10.1371/journal.pone.0111226] [Medline: 25379899]
38. Zhang Y, Wang X, Lai S, He S, Liu K, Zhao J, et al. Ontology matching with word embeddings. In: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Cham, Switzerland: Springer; 2014:34-45.
39. Wang L, Bhagavatula C, Neumann M, Lo K, Wilhelm C, Ammar W. Ontology alignment in the biomedical domain using entity definitions and context. arXiv. 2018. URL: https://arxiv.org/abs/1806.07976 [accessed 2021-07-28]
40. Grau BC, Parsia B, Sirin E, Kalyanpur A. Modularity and Web Ontologies. 2006 Jun. URL: https://www.aaai.org/Papers/KR/2006/KR06-022.pdf [accessed 2021-07-28]
41. Del Vescovo C, Parsia B, Sattler U, Schneider T. The modular structure of an ontology: Atomic decomposition. Palo Alto, CA: AAAI Press; 2011 Jul Presented at: Proceedings of the 22nd International Joint Conferences on Artificial Intelligence; July 16-22, 2011; Barcelona, Catalonia, Spain.
42. Grau BC, Horrocks I, Kazakov Y, Scattler U. Just the right amount: Extracting modules from ontologies. New York, NY: Association for Computing Machinery; 2007 May Presented at: Proceedings of the 16th International Conference on World Wide Web; May 8-12, 2007; Banff, AB. [doi: 10.1145/1242572.1242669]
43. Babalou S, Kargar M, Davarpanah S. Large-scale ontology matching: A review of the literature. 2016 Apr Presented at: Proceedings of the 2nd International Conference on Web Research; April 2016; Tehran, Iran p. 27-28. [doi: 10.1109/icwr.2016.7498461]
44. Abbes H, Gargouri F. MongoDB-Based Modular Ontology Building for Big Data Integration. J Data Semant 2017 Oct 27;7(1):1-27. [doi: 10.1007/s13740-017-0081-z]
45. Algergawy A, Babalou S, Klan F, König-Ries B. Ontology Modularization with OAPT. J Data Semant 2020 Aug 24;9(2-3):53-83. [doi: 10.1007/s13740-020-00114-7]
46. Paulheim H. On applying matching tools to large-scale ontologies. 2008 Oct 26 Presented at: Proceedings of the 3rd International Workshop on Ontology Matching collocated with the 7th International Semantic Web Conference; October 26, 2008; Karlsruhe, Germany.
47. Stuckenschmidt H, Klein M. Structure-based partitioning of large concept hierarchies. Heidelberg, Germany: Springer; 2004 Nov Presented at: Proceedings of the 3rd International Semantic Web Conference; November 2004; Hiroshima, Japan. [doi: 10.1007/978-3-540-30475-3_21]
48. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biol 2012 Jan 31;13(1):R5 [FREE Full text] [doi: 10.1186/gb-2012-13-1-r5] [Medline: 22293552]
49. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv. 2018. URL: https://arxiv.org/abs/1810.04805 [accessed 2020-07-22]
50. 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]
51. Sung M, Jeon H, Lee J, Kang J. Biomedical entity representations with synonym marginalization. arXiv. 2020. URL: http://arxiv.org/abs/2005.00239 [accessed 2021-07-22]
52. Wang P, Xu B, Zhou Y. Extracting semantic subgraphs to capture the real meanings of ontology elements. Tinshhua Sci. Technol 2010 Dec;15(6):724-733. [doi: 10.1016/s1007-0214(10)70121-8]
53. Hayamizu T, Mangan M, Corradi J, Kadin JA, Ringwald M. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. Genome Biol 2005;6(3):R29 [FREE Full text] [doi: 10.1186/gb-2005-6-3-r29] [Medline: 15774030]
54. Hertling S, Paulheim H. WikiMatch: Using Wikipedia for ontology matching. 2012 Nov 11 Presented at: Proceedings of the 7th International Workshop on Ontology Matching collocated with the 11th International Semantic Web Conference; November 11, 2012; Boston, MA. [doi: 10.1007/978-3-642-38288-8_3]
55. Portisch J, Paulheim H. ALOD2Vec matcher. 2018 Oct 08 Presented at: Proceedings of the 13th International Workshop on Ontology Matching collocated with the 17th International Semantic Web Conference; October 8, 2018; Monterey, CA.

XSL•FO
RenderX

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BIC: Bayesian information criterion
FMA: Foundation Model of Anatomy
MELT: Matching EvaLuation Toolkit
NCI: National Cancer Institute
OAEI: Ontology Alignment Evaluation Initiative
OWL: Ontology Web Language
RDF: Resource Description Framework
RDFS: Resource Description Framework Schema
TF-IDF: term frequency-inverse document frequency
URI: uniform resource identifier

Edited by T Hao; submitted 26.02.21; peer-reviewed by W Heng, O Bodenreider, J Li; comments to author 30.03.21; revised version received 23.04.21; accepted 19.05.21; published 19.08.21.

Please cite as:
Wang P, Hu Y, Bai S, Zou S
Matching Biomedical Ontologies: Construction of Matching Clues and Systematic Evaluation of Different Combinations of Matchers
JMIR Med Inform 2021;9(8):e28212
URL: https://medinform.jmir.org/2021/8/e28212
doi:10.2196/28212
PMID:34420930

©Peng Wang, Yunyan Hu, Shaochen Bai, Shiyi Zou. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

XSL•FO
RenderX