Contents

Reviews

Ontologies Applied in Clinical Decision Support System Rules: Systematic Review (e43053)
Xia Jing, Hua Min, Yang Gong, Paul Blondich, David Robinson, Timothy Law, Christian Nohr, Arild Faxvaag, Lior Rennert, Nina Hubig, Ronald Gimbel.......................................................... 6

The Current Status of Secondary Use of Claims, Electronic Medical Records, and Electronic Health Records in Epidemiology in Japan: Narrative Literature Review (e39876)
Yang Zhao, Tadashi Tsubota ................................................................................................................ 21

Smart Glasses for Supporting Distributed Care Work: Systematic Review (e44161)
Zhan Zhang, Enze Bai, Karen Joy, Parth Ghelaa, Kathleen Adelgais, Mustafa Ozkaynak ......................... 36

Mining Sensor Data to Assess Changes in Physical Activity Behaviors in Health Interventions: Systematic Review (e41153)
Claudio Diaz, Corinne Caillaud, Kalina Yacef ........................................................................................ 52

Methods Used in the Development of Common Data Models for Health Data: Scoping Review (e45116)
Najia Ahmadi, Michele Zoch, Patricia Kelbert, Richard Noll, Jannik Schaaf, Markus Wolfien, Martin Sedlmayr .......................................................... 69

Designing Interoperable Health Care Services Based on Fast Healthcare Interoperability Resources: Literature Review (e44842)
Jingwen Nan, Li-Qun Xu ...................................................................................................................... 86

Scalable Causal Structure Learning: Scoping Review of Traditional and Deep Learning Algorithms and New Opportunities in Biomedicine (e38266)
Pulakesh Upadhayaya, Kai Zhang, Can Li, Xiaojian Jiang, Yejin Kim ........................................................................ 147

Systematized Nomenclature of Medicine—Clinical Terminology (SNOMED CT) Clinical Use Cases in the Context of Electronic Health Record Systems: Systematic Literature Review (e43750)
Riikka Vuokko, Anne Vakkuri, Sari Palojoki ...................................................................................... 433

Viewpoints

Developing a Capsule Clinic—A 24-Hour Institution for Improving Primary Health Care Accessibility: Evidence From China (e41212)
Dongliang Li, Rujia Zhang, Chun Chen, Yuyun Huang, Xiaoyi Wang, Qingren Yang, Xuebo Zhu, Xiangyang Zhang, Mo Hao, Liming Shui . .
One Digital Health Intervention for Monitoring Human and Animal Welfare in Smart Cities: Viewpoint and Use Case (e43871)
Arriel Benis, Mostafa Haghi, Thomas Deserno, Oscar Tamburis.

Practical Considerations for Developing Clinical Natural Language Processing Systems for Population Health Management and Measurement (e37805)
Suzanne Tamang, Marie Humbert-Droz, Milena Gianfrancesco, Zara Izadi, Gabriela Schmajuk, Jinoos Yazdany.

Original Papers

Monitoring the Implementation of Tobacco Cessation Support Tools: Using Novel Electronic Health Record Activity Metrics (e43097)
Jinying Chen, Sarah Cutrona, Ajay Dharod, Stephanie Bunch, Kristie Foley, Brian Ostasiewski, Erica Hale, Aaron Bridges, Adam Moses, Eric Donny, Erin Sutfin, Thomas Houston, iDAPT Implementation Science Center for Cancer Control.

Dealing With Missing, Imbalanced, and Sparse Features During the Development of a Prediction Model for Sudden Death Using Emergency Medicine Data: Machine Learning Approach (e38590)
Xiaojie Chen, Han Chen, Shan Nan, Xiangtian Kong, Huilong Duan, Haiyan Zhu.

Barriers and Opportunities for the Use of Digital Tools in Medicines Optimization Across the Interfaces of Care: Stakeholder Interviews in the United Kingdom (e42458)
Clare Tolley, Helen Seymour, Neil Watson, Hamde Nazar, Jude Heed, Dave Belshaw.

Using the H2O Automatic Machine Learning Algorithms to Identify Predictors of Web-Based Medical Record Nonuse Among Patients in a Data-Rich Environment: Mixed Methods Study (e41576)
Yang Chen, Xuejiao Liu, Lei Gao, Miao Zhu, Ben-Chang Shia, Mingchih Chen, Linglong Ye, Lei Qin.

Near Real-time Natural Language Processing for the Extraction of Abdominal Aortic Aneurysm Diagnoses From Radiology Reports: Algorithm Development and Validation Study (e40964)

Deep Learning Approach for Negation and Speculation Detection for Automated Important Finding Flagging and Extraction in Radiology Report: Internal Validation and Technique Comparison Study (e46348)
Kung-Hsun Weng, Chung-Feng Liu, Chia-Jung Chen.

Understanding Views Around the Creation of a Consented, Donated Databank of Clinical Free Text to Develop and Train Natural Language Processing Models for Research: Focus Group Interviews With Stakeholders (e45534)

Acquisition of a Lexicon for Family History Information: Bidirectional Encoder Representations From Transformers—Assisted Sublanguage Analysis (e48072)
Liwei Wang, Huan He, Andrew Wen, Sungrim Moon, Sunyang Fu, Kevin Peterson, Xuguang Ai, Sijia Liu, Ramakanth Kavuluru, Hongfang Liu.

Agreement Between Experts and an Untrained Crowd for Identifying Dermoscopic Features Using a Gamified App: Reader Feasibility Study (e38412)
Toward Individualized Prediction of Binge-Eating Episodes Based on Ecological Momentary Assessment Data: Item Development and Pilot Study in Patients With Bulimia Nervosa and Binge-Eating Disorder (e41513)
Ann-Kathrin Arend, Tim Kaiser, Björn Pannicke, Julia Reichenberger, Silke Naab, Ulrich Voderholzer, Jens Blechert ................................................................. 302

Deployment of Real-time Natural Language Processing and Deep Learning Clinical Decision Support in the Electronic Health Record: Pipeline Implementation for an Opioid Misuse Screener in Hospitalized Adults (e44977)
Majid Afshar, Sabrina Adelaine, Felice Resnik, Marlon Mundt, John Long, Margaret Leaf, Theodore Ampian, Graham Wills, Benjamin Schnapp, Michael Chao, Randy Brown, Cara Joyce, Brihat Sharma, Dmitriy Dilgach, Elizabeth Burnside, Jane Mahoney, Matthew Churpek, Brian Patterson, Frank Liao ................................................................................................................................. 315

A Comprehensive and Improved Definition for Hospital-Acquired Pressure Injury Classification Based on Electronic Health Records: Comparative Study (e40672)
Mani Sotoodeh, Wenhui Zhang, Roy Simpson, Vicki Hertzberg, Joyce Ho .......................................................... 329

Identification of Postpartum Depression in Electronic Health Records: Validation in a Large Integrated Health Care System (e43005)
Jeff Slezak, David Sacks, Vicki Chiu, Chantal Avila, Neha Khadka, Jiu-Chuan Chen, Jun Wu, Darios Getahun ................................................................. 341

Perspectives on Challenges and Opportunities for Interoperability: Findings From Key Informant Interviews With Stakeholders in Ohio (e43848)
Daniel Walker, Willi Tarver, Pallavi Jonnalagadda, Lorin Ranbom, Eric Ford, Saurabh Rahurkar ................................................................. 350

The Effect of Implementation of Guideline Order Bundles Into a General Admission Order Set on Clinical Practice Guideline Adoption: Quasi-Experimental Study (e42736)
Justine Mrosak, Swaminathan Kandaswamy, Claire Stokes, David Roth, Jenna Gorbaktn, Ishaan Dave, Scott Gillespie, Evan Orenstein ............................................. 364

Data Analysis of Physician Competence Research Trend: Social Network Analysis and Topic Modeling Approach (e47934)
So Yune, Youngjon Kim, Jea Lee ......................................................................................................................................................... 389

Visual Analytics of Multidimensional Oral Health Surveys: Data Mining Study (e46275)
Ting Xu, Yuming Ma, Tianya Pan, Yifei Chen, Yuhua Liu, Fudong Zhu, Zhiguang Zhou, Qianming Chen ................................................................. 406

A Linked Open Data–Based Terminology to Describe Libre/Free and Open-source Software: Incremental Development Study (e38861)
Franziska Jahn, Elske Ammenwerth, Verena Dornauer, Konrad Höfner, Michelle Bindel, Thomas Karopka, Alfred Winter ................................................................. 420

Structure of Health Information With Different Information Models: Evaluation Study With Competency Questions (e46477)
Anna Rossander, Daniel Karlsson ......................................................................................................................................................... 442

The Impact of an Electronic Portal on Patient Encounters in Primary Care: Interrupted Time-Series Analysis (e43567)
Karen Ferguson, Mark Fraser, Meltem Tuna, Charles Bruntz, Simone Dahrouge ................................................................. 458

An Electronic Dashboard to Improve Dosing of Hydroxychloroquine Within the Veterans Health Care System: Time Series Analysis (e44455)
Anna Montgomery, Gary Tarasovsky, Zara Izadi, Stephen Shiboski, Mary Whooley, Jo Dana, Iziegbe Ehiorobo, Jennifer Barton, Lori Bennett, Lorinda Chung, Kimberly Reiter, Elizabeth Wahl, Meera Subash, Gabriela Schmajuk ......................................................................................................................................................... 471

Integrated Personal Health Record in Indonesia: Design Science Research Study (e44784)
Nabilah Harahap, Putu Handayani, Achmad Hidayanto ......................................................................................................................................................... 482
Using a Clinical Data Warehouse to Calculate and Present Key Metrics for the Radiology Department: Implementation and Performance Evaluation (e41808)
Leon Liman, Bernd May, Georg Fette, Jonathan Krebs, Frank Puppe. .......................................................... 507

Assessment and Improvement of Drug Data Structuredness From Electronic Health Records: Algorithm Development and Validation (e40312)
Ines Reinecke, Joscha Siebel, Saskia Fuhrmann, Andreas Fischer, Martin Sedlmayr, Jens Weidner, Franziska Bathelt. .......................................................... 526

An Ontology-Based Approach for Consolidating Patient Data Standardized With European Norm/International Organization for Standardization 13606 (EN/ISO 13606) Into Joint Observational Medical Outcomes Partnership (OMOP) Repositories: Description of a Methodology (e44547)
Santiago Frid, Xavier Pastor Duran, Guillen Bracons Cucó, Miguel Pedrera-Jiménez, Pablo Serrano-Balazote, Adolfo Muñoz Carrero, Raimundo Lozano-Rubi. ................................. 542

Data-Driven Identification of Unusual Prescribing Behavior: Analysis and Use of an Interactive Data Tool Using 6 Months of Primary Care Data From 6500 Practices in England (e44237)
Lisa Hopcroft, Jon Massey, Helen Curtis, Brian Mackenna, Richard Croker, Andrew Brown, Thomas O'Dwyer, Orla Macdonald, David Evans, Peter Inglesby, Sebastian Bacon, Ben Goldacre, Alex Walker. .......................................................... 554

Chinese Clinical Named Entity Recognition From Electronic Medical Records Based on Multisemantic Features by Using Robustly Optimized Bidirectional Encoder Representation From Transformers Pretraining Approach Whole Word Masking and Convolutional Neural Networks: Model Development and Validation (e44597)
Weijie Wang, Xiaoying Li, Huiling Ren, Dongping Gao, An Fang. .......................................................... 565

Improving an Electronic Health Record–Based Clinical Prediction Model Under Label Deficiency: Network-Based Generative Adversarial Semisupervised Approach (e47862)
Runze Li, Yu Tian, Zhuyi Shen, Jin Li, Jun Li, Keleng Ding, Jingsong Li. .......................................................... 586

Unique Device Identification–Based Linkage of Hierarchically Accessible Data Domains in Prospective Surgical Hospital Data Ecosystems: User-Centered Design Approach (e41614)
Karol Kozak, André Seidel, Natalia Matvieieva, Constanze Neupetsch, Uwe Teicher, Gordon Lemme, Anas Ben Achour, Martin Barth, Steffen Ihlenfeldt, Well-Guntram Drossel. .......................................................... 605

A Standardized Clinical Data Harmonization Pipeline for Scalable AI Application Deployment (FHIR-DHP): Validation and Usability Study (e43847)
Elena Williams, Manuel Kienast, Evelyn Medawar, Janis Reinelt, Alberto Merola, Sophie Klopfenstein, Anne Flint, Patrick Heeren, Akira-Sebastian Poncette, Felix Balzer, Julian Beimes, Paul von Bünau, Jonas Chromik, Bert Amrigh, Nico Scherf, Sebastian Niehaus. .......................................................... 637

A SNOMED CT Mapping Guideline for the Local Terms Used to Document Clinical Findings and Procedures in Electronic Medical Records in South Korea: Methodological Study (e46127)
Sumi Sung, Hyeoun-Ae Park, Hyesil Jung, Hannah Kang. .......................................................... 648

An End-to-End Natural Language Processing Application for Prediction of Medical Case Coding Complexity: Algorithm Development and Validation (e38150)
He Xu, Bernard Maccari, Hervé Guillain, Julien Herzen, Fabio Agri, Jean Raisaro. .......................................................... 666

Corrigenda and Addenda

Correction: Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study (e45742)
Lama Alsudias, Paul Rayson. .......................................................... 600
Editorial

“To Err Is Evolution”: We Need the Implementation Report to Learn (e47695)

Caroline Perrin Franck, Antoine Geissbuhler, Christian Lovis.
Review

Ontologies Applied in Clinical Decision Support System Rules: Systematic Review

Xia Jing¹, MD, PhD; Hua Min², MD, PhD; Yang Gong³, MD, PhD; Paul Biondich⁴, MD; David Robinson⁵, MD; Timothy Law⁶, DO; Christian Nohr⁷, MD; Arild Faxvaag⁸, MD, PhD; Lior Rennert¹, PhD; Nina Hubig⁹, PhD; Ronald Gimbel¹, PhD

¹Department of Public Health Sciences, Clemson University, Clemson, SC, United States
²College of Public Health, George Mason University, Fairfax, VA, United States
³School of Biomedical Informatics, The University of Texas Health Sciences Center at Houston, Houston, TX, United States
⁴Clem McDonald Biomedical Informatics Center, Regenstrief Institute, Indianapolis, IN, United States
⁵Loweswater Consulting, Combria, United Kingdom
⁶Ohio Musculoskeletal and Neurologic Institute, Ohio University, Athens, OH, United States
⁷Department of Planning, Aalborg University, Aalborg, Denmark
⁸Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway
⁹School of Computing, Clemson University, Clemson, SC, United States

*these authors contributed equally

Corresponding Author:
Xia Jing, MD, PhD
Department of Public Health Sciences
Clemson University
519 Edwards Hall
Clemson, SC, 29634
United States
Phone: 1 8646563347
Email: xjing@clemson.edu

Abstract

Background: Clinical decision support systems (CDSSs) are important for the quality and safety of health care delivery. Although CDSS rules guide CDSS behavior, they are not routinely shared and reused.

Objective: Ontologies have the potential to promote the reuse of CDSS rules. Therefore, we systematically screened the literature to elaborate on the current status of ontologies applied in CDSS rules, such as rule management, which uses captured CDSS rule usage data and user feedback data to tailor CDSS services to be more accurate, and maintenance, which updates CDSS rules. Through this systematic literature review, we aim to identify the frontiers of ontologies used in CDSS rules.

Methods: The literature search was focused on the intersection of ontologies; clinical decision support; and rules in PubMed, the Association for Computing Machinery (ACM) Digital Library, and the Nursing & Allied Health Database. Grounded theory and PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines were followed. One author initiated the screening and literature review, while 2 authors validated the processes and results independently. The inclusion and exclusion criteria were developed and refined iteratively.

Results: CDSSs were primarily used to manage chronic conditions, alerts for medication prescriptions, reminders for immunizations and preventive services, diagnoses, and treatment recommendations among 81 included publications. The CDSS rules were presented in Semantic Web Rule Language, Jess, or Jena formats. Despite the fact that ontologies have been used to provide medical knowledge, CDSS rules, and terminologies, they have not been used in CDSS rule management or to facilitate the reuse of CDSS rules.

Conclusions: Ontologies have been used to organize and represent medical knowledge, controlled vocabularies, and the content of CDSS rules. So far, there has been little reuse of CDSS rules. More work is needed to improve the reusability and interoperability of CDSS rules. This review identified and described the ontologies that, despite their limitations, enable Semantic Web technologies and their applications in CDSS rules.
### Introduction

For more than half a century, clinical decision support systems (CDSSs) have been developed and used in clinical care delivery [1-5]. Some early CDSS examples include Dialog [6], INTERNIST-1 [7-9], Quick Medical Reference [8], and Iliad [10-12]. The effectiveness of CDSSs in clinical care has been established [13-15], with some pioneering researchers’ work on CDSS effectiveness particularly noteworthy [16]. Researchers have examined CDSS users’ and developers’ experiences, discussed their CDSS vision for the future [17], and recommended best practice guidelines in CDSSs [18-22]. Meanwhile, the challenges of CDSSs have been well documented [23]. Meeting clinician information needs is one way a CDSS can help health care providers improve clinical care quality. Many studies, such as Infobutton [13,24], have demonstrated the effectiveness of CDSSs in this aspect. CDSSs are currently routinely used in clinical care, with rates ranging from 68.5% to 100% in primary care settings based in offices [25] in the United States as part of electronic health record (EHR) systems. CDSSs can take many forms, including but not limited to reminders for preventive services (eg, immunizations and screening tests) [26-28], alerts for drug-drug interactions [22,29,30], diagnostic or treatment plan recommendations [31-33], clinician content assistance [34-38], and recommendations for adhering to current clinical practice guidelines [39-41]. CDSSs have played an important role and are widely used in practice to provide safer and better clinical care services.

CDSS rules, which function similarly to the human central nervous system, direct the behaviors of a CDSS during operations by incorporating patient data, contextual information, and medical domain knowledge. The central role of CDSS rules is a decisive factor in the relevance and usefulness of a CDSS in the overall clinical workflow, which impacts whether a CDSS is adopted and routinely used. CDSS rules can be written in Arden syntax [42], Semantic Web Rule Language (SWRL), Jess, Jena, and other programming languages, and the processes are labor intensive. Only specially trained personnel are qualified to write such rules. Moreover, regular updating of CDSS rules is required to keep CDSSs relevant and useful in clinical care delivery. However, the process of developing, updating, and maintaining CDSS rules is time-consuming and resource intensive [4,43], making it difficult for both large institutions and resource-constrained small-scale practices. CDSS rule usage data, such as rule fire rates, overwrite rates, successful rates, and user feedback data, can be collected to improve and customize CDSSs and manage CDSS rules. Typically, CDSS rule maintenance entails adding, deleting, and updating CDSS rules.

Ontologies have been successfully applied to generate and supply domain knowledge in the use, reuse, sharing, and interoperability of information. Ontologies are seen as promising solutions to the challenges of managing and maintaining CDSS rules across institutional boundaries. The Semantic Web is a technology enabled by ontology [44] that is critical in information sharing and reuse [45,46], medicine [47], and CDSSs [48,49]. Although there are numerous definitions of ontology, we used Gruber’s definition in this manuscript: “an ontology is a specification of conceptualization” [45]. Interoperability has been identified as a major challenge for health care information technologies, particularly when it comes to sharing health information across institutional or national boundaries. Ontologies have the potential to shorten the interoperability gap.

Reusing and sharing CDSS rules are important, but they are not yet routine operations; thus, we conducted this systematic literature review. This study aims to expand on the current state of using ontologies in CDSS rules by conducting a systematic review of the literature on the intersection of CDSS rules, Semantic Web technologies (particularly ontologies), and use of ontologies in CDSSs. The review is expected to provide a comprehensive view of using ontologies in CDSS rules, with granular details. The results could serve as a basis to form a knowledge framework of the topic that may inspire future research. The research question we intend to answer with this systematic literature review is as follows: What is the current state of using semantic technologies, particularly ontologies, to leverage CDSS rule interoperability? Furthermore, the manually annotated results of selected publications could serve as gold standards for automatically identifying relevant entities in the literature.

### Methods

#### Databases and Search Strategies

Figure 1 illustrates the general workflow we used to conduct this literature review. An initial set of literature searches was conducted on June 2, 2020, which was followed by a review and discussions. The reviewers (XJ, HM, and YG) refined and agreed with the search strategies and searched PubMed, the Association for Computing Machinery (ACM) Digital Library, and the Nursing & Allied Health Database (NAHD) for literature, using the search strategies mentioned below. A final search was conducted on January 5, 2022, in the 3 literature databases as an update.

For PubMed, the following search was conducted: (clinical decision support systems[MeSH Terms] AND (ontolog*[Title/abstract] OR rule*[Title/abstract])). For ACM Digital Library, the following search was conducted within the scope of the ACM Guide to Computing Literature: [[Publication Title: “clinical decision support*”] OR [Publication Title: cds*]] AND [[Publication Title: ontolog*] OR [Abstract: ontolog*] OR [Publication Title: rule*] OR [Abstract: rule*]]. For the NAHD, the following search was limited to peer-reviewed publications: mesh(clinical decision support) AND (ti(ontology))
Inclusion and Exclusion Criteria

The inclusion criteria were as follows: text was written in English; full-text publication was available; ontologies were designed to be implemented or were already implemented in CDSSs, particularly related to CDSS rules; content included the granularity of CDSS rules; ontologies were designed to be integrated or were already integrated with health information systems (eg, EHRs), either in a production system or a prototype, with at least one architecture diagram, applied in clinical domains or designed for clinical domains to support health care providers; the publication was peer-reviewed; and details on the integration of CDSSs and EHRs were present for evaluation studies.

The exclusion criteria were as follows: only CDSS rules were included, regardless of the stage of the CDSS rule lifecycle (ie, development, identification, refinement, validation, evaluation, or implementation) or there was no mention of integration or ontologies; only ontologies were developed, evaluated, and validated, or there was no mention of integration or a CDSS; the system was designed without mentioning the granularity of CDSS rules or ontologies; and nonclinical decisions, such as administrative or management decisions (eg, supply chain management), were described.

General Workflow for Screening Papers

The first 100 papers were screened by all 3 authors (XJ, HM, and YG) independently. The first 100 retrieved papers were initially screened by 1 author (XJ) to draft initial inclusion and exclusion criteria. The inclusion and exclusion criteria were refined and adjusted by 2 authors (HM and YG) during the iterative screening, review, analysis, and discussions. Further, 2 authors (HM and YG) replicated the screening, and all 3 authors discussed and validated the results. The rest of the papers were then screened by at least 2 authors (XJ and HM, or XJ and YG) independently to determine inclusion. Disagreements were discussed and resolved via iterative rounds of group meetings.

The screening and manual review processes were conducted independently and approved by at least 2 authors. The literature was first screened based on titles, abstracts, and full-text publications when needed. The papers that were included were then manually coded to provide more content analysis and synthesized evidence. The final results were shared among all the authors. All disagreements were settled through group discussions.

Reviewing, Coding, Analyzing, and Synthesizing Processes

We followed grounded theory during the reviewing and manual coding of the included publications. One author (XJ) randomly selected 10 papers from the included 81 papers to start the coding (annotating) based on the focus of this literature review. ATLAS.ti 9 (desktop and web versions; ATLAS.ti Scientific Software Development GmbH), a qualitative data analytic tool, was used for coding. The coding results were discussed by 3 authors (XJ, HM, and YG). The discussion results formed the first draft of codes and code groups (Multimedia Appendix 1), that is, data items. Three coders (XJ, HM, and YG) then reviewed and coded the first 40 of the included papers using the initial principles and code groups, and added new codes and code groups when needed. Then, a second set of meetings was used to obtain consensus on updated principles and code groups. Refined codes and code groups were used to code the remaining papers. Every paper was coded by at least 2 coders independently. The coding results were then compared, and any discrepancies were resolved by group discussions. The code groups and codes were revised, consolidated, and updated during each discussion. Multimedia Appendix 2 presents the refined code groups and examples. Data items emerged during the review and were refined via discussions instead of predefinition before reviewing. Multimedia Appendix 3 lists all included papers.

After coding, the literature was analyzed and synthesized with a focus on several aspects, including CDSS application domains, CDSS mechanisms used in clinical settings, CDSS rule formats, authoring, management, and the roles of ontologies. The 3 authors worked together in an iterative process of analysis and synthetization. After obtaining consensus among all 3 authors, the results were then shared and discussed among all authors. Any concerns, confusions, or disagreements among the authors were resolved through iterative discussions. We followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist [50] for reporting the systematic review with all relevant items (Multimedia Appendix 4 and Multimedia Appendix 5).

Results

Overview

By January 5, 2022, literature searches retrieved 1235 publications from 3 sources. After removing duplicates and examining according to the inclusion and exclusion criteria, 81
publications (Multimedia Appendix 3) were included in the final review and analysis [26,27,29,31-33,51-125]. Figure 2 depicts the literature search, screening, selection flow, and results. Figure 3 summarizes the main components covered by the literature review and the summary findings, and serves as an initial knowledge framework on CDSSs, CDSS rules, and ontology applications in CDSSs.

**Figure 2.** Flowchart of the literature search, screening, and selection. ACM: Association for Computing Machinery.

The majority of the publications (73/81, 90%) included in the review were from PubMed, a dominant source. After removing duplicates, the ACM Library added 8 new publications. After cleaning, discussion, and consolidation, 30 code groups and...
221 final codes were used in ATLAS.ti (Multimedia Appendix 2). These codes and code groups guided our analysis and synthesis of the results. Multimedia Appendix 6 shows a word cloud image generated by ATLAS.ti that reflects the codes coded in the publications included.

PRISMA 2020 is designed to guide the reporting of outcome-oriented studies. Our systematic literature review focused on the design, development, and implementation of CDSSs, particularly related to CDSS rules and ontologies. Therefore, effect measures or certainty assessments were irrelevant items. We reported 19 categorical items (out of 27 categorical items, 26 items out of 42 items; Multimedia Appendix 4) for the full-text papers and 10 for the abstracts (out of 12 items; Multimedia Appendix 5).

Results Related to CDSS Characteristics

Over one-third (29/81, 36%) of CDSSs were designed and used for chronic condition management, prediction, or risk assessment, including but not limited to type 1 and 2 diabetes, hypertension, and asthma. Medication prescriptions (13/81, 16%), such as medication ordering, detection of adverse drug events, drug-drug interactions, and cancer care (8/81, 10%), were also significant application domains. Multimedia Appendix 7 illustrates the clinical domains of CDSSs within the included publications. Most CDSSs were designed for health care providers, but only 11% (9/81) were intended for patients. Most CDSSs provided recommendations, suggestions, alerts, or reminders. Among all the items in our comparison (Multimedia Appendix 8), EHR evaluation studies within the operational systems or prototypes exhibited the least complete information. Evaluations of CDSSs have been listed in multiple columns in Multimedia Appendix 8. Some CDSSs were implemented in production systems (31/81, 38%), whereas others were implemented in prototypes (30/81, 37%), which included experimental systems. Multimedia Appendix 8 summarizes the key features of CDSSs identified in the publications. In all tables, we adopted the original terms used in the corresponding papers. Some papers, for example, referred to “physicians” as CDSS users, whereas others referred to “clinicians” as CDSS users.

Results Related to CDSS Rules

Most CDSS rules were written in Web Ontology Language (OWL; 11/81, 14%), Extensive Markup Language (XML; 10/81, 12%), SWRL (9/81, 11%), Jena rules (5/81, 6%), and medical logic module (MLM; 3/81, 4%). Moreover, 2 publications [117,119] used N3 Language and 2 [90,117] used Natural Rule Language (NRL). Multimedia Appendix 9 presents 54 publications with more details on the CDSS rules, that is, publications that can fill out 3 or more cells (except for authors and publication year).

The most significant CDSS rule source is from clinical practice guidelines (36/81, 44%). Other sources of CDSS rules included domain expert input, publications (eg, textbooks and papers), multimedia sources, and internet resources. Data mining results were involved in CDSS rule sources [67,73]. CDSS rule authoring and editing tools were not routinely specified in the publications. Protégé [115] was the most prevalent tool to edit and author CDSS rules. Several publications also described developing authoring and editing tools [57,65,91].

There was a lack of technical details regarding rule engines, among which Jena (6/81, 7%), inference engine (6/81, 7%), Jess (4/81, 5%), JBob (3/81, 4%), guideline engine (3/81, 4%), Drools (2/81, 3%), and Bayes (2/81, 3%) were frequently mentioned. Multimedia Appendix 9 summarizes how the CDSS rule (operation) works in a simplified manner. Many publications did not specify the working mechanism of CDSS rules within the EHR, electronic medical record (EMR), or hospital information system (HIS) context.

The majority of the publications did not appear to be focused on interoperability. Few papers that discussed interoperability (Multimedia Appendix 9) used HL7 CDA (Health Level 7, Clinical Document Architecture) or HL7 FHIR (Health Level 7, Fast Healthcare Interoperability Resources) standards. However, it is worth noting that such HL7 measures were not specifically designed for CDSS rules but rather for CDSS input and output.

Furthermore, some publications lacked necessary information for explaining the mechanisms of the systems, which can be critical barriers to reproducibility. Some publications lacked critical information, such as CDSS architecture diagrams; CDSS rule engines; CDSS rule languages; backend management methods for CDSS rules; and integration mechanisms among CDSS rules, ontologies, and EHR, EMR, or HIS systems.

Results Related to Ontologies

In the included publications, ontologies were primarily used as knowledge sources for CDSSs (32/81, 40%) to facilitate classification (7/81, 9%), reasoning, and inference (6/81, 7%; eg, identification recommendations or relationships). Moreover, ontologies were used to specify CDSS rules (12/81, 15%) or to provide general knowledge for the EMR or EHR systems. These 2 applications overlapped in some cases (19/81, 24%; ie, the ontologies were used to provide specified CDSS rules and general knowledge).

In the included publications, the terms “reasoner” and “rule engine” were used interchangeably. Reasoner, in our opinion, refers to the inference for a consistency check or classification for an ontology. A reasoner can be part of an ontology tool or can be external. For CDSSs, a rule engine is the mechanism that generates or provides recommendations by incorporating a patient’s data, contextual information, and medical knowledge (typically from an ontology or knowledge base). However, we kept the authors’ choice of terms in tables without modification. Among the included publications, the most common reasoners were Pellet (11/81, 14%), Jena (4/81, 5%), OWL reasoner (3/81, 4%), Jess (2/81, 3%), and the Euler/EYE inference engine (2/81, 3%).

The content and code systems used to represent the content should be included as ontology sources. The content could come from a popular textbook or a clinical practice guideline. The content can be coded in a specific code system, such as SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms). Multimedia Appendix 10 includes code systems that served as ontology sources. The most often
Discussion

Summary of the Results

Although ontologies contribute to the content of CDSS rules and have the potential to facilitate interoperable CDSS rules, our systematic review showed that reusing and sharing of CDSS rules have not been achieved. CDSSs have a wide range of clinical application domains, primarily for health care providers, such as chronic condition management, medication ordering, and cancer care. CDSS rules are primarily based on clinical practice guidelines.

Although reusing and sharing CDSS artifacts are well-recognized challenges [1,109], reusability, customization, and shareability of CDSS rules are not yet a common focus, even in publications focusing on CDSS rule editing [43,126,127]. These are important topics to cover in a literature review. Marco-Ruiz et al [109] demonstrated how to use CDSS artifacts in the Linked Data framework [128] by leveraging Semantic Web technologies, particularly ontology. However, that work was at a higher level, describing concepts without tangible tools implemented in clinical practice. To fill this gap, one approach is to build an upper-level CDSS ontology [129] to encourage the reuse of CDSS rules and demonstrate the potential of ontologies. Our effort is in alignment with their vision, as well as other efforts in reusing and sharing CDSS artifacts [1,109].

Ontologies were not at the center of any early examples of CDSSs [6-12]. An early demonstration of using medical terminology in CDSSs was the adoption of Current Medical Information and Terminology (CMIT) in a diagnostic engine [130,131]. Even under our “loose use of ontology” during our systematic literature search, there was no case in which ontology played a central role in sharing CDSS rules, particularly for rule management and maintenance.

Over the years, CDSSs have been successfully applied in clinical care. Unfortunately, CDSS rules are not yet portable. Making CDSS rules more portable is therefore significant work that could be leveraged by ontologies, and our systematic literature review brings us one step closer to that goal. Marco-Ruiz et al also conducted a very relevant systematic literature review. However, their focus was on the interoperability mechanisms used in CDSSs [132,133]. According to the results of their systematic literature review, 32% of the included papers used ontologies and 46% used standard terminologies. The findings related to ontologies are similar [132,133] to those of our paper. However, we presented a more detailed and thorough analysis of these technologies used in CDSS rules. Nevertheless, both papers concluded that complete CDSS interoperability is not a reality. Thus, additional efforts are required to achieve interoperable and reusable CDSS artifacts, such as CDSS rules.

Interpretation of the Results

Rule engines, which execute rules, patient data, and context information to produce a result, such as an alert or a recommendation, are critical components of CDSSs [1]. Jess, a rule engine and development environment in Java [134], was frequently mentioned in the included publications as a tool for developing rule-based CDSSs. SWRL rules can be converted to Jess rules in the popular tool Protégé, using a plug-in application programming interface (API) SWRLJessTab. Jess rules can be used by the Jess rule engine, which is widely used in rule-based expert systems [134]. In addition to Jess, Jena and Drools were used frequently in the publications included. Jena is a Java API that supports rule-based inference and makes use of resource description framework (RDF) graphs [135]. Jena.java API is a popular framework for managing RDF/OWL descriptions and can handle OWL models [96]. Drools is a business rule management system that includes a rule engine [136]. Drools also has the SWRL API that supports SWRL and Semantic Query-Enhanced Web Rule Language (SQWRL). SWRL can be queried by SQWRL.

Reasoning via a reasoner is a critical characteristic of many ontologies, even though the current reasoning is still in first-degree logic. Reasoning can be used for the following 3 main functions: consistency check, classification, and realization [137]. Several publications specified the classification roles of the ontologies and reasoners (Multimedia Appendix 10). The Manchester University OWL group has curated an updated list of OWL reasoners [137]. Parsia et al [137] compiled and compared the current OWL reasoners and their performances via the competition report. Both Pellet and Jena are popular reasoners (Multimedia Appendix 10), and other reasoners include FaCT++ [98], Z3 Solver reasoner [105], Euler/EYE inference engine [117,119], OWL Horst [109], and OWL Cerebra [63] among the included publications. Among these reasoners, Pellet [138] is Java based, and it can work on SWRL rules and ontologies written in OWL2. SWRL was initially designed as a rule language for Semantic Web technologies [139]. A user needs the rule language and an editor (eg, Protégé SWRL tab) to write, revise, and query the rules. SWRL can be queried by SQWRL (a query language for OWL) or SPARQL (SPARQL Protocol and RDF query language). Reasoners can then be used to conduct reasoning based on the rules and facts defined in the ontology or knowledge base. Protégé-OWL [140] provides an editor for SWRL rules. Protégé SWRL editor is another example.

This review has demonstrated unique insights about CDSS rules, ontologies, and ontology applications, particularly in CDSS rule management and maintenance, and has presented several distinct characteristics that complement the existing literature.
An earlier review [40] focused on clinical decision-making in forming ontologies to support complex cognitive processes and reasoning processes comparing evaluation metrics but did not cover the implementation of EHR, EMR, or HIS systems and the mechanisms of these characteristics.

Significance of the Work

Our systematic review demonstrated the state-of-the-art applications of ontologies in CDSS rules. These applications have a lot of potential for reusing and sharing CDSS artifacts. However, none of the existing papers elaborated or demonstrated how ontologies enable portable CDSS rules. Although some authors recognized this benefit [1,43,109], none have conducted a systematic review. Our literature review thoroughly examined the topic, outlined the current frontlines on CDSS rules and ontology uses in CDSSs, established the knowledge framework, and compiled a comprehensive collection of relevant publications that can inform future efforts to design or improve CDSSs. This systematic review focused on the mechanisms of CDSSs in clinical practices or prototypes, CDSS rules, and ontology roles in CDSSs. The detailed information provided in each included publication (Multimedia Appendix 8, Multimedia Appendix 9, and Multimedia Appendix 10) about the reasoners, rule engines, ontologies, and CDSS rule formats used provided valuable references for designing or improving systems. The side-by-side comparison of publications (Multimedia Appendix 8, Multimedia Appendix 9, and Multimedia Appendix 10) also provided structured guidance for preparing future designs and publications or teaching references on the topics in tangible ways.

Missing Information in the Publications and Our Recommendations

Inconsistent or missing information about CDSS rule languages, CDSS rule engines, and CDSS evaluation details was identified. In CDSS evaluation, there was commonly no information about how the evaluation was conducted or who performed the evaluation. There were also inconsistencies in technical details related to ontology purposes, reasoners, connection mechanisms, or communications between CDSSs and EHR, EMR, or HIS systems. Inconsistent or missing information hampered reproducibility and further improvement of published work. We are obviously not the only group that has identified missing critical information as a problem in technical papers on similar topics [141].

Another missing piece is the evaluation and validation of ontologies or knowledge bases. Only 25% of publications mentioned that domain experts conducted evaluation or validation. A formal assessment or validation is critical to ensure the validity of the results from automated processes for some ontologies (or knowledge bases) derived from other automatic methods (eg, machine learning algorithms). Testing has not been conducted consistently across the publications. Some ontologies were authored by domain experts, which provides greater validity than those involving nondomain experts while constructing ontologies.

Thus, it is recommended that authors include essential technical details in publications. These technical details include CDSS application domains, intended CDSS users, CDSS notification types, CDSS evaluations (what, how, and by whom), CDSS rule sources, CDSS rule languages, CDSS rule engines, CDSS operation mechanisms, ontology use purposes, ontology sources (both content and code systems), ontology validation, reasoners, and connection or communication mechanisms between CDSSs and EMR, EHR, or HIS systems. Authors are highly encouraged to include such details to help readers reference, compare, and increase the reproducibility of the reported work.

Limitations

Our review has limitations. Non-English publications or full-text unavailable publications were not included. Publications that focused only on CDSS rules [43,126,127] were also excluded. Moreover, publications without specifying an ontology component were excluded, although such publications had a similar focus to one aspect of our systematic review. We also noticed that most of the publications on CDSS rule authoring and managing tools were from Partners HealthCare/Harvard Medical School. The strengths of Partners HealthCare/Harvard Medical School were shown. On the other hand, a lack of broad adoption, implementation, or publication of such topics was shown.

When “CDSS” is not specified as a keyword, the search results may exclude publications. For example, our 2 previous papers [142,143] were not found via the search strategy because “CDSS” was not used as a keyword, although the content was undoubtedly within the scope of this review. This challenge is common to how our current literature databases are organized and how we conduct a literature search. Even with MeSH (Medical Subject Headings; the controlled vocabulary for PubMed), publications can still be missed without using commonly recognized keywords. This challenge could be minimized and mitigated by carefully developing an exhaustive list of keywords to maximize the possibilities found during a literature search in the future.

Conclusions

The reuse, management, and maintenance of CDSS rules are critical yet challenging for their clinical application. Although ontologies have been used to contribute to the content of CDSS rules, they have not been used to facilitate CDSS rule reuse and sharing. Building a CDSS ontology, which could be the first tangible step, requires bridging high-level visions and operational efforts. Semantic interoperability remains a major challenge that must be overcome to achieve reuse of CDSS artifacts, including CDSS rules. The realization of semantic interoperability will not only allow for the reuse of CDSS artifacts, which are resource intensive to develop and maintain, but also provide practical insights to achieve interoperable patient records. This has been a long-lost aspect, and health care providers will be able to access patients’ complete records to provide safer and higher quality care every time to every patient. We believe that making CDSS rules interoperable can provide insightful guidance for interoperable patient records.

Incomplete technical details on CDSS rules and ontologies presented in publications should be addressed in future publications by including more detailed information about
architectural diagrams; the mechanisms of connection among ontologies, CDSS rules, and EHR, EMR, or HIS systems; CDSS rule languages; reasoners; rule engines; the validation or authorization of ontologies and CDSS rules; the purposes of ontologies; ontology sources; and the management and maintenance of CDSS rules. Such information can help researchers to optimize design and development while also increasing reproducibility. Finally, the knowledge framework and the summarization of included publications are expected to guide future CDSS improvements and innovations, CDSS rules, and the integration and communication of CDSSs with EHR, EMR, or HIS systems.

Acknowledgments
The authors thank professors James J Cimino, Dean F Sittig, and Adam Wright for their valuable feedback and suggestions on the manuscript. This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (under award number R01GM138589 and partially under award number P20GM121342). The content is solely the authors’ responsibility and does not necessarily represent the official views of the National Institutes of Health.

Data Availability
A request for additional data in addition to the results and appendices of this manuscript can be made to the corresponding author, and the final decision on data release will be made on a case-by-case basis, as appropriate.

Authors' Contributions
XJ, HM, and YG designed the study, conducted the review and coding, drafted the first version, analyzed and interpreted the results, and revised the manuscript significantly. PB, DR, TL, CN, AF, LR, NH, and RG participated in the design of the study, analyzed and interpreted the results, and revised the manuscript significantly.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Initial draft of codes and code groups used in reviewing/coding. [PDF File (Adobe PDF File), 195 KB - medinform_v11i1e43053_app1.pdf]

Multimedia Appendix 2
Refined version of codes and code groups used in reviewing/coding. [PDF File (Adobe PDF File), 215 KB - medinform_v11i1e43053_app2.pdf]

Multimedia Appendix 3
List of papers included in this systematic literature review (n=81). [PDF File (Adobe PDF File), 222 KB - medinform_v11i1e43053_app3.pdf]

Multimedia Appendix 4
PRISMA 2020 item checklist reported in our systematic literature review. [PDF File (Adobe PDF File), 155 KB - medinform_v11i1e43053_app4.pdf]

Multimedia Appendix 5
PRISMA 2020 for Abstracts checklist reported in our systematic literature review. [PDF File (Adobe PDF File), 37 KB - medinform_v11i1e43053_app5.pdf]

Multimedia Appendix 6
Word cloud generated from ATLAS.ti based on codes in the included publications. [PNG File, 353 KB - medinform_v11i1e43053_app6.png]

Multimedia Appendix 7
Bar chart generated from ATLAS.ti showing the clinical domains of clinical decision support systems in included publications. [PNG File, 122 KB - medinform_v11i1e43053_app7.png]

Multimedia Appendix 8
Basic clinical decision support system profiles in included publications (n=81).

Multimedia Appendix 9
Comparison of clinical decision support system rule characteristics in included publications (n=54).

Multimedia Appendix 10
Comparison of ontology roles in included publications (n=36).

References


129. Jing X, Min H, Gong Y. A clinical decision support system (CDSS) ontology to facilitate portable vaccination CDSS rules: preliminary results. 2021 Presented at: AMIA 2021 Annual Symposium; October 30-November 3, 2021; San Diego, CA.


136. Drools. URL: https://www.drools.org/ [accessed 2022-12-21]


Abbreviations

ACM: Association for Computing Machinery
API: application programming interface
CDSS: clinical decision support system
EHR: electronic health record
EMR: electronic medical record
HIS: hospital information system
HL7: Health Level 7
NAHD: Nursing & Allied Health Database
OWL: Web Ontology Language
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RDF: resource description framework
SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms
SQWRL: Semantic Query-Enhanced Web Rule Language
SWRL: Semantic Web Rule Language
Review

The Current Status of Secondary Use of Claims, Electronic Medical Records, and Electronic Health Records in Epidemiology in Japan: Narrative Literature Review

Yang Zhao¹, PhD; Tadashi Tsubota¹, PhD
Audit & Assurance Deloitte Analytics R&D, Deloitte Touche Tohmatsu LLC, Tokyo, Japan

Corresponding Author:
Yang Zhao, PhD
Audit & Assurance Deloitte Analytics R&D
Deloitte Touche Tohmatsu LLC
Marunouchi Nijubashi Building
3-2-3 Marunouchi, Chiyoda-ku
Tokyo, 1008360
Japan
Phone: 81 80 9350 0848
Email: yang1.zhao@tohmatsu.co.jp

Abstract

Background: Real-world data, such as claims, electronic medical records (EMRs), and electronic health records (EHRs), are increasingly being used in clinical epidemiology. Understanding the current status of existing approaches can help in designing high-quality epidemiological studies.

Objective: We conducted a comprehensive narrative literature review to clarify the secondary use of claims, EMRs, and EHRs in clinical epidemiology in Japan.

Methods: We searched peer-reviewed publications in PubMed from January 1, 2006, to June 30, 2021 (the date of search), which met the following 3 inclusion criteria: involvement of claims, EMRs, EHRs, or medical receipt data; mention of Japan; and published from January 1, 2006, to June 30, 2021. Eligible articles that met any of the following 6 exclusion criteria were filtered: review articles; non-disease-related articles; articles in which the Japanese population is not the sample; articles without claims, EMRs, or EHRs; full text not available; and articles without statistical analysis. Investigations of the titles, abstracts, and full texts of eligible articles were conducted automatically or manually, from which 7 categories of key information were collected. The information included organization, study design, real-world data type, database, disease, outcome, and statistical method.

Results: A total of 620 eligible articles were identified for this narrative literature review. The results of the 7 categories suggested that most of the studies were conducted by academic institutes (n=429); the cohort study was the primary design that longitudinally measured outcomes of proper patients (n=533); 594 studies used claims data; the use of databases was concentrated in well-known commercial and public databases; infections (n=105), cardiovascular diseases (n=100), neoplasms (n=78), and nutritional and metabolic diseases (n=75) were the most studied diseases; most studies have focused on measuring treatment patterns (n=218), physiological or clinical characteristics (n=184), and mortality (n=137); and multivariate models were commonly used (n=414). Most (375/414, 90.6%) of these multivariate modeling studies were performed for confounder adjustment. Logistic regression was the first choice for assessing many of the outcomes, with the exception of hospitalization or hospital stay and resource use or costs, for both of which linear regression was commonly used.

Conclusions: This literature review provides a good understanding of the current status and trends in the use of claims, EMRs, and EHRs in clinical epidemiology in Japan. The results demonstrated appropriate statistical methods regarding different outcomes, Japan-specific trends of disease areas, and the lack of use of artificial intelligence techniques in existing studies. In the future, a more precise comparison of relevant domestic research with worldwide research will be conducted to clarify the Japan-specific status and challenges.

(JMIR Med Inform 2023;11:e39876) doi:10.2196/39876

KEYWORDS
claims; electronic medical records; EMRs; electronic health records; EHRs; epidemiology; narrative literature review
Introduction

Background
Medical claims data, electronic medical records (EMRs), and electronic health records (EHRs) are familiar sources of real-world data (RWD). They are often used secondarily to complement limitations in clinical trials. For example, they can characterize patient subgroups that are excluded from clinical trials by following eligibility criteria such as comorbidities or age. Findings obtained through long-term, naturalistic observations of a large and diverse patient population can be easily generalized to other populations. Other advantages are that these data have high external validity, a single data source can be used for different study purposes, and prospective data collection is not required.

Claims data are electronic records of transactions between patients and health care providers. They include information on bills (claims) submitted by providers (hospitals, clinics, and pharmacies) to third-party payers (health insurance associations). There are already some large-scale commercial and nonprofit claims databases available in Japan [1-6] that aggregate information from multiple health care providers for secondary use. Recently, the EMR and EHR data have become widely available. The EMR data are the details of the encounters with patients recorded by physicians through EMR systems. They contain rich clinical information such as laboratory test results, diagnostic images, pathology findings, and patient symptoms. As different facilities may use different EMR systems, domestic EMR data are currently available from ≥1 medical institution. The EHR data are electronic records of all health-related information of individual patients created and managed by clinical professionals, which can be shared and used among various medical facilities. Current EHR databases in Japan include both patient claims data and medical records.

In recent years, claims data, EMRs, and EHRs have been increasingly used in clinical epidemiology studies. Such studies include cost-effectiveness analysis of drugs (including disease burden and assessment of medical technology), risk factor analysis, investigation of the actual status of drugs (including preclinical feasibility valuation, marketability study, and detection of prescription patterns), and evaluation of drug efficacy in actual clinical practice. Because these data are not designed for research purposes, the secondary use requires an understanding of their limitations and the ability to generate clinical questions, epidemiological skills to construct a study design, and statistical skills to analyze retrospective observational data. Previous approaches have addressed the limitations and challenges of using these data [7-12]. Understanding their application status based on these advanced guidelines is essential. However, investigations of existing epidemiological studies based on these data are lacking.

Objective
We conducted a comprehensive narrative literature review to clarify the secondary use of claims, EMRs, and EHRs in clinical epidemiology in Japan. We focused on 7 categories of key information, including organization, study design, RWD type, database, disease, outcome, and statistical method. We expect that this review would help in the design of high-quality epidemiological studies.

Methods

Overview
This is a comprehensive narrative literature review that investigated the secondary use of claims data, EMRs, and EHRs in epidemiology in Japan. Referring to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [13] and procedures used in previous review studies [14-18], we conducted this review by searching for biomedical articles in PubMed.

Information Source
We searched peer-reviewed publications that satisfied the eligibility criteria for this narrative literature review in PubMed from January 1, 2006, to June 30, 2021 (the date of search).

Search Strategy
Keywords used to search PubMed consisted of “real world,” “database,” “claim,” “receipt,” “administrative,” “emr,” “ehr,” “japan,” “electronic medical record,” “electronic health record,” and Medical Subject Headings (MeSH) terms including, “Electronic Health Record,” “Administrative Claims, Healthcare,” “Insurance Claim Review/statistics and numerical data,” and “Japan/epidemiology.” We initially identified related articles by using various combinations of these keywords. The details of the search string are available in Multimedia Appendix 1.

Eligibility Criteria
On the basis of the search strategy, we identified articles whose titles and abstracts satisfied the following three inclusion criteria: (1) involvement of claims, EMRs, EHRs, or medical receipt data; (2) mention of Japan; and (3) published from January 1, 2006, to June 30, 2021. Eligible articles were then filtered out by satisfying any of the following six exclusion criteria: (1) review articles; (2) non-disease-related articles; (3) articles in which the Japanese population is not the sample; (4) articles without claims, EMRs, or EHRs; (5) unavailability of full-text articles; and (6) articles without statistical analysis.

Selection Process
The second author (TT) conducted the article search based on the search strategy. Both authors jointly reviewed all searched publications and performed 2 rounds of screening to identify target eligible articles. In the first round, we removed duplicates and articles that met any of the 6 exclusion criteria by screening the titles and abstracts. Review articles were automatically identified by a section classification model [19] trained on the PubMed 200k data set [20], which classified sentences in the abstracts into 5 sections (introduction, objective, method, result, and conclusion). On the basis of the hypothesis that review articles do not have sentences describing the results, we considered those without result sentences as review articles and removed them from the target articles. Artificially, we filtered out articles that met the exclusion criteria (2)-(5). In the second round of screening, the first author (YZ) reviewed the full text articles and made the final decision on eligibility.
of the remaining articles and removed those that did not include statistical analysis. The 2 authors double checked the results to ensure accuracy and finalized the eligible articles.

**Data Collection**

**Overview**

Investigations of the titles, abstracts, and full texts were conducted for eligible articles, from which 7 categories of key information were collected. The information included organization, study design, RWD type, database, disease, outcome, and statistical methods. Details regarding the classifications for each category are provided in Multimedia Appendix 2.

**Automated Data Extraction**

Four of these categories, including organization, study design, RWD type, and disease, were automatically extracted by keywords matching on the titles and abstracts. Two authors coded the data collection together.

On the basis of authors’ address information, organization was classified into 3 groups: “academic,” “nonacademic,” and “collaboration,” which denote that a study was conducted by academia, enterprises (including pharmaceutical companies, biotechnology companies, medical device companies, voluntary associations, and other health care–related companies), or collaboration of academia and nonacademic enterprises, respectively. Study design information was extracted by matching sentences in the abstracts to the categories listed in Multimedia Appendix 2, which consists of cohort studies, case-control studies, case-crossover studies, and cross-sectional studies. Similarly, RWD-type information was extracted by matching sentences in the abstract with 3 keywords, including claims, EMRs, and EHRs. Disease information was classified according to tree codes C01-C26 of MeSH terms [21]. For articles without the corresponding MeSH terms, disease information was collected from their titles using MetaMap [22] and pyMeSHSim [23].

**Manual Data Extraction**

Subsequently, the first author (YZ) conducted a full-text investigation to collect information on the database, outcome, and statistical method used in the target articles. The second author (TT) cross-checked the results of this data collection.

Database information was collected directly from the full texts. For those articles that did not use a specific database, we categorized them uniformly according to their data source as “other database” or “municipal claims database,” where “other database” indicates data from 1 or more medical facilities and “municipal claims database” indicates claims data provided by regional administrative agencies. Because there is no familiar way of categorizing outcomes for RWD studies, we defined 8 classifications of outcomes by referring to the article by Abaho et al [24]. The explanations for these classifications are detailed in Multimedia Appendix 2. We defined a hierarchical approach to collect information on statistical methods in the text. As shown in Figure 1, the method used in these articles was first categorized as multivariate modeling, simple statistical analysis, or descriptive analysis. Then, multivariate modeling was subdivided according to the purposes of confounding adjustment, clustered data modeling, factor exploration, or cost-effectiveness analysis, where confounding adjustment was further classified according to whether propensity score (PS) analysis was conducted.

**Analysis**

We performed a descriptive statistical analysis of the collected data by describing their counts and percentages. In addition, we calculated the percentages of outcomes and databases for each disease. The percentages of statistical methods used to assess different outcomes were also analyzed. All codes used for data
collection and descriptive analyses were performed using Python (version 3.8.8, 2021).

## Results

### Study Selection

A total of 620 eligible articles were identified for this narrative literature review. Figure 2 [13-18] illustrates the selection process and the results of each screening step. We also illustrate the publication years of these articles in Multimedia Appendix 3. The distribution indicated that 68.7% (426/620) of the articles were published after 2018, suggesting that the secondary use of the 3 RWD types in epidemiological research in Japan was prevalent in approximately the last 5 years.

Figure 2. Search and screening process [13-18]. EHR: electronic health record; EMR: electronic medical record.

### Summary of Findings

#### Overview

We summarize the counts and percentages of information in the 7 categories and illustrate the top-ranked items for each category in Tables 1 and 2. All results are detailed in Multimedia Appendix 4. It should be noted that for an article with multiple diseases, data types, study designs, databases, outcomes, or modeling purposes, it was double counted in each classification to which it belongs. Therefore, the total percentage of these categories may not be 100%. The following subsections present the results for each category.
### Table 1. Results of counts and percentages of the 7 categories (n=620).

<table>
<thead>
<tr>
<th>Category</th>
<th>Count, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organization</strong></td>
<td></td>
</tr>
<tr>
<td>Academic</td>
<td>429 (69.2)</td>
</tr>
<tr>
<td>Nonacademic</td>
<td>153 (24.7)</td>
</tr>
<tr>
<td>Collaboration</td>
<td>35 (5.6)</td>
</tr>
<tr>
<td><strong>Study design</strong></td>
<td></td>
</tr>
<tr>
<td>Cohort study</td>
<td>533 (86)</td>
</tr>
<tr>
<td>Case-control study</td>
<td>30 (4.8)</td>
</tr>
<tr>
<td>Case-crossover study</td>
<td>23 (3.7)</td>
</tr>
<tr>
<td>Cross-sectional study</td>
<td>6 (1)</td>
</tr>
<tr>
<td><strong>RWD(^a)</strong> type</td>
<td></td>
</tr>
<tr>
<td>Claim</td>
<td>594 (95.8)</td>
</tr>
<tr>
<td>EMR(^b)</td>
<td>30 (4.8)</td>
</tr>
<tr>
<td>EHR(^c)</td>
<td>4 (0.6)</td>
</tr>
<tr>
<td><strong>Database</strong></td>
<td></td>
</tr>
<tr>
<td>JMDC(^d)</td>
<td>181 (29.2)</td>
</tr>
<tr>
<td>DPC(^e) database (MHLW(^f))</td>
<td>141 (22.7)</td>
</tr>
<tr>
<td>MDV(^g)</td>
<td>103 (16.6)</td>
</tr>
<tr>
<td>NDB(^h)</td>
<td>65 (10.5)</td>
</tr>
<tr>
<td>Other databases</td>
<td>26 (4.2)</td>
</tr>
<tr>
<td>JROAD-DPC(^i)</td>
<td>17 (2.7)</td>
</tr>
<tr>
<td>Municipal claims database</td>
<td>12 (1.9)</td>
</tr>
<tr>
<td>QIP(^j)</td>
<td>10 (1.6)</td>
</tr>
<tr>
<td><strong>Disease</strong></td>
<td></td>
</tr>
<tr>
<td>Infections</td>
<td>105 (16.9)</td>
</tr>
<tr>
<td>Cardiovascular diseases</td>
<td>100 (16.1)</td>
</tr>
<tr>
<td>Neoplasms</td>
<td>78 (12.6)</td>
</tr>
<tr>
<td>Nutritional and metabolic diseases</td>
<td>75 (12.1)</td>
</tr>
<tr>
<td>Digestive system diseases</td>
<td>68 (11)</td>
</tr>
<tr>
<td>Pathological conditions, signs and symptoms</td>
<td>63 (10.2)</td>
</tr>
<tr>
<td>Nervous system diseases</td>
<td>62 (10)</td>
</tr>
<tr>
<td>Musculoskeletal diseases</td>
<td>42 (6.8)</td>
</tr>
<tr>
<td>Mental disorders</td>
<td>38 (6.1)</td>
</tr>
<tr>
<td>Wounds and injuries</td>
<td>33 (5.3)</td>
</tr>
<tr>
<td>Male urogenital diseases</td>
<td>30 (4.8)</td>
</tr>
<tr>
<td>Respiratory tract diseases</td>
<td>27 (4.4)</td>
</tr>
<tr>
<td>Hemic and lymphatic diseases</td>
<td>16 (2.6)</td>
</tr>
<tr>
<td>Eye diseases</td>
<td>14 (2.3)</td>
</tr>
<tr>
<td>Skin and connective tissue diseases</td>
<td>10 (1.6)</td>
</tr>
<tr>
<td><strong>Outcome</strong></td>
<td></td>
</tr>
<tr>
<td>Treatment patterns</td>
<td>218 (35.2)</td>
</tr>
</tbody>
</table>
### Table 2. Results of modeling purposes as defined in Figure 1 and specific models used in the 414 multivariate modeling studies.

<table>
<thead>
<tr>
<th>Category of multivariate modeling studies</th>
<th>Count (n=414), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Modeling purpose</strong></td>
<td></td>
</tr>
<tr>
<td>Confounding adjustment</td>
<td>375 (90.6)</td>
</tr>
<tr>
<td>Propensity score matching analysis</td>
<td>96 (23.2)</td>
</tr>
<tr>
<td>Covariate adjustment</td>
<td>279 (67.4)</td>
</tr>
<tr>
<td>Clustered data modeling</td>
<td>69 (16.7)</td>
</tr>
<tr>
<td>Factor exploration</td>
<td>68 (16.4)</td>
</tr>
<tr>
<td>Cost-effectiveness analysis</td>
<td>8 (1.9)</td>
</tr>
<tr>
<td><strong>Specific method</strong></td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>249 (60.1)</td>
</tr>
<tr>
<td>Cox proportional hazards regression</td>
<td>87 (21)</td>
</tr>
<tr>
<td>Linear regression</td>
<td>57 (13.8)</td>
</tr>
<tr>
<td>Poisson regression</td>
<td>23 (5.6)</td>
</tr>
<tr>
<td>GLM(^a)</td>
<td>18 (4.3)</td>
</tr>
</tbody>
</table>

\(^a\)GLM: generalized linear model.

**Organization**

In Table 1, the results of organization show that most (429/620, 69.2%) target articles were conducted by academics, whereas nonacademic firms preferred to collaborate with academic institutions (153/620, 24.7%).

**Study Design**

The results of study design show 86% (533/620) of the articles that performed cohort studies, whereas only a few (30/620, 4.8%) studies were case-control studies, cross-sectional studies (23/620, 3.7%), and case-crossover studies (6/620, 1%).
**RWD Type**

Most (594/620, 95.8%) studies used claims data. Only a small number (30/620, 4.8%) of studies used EMRs and (4/620, 0.6%) EHRs. According to the articles that used EMRs or EHRs, we found that these studies commonly collected EMRs or EHRs from private databases (1 or some specific hospitals), which did not have large patient populations.

**Database**

Table 1 shows the top-ranked databases (n≥10) used in the target articles. The Japan Medical Data Center Claims (JMDC) database, a well-known, large-scale commercial insurance-based claims database operated by JMDC Inc [3,4], was the most used database. JMDC was used in 29.2% (181/620) of the total articles. The second most used database is composed of claims data from diagnosis procedure combination (DPC) hospitals provided by the Ministry of Health, Labour and Welfare (MHLW) [25,26], which we called the DPC database (MHLW). A total of 22.7% (141/620) of articles used the DPC database (MHLW). Medical data vision (MDV) [5], another commercial hospital claims-based database, was used for 16.6% (103/620) of the total articles. Fourth in the ranking is the National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB) data, which was established by the MHLW in 2009, covering almost the whole population in Japan [1,2]. NDB was used in 10.5% (65/620) of the total articles.

**Disease**

According to the information on diseases in Table 1, we found that most studies have focused on infections (105/620, 16.9%), cardiovascular diseases (100/620, 16.1%), neoplasms (78/620, 12.6%), and nutritional and metabolic diseases (75/620, 12.1%). In addition, there were a number of studies on psychiatric disorders, indicated here as nervous system diseases (62/620, 10%) and mental disorders (38/620, 6.1%).

**Outcome**

The results of outcome show that treatment patterns (218/620, 35.2%), physiological or clinical outcomes (184/620, 29.7%), and mortality (137/620, 22.1%) were the most assessed outcomes. Comparatively, few (32/620, 5.2%) articles assessed guideline adherence. Only few studies measured quality indicators (5/620, 0.8%).

**Statistical Method**

Table 1 also suggests that most (414/620, 66.8%) studies were performed using multivariate modeling. In addition, we investigated the counts and percentages of modeling purposes (Figure 1) and specific models used in the 414 multivariate modeling studies in Table 2. The results show that most (375/414, 90.6%) of the multivariate modeling studies were performed for confounder adjustment. Some were conducted for clustered data modeling (69/414, 16.7%) and factor exploration (68/414, 16.4%). Two types of models were used for clustered data modeling: the generalized estimating equations (GEE) method and multilevel models. The GEE methods adjust for the clustering nature of the data and correctly estimate the SE of the estimated parameters. Multilevel models are often used with random effects to estimate the predictor effects for patients in specific clusters. Our results indicate a greater tendency to use multilevel regression (43/414, 10.4%) than GEE (26/414, 6.3%) in clustered data modeling studies. Only a few (8/414, 1.9%) studies analyzed cost-effectiveness. Regarding the specific models used in the multivariate modeling studies, logistic regression (249/414, 60.1%), Cox proportional hazards regression (87/414, 21%), and linear regression (57/414, 13.8%) were the most used.

**Diseases and Outcomes**

We investigated the percentage of each outcome measured for different diseases. As shown in Figure 3, most (10/14, 71%) studies on eye diseases have focused on assessing their treatment patterns. Similarly, a number of studies on mental disorders (21/38, 55%), musculoskeletal diseases (21/42, 50%), and respiratory tract diseases (11/27, 40%) have also focused on assessing treatment patterns. Among the studies on hemic and lymphatic diseases, mortality accounted for the highest percentage (10/16, 63%), whereas few studies assessed adverse events. Furthermore, mortality has not been assessed in studies of mental disorders, eye diseases, and skin and connective tissue diseases. In addition, no study has assessed hospitalization or hospital stay in musculoskeletal, eye, and skin and connective tissue diseases.
**Statistical Methods and Outcomes**

We also calculated the percentages of statistical methods used to assess different outcomes. **Figure 4A** shows the percentages of the 3 types of statistical analyses used for each outcome; **Figure 4B** shows the percentages of multivariate modeling studies for different purposes for assessing these outcomes, and **Figure 4C** shows the percentage of each detailed multivariate model used for these outcomes. Multivariate modeling was used most frequently to assess mortality (116/137, 84.7%). Although the treatment patterns were the most assessed by the target studies (n=218), not many of them used multivariate modeling (97/218, 44.5%). **Figure 4B** indicates that almost all outcomes were measured with confounding adjustments. As shown in **Figure 4C**, logistic regression was the first choice for assessing mortality (96/116, 82.8%), physiological or clinical outcomes (60/110, 54.5%), treatment patterns (56/97, 58%), and guideline adherence (17/19, 90%). The results also suggest the use of Cox proportional hazards regression to assess these outcomes. In contrast, linear regression was the most commonly used model for assessing hospitalization or hospital stay (31/74, 42%) and resource use or costs (28/66, 42%).
Discussion

Principal Findings

A comprehensive narrative literature review was conducted to understand the secondary use of nationwide claims data, EMRs data, and EHRs data in clinical epidemiology in Japan. On the basis of the search strategy and eligibility criteria, a total of 620 eligible articles were identified from PubMed between January 1, 2006, and June 30, 2021 (the date of search).

We quantified 7 categories of key information from these 620 eligible articles. The main findings were that (1) most of the research has been done by academic institutions, whereas nonacademic institutions tend to collaborate with academic institutions; (2) the cohort study was the major design that longitudinally measured outcomes of proper patients; (3) most studies used claims data; (4) the JMDC, DPC database (MHLW), MDV, and NDB were mostly used, whereas only a few studies used EMRs or EHRs from a single hospital or multiple hospitals, which do not have a large patient population; (5) the top rank of diseases studied in the current research were infections, cardiovascular diseases, neoplasms, and nutritional and metabolic diseases; (6) treatment patterns, physiological or clinical outcomes, and mortality were the most assessed in these articles; and (7) multivariate models were commonly used, during which logistic regression and linear regression were shown to be the first choice for analyzing categorical variables and continuous variables, respectively.

The findings on the percentage of outcomes for different diseases hint at the tendency of existing studies to examine different diseases. For some common, chronic, and psychiatric diseases, current studies tended to assess their treatment patterns, whereas for some sudden onset severe diseases, patient mortality and hospitalization or hospital stay were assessed more often. Existing studies have focused more on assessing treatment modalities, physiological or clinical outcomes, and mortality when targeting diseases such as infections, cardiovascular diseases, and neoplasms. Furthermore, although strong trends were detected between eye diseases and treatment patterns, hemic and lymphatic diseases versus mortality, and mental disorders versus mortality (Figure 3), it was difficult to draw any conclusions that reflect clinical importance because of the small sample size. However, these results indicated different distributions of outcomes measured in different diseases, from which we can learn the focus and shortcomings of the existing studies. In addition, the total number of studies measuring guideline adherence was relatively small (n=32). During this

---

**Figure 4.** (A) Percentages of statistical analysis types for each outcome, (B) modeling purposes for each outcome, and (C) specific models for each outcome. GLM: generalized linear model.
period, 63% (20/32) of the studies were conducted on “cardiovascular diseases” and “nutritional and metabolic diseases.” These results also revealed a relative lack of studies measuring guideline adherence in infections. We expect that RWD research on guideline adherence would receive more attention in future.

The percentage of databases used for different diseases implied the selection of databases for observing different diseases. The JMDC databases and DPC database (MHLW) showed opposite use trends in diseases, especially nutritional and metabolic diseases, musculoskeletal diseases, mental disorders, hemic and lymphatic diseases, eye diseases, and skin diseases.

According to the investigation of statistical methods used to assess different outcomes, multivariate models were the most commonly used in assessing mortality. Regardless of the outcome, multivariate modeling was accompanied by adjustments for various confounders (Figure 4B). Mortality, hospitalization or hospital stay, and resource use or costs have been analyzed using multilevel models or marginal models (eg, GEE) more than others. This implies that hospital-related outcomes tended to be assessed by models that took clustering into account. Logistic regression was the first choice for measuring many of the outcomes, with the exception of hospitalization or hospital stay and resource use or costs, for which linear regression was commonly used. Cox proportional hazards regression was suggested as the second choice when assessing mortality, physiological or clinical outcomes, and treatment patterns. Although the PS technique has been proven effective in balancing confounders between groups, it has not been widely used in existing studies. There is a relative preference for this technique in studies assessing mortality.

Comparison With Prior Work

In this subsection, we compare this review with 2 similar studies [27,28]. Hirose et al [27] conducted a narrative review of 68 studies on the secondary use of claims data in a specific database, NDB, from October 2016 to June 2019. They summarized 5 key pieces of information, including study design, research area, setting or sample, outcomes, and strengths and limitations. Subsequently, Fujinaga and Fukuoka [28] conducted a similar narrative review of 643 studies on the secondary use of claims data in 4 large-scale domestic databases: NDB, DPC database (MHLW), JMDC, and MDV, from January 2015 to October 2020, from which 3 categories of research type, design, and area were analyzed descriptively. Both studies used a classification of the journals in which the target articles were published to extract information about the research area [29]. These classifications mixed disciplinary categories, such as clinical medicine, pharmacology and pharmacy, pharmacology and toxicology, and immunology; disease categories, such as infectious diseases; and general categories, such as social sciences and public environmental health. In addition, only the primary outcomes were analyzed in these 2 studies. As a result, the distribution of articles in each category was summarized in these studies.

Because of the partial overlap in search periods, as well as the fact that PubMed was used for the search, there were some articles that were reviewed in both this study and these 2 prior studies. In contrast to these 2 studies, which used 1 or more specific claim databases without specifying a research area, our review investigated domestic epidemiological studies based on the secondary use of 3 types of RWD: claims, EMRs, and EHRs. A further difference is that we defined 7 categories for data collection to assess the status and trends of the existing studies. One of the novelties is that we classified the outcomes with reference to the paper by Abaho et al [24] paper and collected information on all the outcomes measured in the target articles. The advantage of this classification is that these outcomes are also applicable to clinical trial studies and can be automatically identified from biomedical articles [24]. Another innovative point is that we proposed a hierarchical approach to classify the statistical methods that appear in the target articles. For the results of the data collection, we summarized the distribution of the target articles in each category. Additional comparative analyses were performed for diseases versus outcomes (Figure 3), outcomes versus statistical methods (Figure 4), and diseases versus databases (Multimedia Appendix 5), which revealed trends in the assessment of outcomes across different diseases, trends of statistical methods used for different outcomes, and trends in database selection when analyzing different diseases. Moreover, our findings shed light on the focus and shortcomings of previous studies.

In addition, we identified several other review studies on the secondary use of RWD data [30-32]. The paper by Ferver et al [30] provided a narrative review of 1956 claims-based studies in 5 health care journals from 2000 to 2005 by summarizing the research types and areas. The paper by Hutchings et al [31] provided a systematic literature review of 18 studies to investigate the attitudes of relevant practitioners toward the secondary use and sharing of health administrative and clinical trial data. Schlegel et al [32] conducted a literature review of 941 studies on the secondary use of health care data in 2016 to select the best performing articles. We summarized these additional studies to understand other investigations on the secondary use of RWD data. Comparisons were not made because of the survey years or different research purposes.

Limitations

The first limitation of this review is that we only searched the literature in PubMed, which may have led to significant publication bias. Second, we only investigated studies conducted in Japan. In the future, a comparison of studies from other countries, such as the United States, will be necessary to understand the Japan-specific trends of such studies. In addition, searches of multiple electronic databases should be considered to reduce potential publication bias.

Future Directions

In this subsection, we discuss the future perspectives for the use of claims, EMRs, and EHRs in epidemiology in the Japanese context, in terms of the findings of this large narrative literature review.

Organization

Regarding collaborative aspects, with strong national promotion for RWD use and high level of interest from health care firms, collaborative research, involving multiple stakeholders and
academic researchers, is seen to be necessary to leverage academic results and accelerate clinical applications.

**RWD Type**

Notably, only a few studies have used EHRs. EHRs have not been widespread in Japan because of the high cost of implementation and the difficulties in bridging different EHR service vendors. With the promotion of “cloud-based EHR” development by the Japanese Ministry of Internal Affairs and Communications, EHRs are expected to become widely used in the future.

**Disease**

With regard to the disease trend detected in this review, we made a rough comparison with worldwide trends. As we did not find a quantitative survey of RWD research on different diseases, the worldwide trend was roughly estimated by counting the number of related publications for different diseases. We focused on the top-ranked disease areas identified in this review, including infections, cardiovascular diseases, and neoplasms. The number of publications for these diseases was obtained by searching for electronic databases, such as PubMed or PubMed Central with search keywords: combinations of “claims,” “EHR,” “EMR,” to “infection,” “cardiovascular disease,” and “cancer.” We retrieved 18,847 publications on cancer, 7517 publications on infections, and 6624 publications on cardiovascular diseases from PubMed. The same trend was detected in PubMed Central. According to these counts, we estimated that the worldwide trend of the disease examined in existing studies was cancer. In contrast, our results revealed a Japan-specific trend in the studies on infections.

It is important to note that the above counts may be subject to bias because we have not designed any eligibility criteria for the precise search of related publications worldwide. In the future, it will be necessary to compare relevant studies with those of other countries to clarify the Japan-specific status and challenges.

**Statistical Method**

On the basis of the statistical skills used in the eligible articles, we summarized the appropriate statistical methods for use under different conditions. First, to design simple statistical analyses, our findings suggest using Fisher’s exact tests or chi-square test to compare categorical variables, and 2-tailed t test, ANOVA, and Mann-Whitney U test were used to compare continuous variables [33-36]. To evaluate variable change trends, the Cochran-Armitage test was used for categorical variables, whereas the Jonckheere-Terpstra test was used for continuous variables [37].

Suggestions for statistical methods to measure different outcomes are summarized in Table 3. For confounding adjustment, there are 2 methods: covariate adjustment and PS analysis. PS analysis is known to be an effective technique for balancing the patient backgrounds between the 2 groups across all putative risk factors or confounders [38-40]. However, referring to the study by Elze et al [41] that PS analysis is not necessarily superior to conventional covariate adjustment, we suggest selecting PS analysis with caution for confounder adjustment. Our findings also demonstrated that most existing studies used covariate adjustment (n=279) rather than PS analysis (n=96; Multimedia Appendix 4). In addition, hospital-based medical data are frequently clustered within medical centers or physicians. For instance, patients treated in a particular hospital may be more alike than those treated in another hospital because of differences in treatment policies. To model such clustered data, multilevel models with random effects have been suggested for use in estimating predictor effects for patients in specific clusters [42,43].

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Method recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment patterns</td>
<td>Logistic regression, Cox proportional hazards regression</td>
</tr>
<tr>
<td>Physiological or clinical</td>
<td>Logistic regression, Cox proportional hazards regression</td>
</tr>
<tr>
<td>Mortality</td>
<td>Kaplan-Meier analysis, log-rank test, logistic regression, Cox proportional hazards regression</td>
</tr>
<tr>
<td>Hospitalization or hospital stay</td>
<td>Linear regression, GLM&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Adverse events</td>
<td>Logistic regression, Cox proportional hazards regression</td>
</tr>
<tr>
<td>Resource use or costs</td>
<td>Linear regression, GLM</td>
</tr>
<tr>
<td>Guideline adherence</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Quality indicators</td>
<td>Logistic regression</td>
</tr>
</tbody>
</table>

<sup>a</sup>GLM: generalized linear model.

In contrast, there were few studies on predictive machine learning models in this review (n=3; Multimedia Appendix 4). However, we roughly retrieved 2223 publications worldwide on PubMed by searching for the keywords of “claims,” “EHR,” “EMR,” and “machine learning.” Notably, we did not design any eligibility criteria for this study. The large difference in the number of articles indicates that epidemiological research based on claims, EMRs, and EHRs in Japan is backward in the use of artificial intelligence techniques.

**Conclusions**

This literature review provides a good understanding of the current status and trends in the use of claims, EMRs, and EHRs in clinical epidemiology in Japan. The results demonstrated appropriate statistical methods regarding different outcomes, Japan-specific trend of disease areas, and lack of use of artificial intelligence techniques in this country.

https://jmir.org/2023/11/e39876
intelligence techniques in existing studies. We hope that the results of this narrative review will provide useful information for researchers to design relevant studies. In the future, a more precise comparison of relevant domestic research with worldwide research will be conducted to clarify the Japan-specific status and challenges.

Data Availability
All data generated or analyzed during this study are included in this published paper and its multimedia appendices.

Conflicts of Interest
None declared.

Multimedia Appendix 1
PubMed search string.
[DOCX File, 28 KB - medinform_v11i1e39876_app1.docx]

Multimedia Appendix 2
Explanation of the 7 categories of information.
[DOCX File, 32 KB - medinform_v11i1e39876_app2.docx]

Multimedia Appendix 3
Distribution of publication years of the 620 eligible articles.
[PNG File, 153 KB - medinform_v11i1e39876_app3.png]

Multimedia Appendix 4
Full results of counts and percentages of the 7 categories.
[XLSX File (Microsoft Excel File), 21 KB - medinform_v11i1e39876_app4.xlsx]

Multimedia Appendix 5
Percentages of databases used in each disease.
[PNG File, 602 KB - medinform_v11i1e39876_app5.png]

References


Smart Glasses for Supporting Distributed Care Work: Systematic Review

Zhan Zhang¹, PhD; Enze Bai¹, MSc; Karen Joy¹, MSc; Partth Naressh Ghelaa¹, MSc; Kathleen Adelgais², MD; Mustafa Ozkaynak³, PhD

¹School of Computer Science and Information Systems, Pace University, New York, NY, United States
²School of Medicine, University of Colorado, Aurora, CO, United States
³College of Nursing, University of Colorado, Aurora, CO, United States

Corresponding Author:
Zhan Zhang, PhD
School of Computer Science and Information Systems
Pace University
1 Pace Plaza
New York, NY, 10078
United States
Phone: 1 3153992627
Email: zzhang@pace.edu

Abstract

Background: Over the past 2 decades, various desktop and mobile telemedicine systems have been developed to support communication and care coordination among distributed medical teams. However, in the hands-busy care environment, such technologies could become cumbersome because they require medical professionals to manually operate them. Smart glasses have been gaining momentum because of their advantages in enabling hands-free operation and see-what-I-see video-based consultation. Previous research has tested this novel technology in different health care settings.

Objective: The aim of this study was to review how smart glasses were designed, used, and evaluated as a telemedicine tool to support distributed care coordination and communication, as well as highlight the potential benefits and limitations regarding medical professionals’ use of smart glasses in practice.

Methods: We conducted a literature search in 6 databases that cover research within both health care and computer science domains. We used the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology to review articles. A total of 5865 articles were retrieved and screened by 3 researchers, with 21 (0.36%) articles included for in-depth analysis.

Results: All of the reviewed articles (21/21, 100%) used off-the-shelf smart glass device and videoconferencing software, which had a high level of technology readiness for real-world use and deployment in care settings. The common system features used and evaluated in these studies included video and audio streaming, annotation, augmented reality, and hands-free interactions. These studies focused on evaluating the technical feasibility, effectiveness, and user experience of smart glasses. Although the smart glass technology has demonstrated numerous benefits and high levels of user acceptance, the reviewed studies noted a variety of barriers to successful adoption of this novel technology in actual care settings, including technical limitations, human factors and ergonomics, privacy and security issues, and organizational challenges.

Conclusions: User-centered system design, improved hardware performance, and software reliability are needed to realize the potential of smart glasses. More research is needed to examine and evaluate medical professionals’ needs, preferences, and perceptions, as well as elucidate how smart glasses affect the clinical workflow in complex care environments. Our findings inform the design, implementation, and evaluation of smart glasses that will improve organizational and patient outcomes.

(JMIR Med Inform 2023;11:e44161) doi:10.2196/44161

KEYWORDS
smart glass; care coordination; telemedicine; distributed teamwork; mobile phone
Introduction

Background

Effective and timely care coordination and communication are critical components of efficient and safe patient care [1,2]. Failure in providing coordinated care and communicating patient data is seen as one of the root causes of adverse events such as delays in patient care and deviations from standard medical procedures [3]. The challenges in maintaining effective care coordination and communication are exacerbated when care providers are distributed (eg, located in different places) [4,5].

Over the past 2 decades, many telemedicine systems have been developed to augment remote clinical consults [6-8]. During the COVID-19 pandemic, the need for such systems became more obvious. Most telemedicine systems are implemented on desktops or tablet devices [6,7]. However, these devices have practical limitations: (1) desktop systems have limited portability because they are installed in a fixed location; and (2) tablet device-based systems rely on manual input and control, which can hinder usability [9,10]. These issues could result in limited use of technology in real time, especially during complex care environments and time-critical patient scenarios because they demand the full cognitive attention and physical involvement of care providers [11].

In recent years, the use of smart glasses—a computing device worn as a conventional pair of glasses (Figure 1)—has been gaining momentum in health care because they allow for real-time visual communication in a hands-free manner [12,13]. In particular, smart glasses can present both imagery and textual information within the wearer’s field of view (FOV) through a prism and enable videoconferencing for consults or second opinions via a front-facing camera. Since the introduction of smart glasses to the market, researchers have explored their applicability and usefulness in various medical settings and clinical scenarios [9], such as broadcasting surgeries to facilitate resident teaching [14], recording encounters with patients in wound care [15,16], assessing patients in mass casualty incidents [17], and supporting communication between prehospital and hospital providers [18,19].

Figure 1. Examples of smart glasses with various hardware components labeled. (A) Google Glass. (B) Vuzix M400.

Objectives

As there is a growing interest in using smart glasses to support care coordination and communication across distributed care providers [9,11,20], the aim of this study was to synthesize the knowledge and experiences in this area, understand the benefits and limitations regarding adopting smart glasses as a telemedicine tool, and inform the design of future smart glass applications to better support remote care coordination. We focused on the use of smart glasses in care coordination in various clinical settings (eg, surgical operation, emergency care,
and intensive care unit). Our specific research questions were as follows:

1. What are the general characteristics of prior research on using smart glasses for care coordination?
2. How was the system designed, used, integrated, and evaluated in supporting communication and care coordination across distributed care providers?
3. What types of challenges were identified by medical providers while they were using or testing the smart glass technology in practice?

These research questions were answered through a systematic literature review covering research within both health care and computer science fields.

Our work contributes the following to the medical informatics community: (1) an in-depth analysis and synthesis of prior research on the use of smart glasses for care coordination and communication; and (2) methodological and design implications for future research on smart glasses to improve distributed care coordination and communication.

### Methods

#### Data Search

Our search started with discussing the search time frame and the most appropriate databases to use as well as search terms with experienced librarians. Using technology keywords such as “smart glasses” and “heads-up display,” along with health care keywords such as “distributed care” and “telemedicine,” a health librarian performed database searches for articles published between January 1, 2000, and March 1, 2022. We chose this time frame to capture the evolution of this technology (ie, from early concepts such as head-worn displays [21] to smart glasses, which became a well-known concept after the introduction of Google Glass in 2013 [22]). The full list of search terms is presented in Textbox 1. We chose the following databases to cover research within both health care and computer science: ACM Digital Library, Cochrane Library, IEEE Xplore, Ovid MEDLINE, Embase, and Web of Science. A sample search strategy for Ovid MEDLINE is illustrated in Textbox 2. The database searches were set to include only studies published in peer-reviewed journals and conference proceedings in English. Literature reviews, dissertations, posters, and extended abstracts were excluded from the literature search. The retrieved citations were stored and managed using EndNote bibliographic management software (version X9; Clarivate).

### Textbox 1. Keywords for literature search.

**Search concepts and specific keywords**

- Smart glass: smart glass, augmented reality glasses, heads-up display, head-mounted, head-worn, virtual reality, augmented reality, mixed reality, wearable technology, Google Glass, Vuzix, Epson Moverio
- Clinical: distributed care, remote care, telehealth, telemedicine, telecare, emergency care, pre-hospital

### Textbox 2. A sample search strategy for MEDLINE.

**Search steps**

1. (“distributed healthcare” or “distributed care” or “remote care” or tele* or nursing or “long term care” or “home health” or “home care” or prehospital or pre-hospital or “emergency medical” or “emergency care” or paramedic* or ((clinical or surg*) adj3 (application* or use* or implementation*))ti,ab,kf. or exp Telemedicine/ or exp Home Care Services/ or exp Emergency Medical Services/
2. ((smart adj1 glass*) or smartglass* or Hololens or picolinker or (google adj1 glass*) or vuzix or “epson moverio” or “augmented reality” or (AR and augmented) or “mixed reality” or “virtual reality” or (VR and virtual) or “wearable technology” or wearables or “heads up” or “head mounted” or “head worn”).ti,ab,kf. or wearable electronic devices/ or smart glasses/ or augmented reality/ or virtual reality/
3. Steps 1 and 2
4. Limit step 3 to (english language and yr="2000-Current")
5. (training* or education* or simulation* or telephon* or teleconferenc* or television*).ti. or exp *education/ or *telephone/ or *television/
6. Step 4 not step 5

**Article Screening and Selection**

We used the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology to search and screen articles [23]. Figure 2 outlines the number of records that were identified, included, and excluded through different phases. More specifically, 5865 articles were identified through database searches, of which 5862 (99.95%) were included for screening after removing duplicates. Article titles were screened first, followed by abstract screening, to identify relevant articles. Of the 5862 articles, after screening of article titles, we excluded 5341 (91.11%); of the remaining 521 studies, 446 (85.6%), were excluded, leaving 75 (14.4%) for full-text review. After reviewing the full text of these 75 articles, we deemed 21 (28%) to be eligible for this systematic review.
Three authors (EB, KJ, and PG) independently screened all papers through the paper stack and selected relevant papers for inclusion. Two senior researchers (ZZ and MO) oversaw the whole article review and selection process. Any conflicts in selection decisions were resolved through discussion among all the authors during weekly group research meetings. The inclusion criteria were peer-reviewed articles that reported the use or testing of any smart glass technology and accompanying software in the context of communication and collaboration across distributed care providers. Articles were excluded if they only reported the use of smart glasses by an individual or in a collocated clinical setting or if they did not provide adequate supporting information, such as what clinical setting the smart glasses were used in and who used the technology.

**Data Extraction, Analysis, and Synthesis**

Guided by the research questions of this study, 2 authors (KJ and EB) used a Microsoft Excel spreadsheet to extract, collate, and summarize data from the included studies, such as the country where the study was conducted, study objectives and scope, clinical scenarios, system evaluation methods, technology specifics, barriers and challenges, and a summary of study findings. **Textbox 3** summarizes these data fields and their brief definitions. In addition to extracting the aforementioned metadata, we also assessed the technology readiness levels (TRLs) [24] of the systems tested in the reviewed studies. There are 9 different TRLs, ranging from level 1 (scientific knowledge generated underpinning hardware and software technology) to level 9 (actual system “flight proven” through successful mission operations). Two authors (KJ and EB) followed the metrics proposed in the study by Engel et al [25] and independently assessed TRLs for each system. They then compared and discussed their TRL evaluations until they reached agreement.

Two senior researchers (ZZ and MO) reviewed all the articles and analyses as a verification step. The research team met regularly to discuss the results. We performed the data analysis iteratively (ie, we went back and forth as more knowledge was obtained), as suggested by prior work [11,26]. A meta-analysis of the study results was not considered in this work owing to the heterogeneity of the study designs and results.

In the following section, we report information that was synthesized from the reviewed articles, including characteristics of the selected studies, system architecture and features, TRLs of the reviewed systems, system evaluation methods, and care providers’ perceived benefits and challenges of using and adopting smart glasses for distributed care coordination.
Textbox 3. Assessed article information and metadata.

Assessed information and brief definition

- Study objectives and scope: the objective of the research and the purpose and scope of the use and test of smart glasses in each study (eg, patient care vs medical training)
- Clinical scenarios: the clinical domain and context in which the study was conducted
- Publication details: the type (eg, journal article vs conference paper), region, and year of the publication
- System infrastructure: the hardware, software, and network setup on both local and remote sites for establishing teleconsultation
- System features: the system features used, developed, or evaluated in each study
- System evaluation: the aspects of the smart glass system that were evaluated in the study and the methods used for system evaluation
- Benefits and challenges: the reported benefits and challenges of using smart glasses in improving communication and care coordination among distributed medical teams
- Major study findings: a summary of the major findings of a study

Results

General Characteristics of the Reviewed Studies

Of the 21 reviewed articles, 10 (48%) were conducted in the United States [18,19,27-34], and 2 (10%) were conducted for surgical teleproctoring between high-income countries and low- and middle-income countries (LMICs), such as between surgeons in the United States and Mozambique [35] and between experienced surgeons recruited from the United States and Germany and novice surgeons in Brazil and Paraguay [36]. The remaining studies (9/21, 43%) were conducted in different countries, such as Spain [37], China [38], Germany [39], France [40], Italy [41], Switzerland [42], Malaysia [43], South Korea [44], and Republic of the Congo [45]. The reviewed studies were conducted to assess the feasibility, effectiveness, and user experience of smart glasses in supporting remote patient evaluation and care procedure operation in a particular medical domain. The study objectives, along with major findings for each reviewed article, are presented in Multimedia Appendix 1 [18,19,27-45].

The clinical foci in these 21 papers vary: 9 (43%) focused on surgical settings [29,30,33-38,44], whereas 6 (29%) focused on the prehospital or emergency medical services domain [18,19,28,31,39,42]. The remaining studies (6/21, 29%) focused on intensive care [40,43], toxicology [27], ophthalmology [32], pediatric cardiology [41], and general medicine [45].

The scope and purpose of the use of smart glasses among these studies vary. As shown in Figure 3A, the majority of the reviewed studies (16/21, 76%) used smart glasses to enable remote patient care and evaluation [18,19,27,28,30-32,37-45]. Of these 16 studies, 8 (50%) [27,28,30,32,37,38,43,45] tested smart glasses with real patients, 6 (38%) [18,19,31,39,40,44] conducted system testing in a simulated environment, and 2 (13%) [41,42] did not specify how the device was tested. The remaining studies (5/21, 24%) [29,33-36] leveraged smart glasses for training and teleproctoring purposes; of these 5 studies, 4 (80%) [29,34-36] tested the device with real patients, whereas 1 (20%) [33] tested the device in a simulated environment.

The reviewed articles were published between 2014 and 2021 (Figure 3B). It is noticeable that almost half of the reviewed articles (9/21, 43%) were published within the first 3 years of the release of Google Glass [22]. Subsequently, the number of studies on the use of smart glasses for supporting distributed care decreased until 2021. One possible explanation for this finding is that the use of smart glasses regained momentum right after the outbreak of the COVID-19 pandemic as researchers started exploring smart glass use to enable medical personnel to participate in remote assessment and consultation, with the aim of safeguarding patients and health care providers during the pandemic.
System Architecture

Although the system architecture implemented in each study varied, there were some similarities across the reviewed studies. Typically, there are two types of technology setups on the local site: (1) smart glasses are connected to a Wi-Fi network, a Wi-Fi hotspot, or a mobile router to directly stream the first-person point-of-view to a remote consultant (Figure 4A); or (2) smart glasses are connected to a smartphone or a laptop via Bluetooth or Wi-Fi for video streaming and audio transmission (Figure 4B). The first approach was adopted by 52% (11/21) of the studies [19,27,31,32,36,38,39,41,43,44], and the second approach was used in 33% (7/21) of the studies [18,29,35,42,45]; for example, in the study by Diaka et al [45], the smart glasses were designed as an extension of a smartphone, which meant that the local wearer needed to initiate the call on the smartphone. Regardless of the system implementation method on the local site, the remote experts were usually equipped with either a computer or a mobile device (eg, a tablet device) to review and access the video stream and other multimedia data shared by the local medical practitioner (Figure 4). However, it is worth mentioning that in the study by Brewer et al [33], where smart glasses were used for surgical training, the remote expert (trainer) also wore a pair of smart glasses to view the video streamed from the learner.

Figure 4. Common system architecture setups in the reviewed studies. (A) Smart glasses connected to a Wi-Fi network, a Wi-Fi hotspot, or a mobile router. (B) Smart glasses connected to a smartphone via Bluetooth or Wi-Fi.

As shown in Figure 5A, the reported brands of smart glasses in these studies included Google Glass [18,19,27-30,32-36,40,42,44], Vuzix [38,43], Iristick [37,45], Pivothead Original Series [32], Intel Recon Jet [31], and Epson Moverio BT-200 [41]. Google Glass was the most frequently used smart glass device (13/20, 65%). Another interesting observation is that all of the studies (21/21, 100%) used off-the-shelf, commercialized videoconferencing software (Figure 5B) such as Pristine Eyesight [19,27], AMA XpertEye [28,35], Livestream [18,36], WebRTC (enabled by Google) [42,44], Livecast Media [38], Skype [29], CrowdOptic [33], Google Hangout [34], and Polycom RealPresence Group 500 [32]. Most of the videoconferencing software used was compliant with the Health Insurance Portability and Accountability Act (HIPAA) rules, except in the case of the study by Cicero et al [18], where the researchers only tested the use of smart glasses in a simulated environment (real patient care was not involved).

Figure 5. (A) Smart glass brands used in the reviewed articles. (B) Videoconferencing tools used in the reviewed articles.

System Features

Although there was variation in the application scopes and domains, there were some common software features across the reviewed studies (Textbox 4). Real-time synchronous video and audio streaming from the local smart glass wearer to the remote consultant is the most common feature among the studies (19/21, 90%). In the case of the exceptions (2/21, 10%), because of technical limitations (eg, limited internet connection), the study by Gupta et al [30] first recorded patient care and evaluation...
using smart glasses and then transmitted the recordings to remote experts at a later time to simulate real-time telemedicine consults, whereas in the study by Hashimoto et al [34], researchers used Google Glass and an Apple iPhone to capture videos of a surgical operation and compared the video quality and its adequacy for safe use in telementoring.

Another noteworthy feature is enabling imagery and text-based remote guidance and annotation; for example, the remote consultant can annotate images captured from the live stream and project them back onto the local glass wearer’s visual field [35,37]. In 19% (4/21) of the studies [19,27,36,44], the remote consultant could use the texting feature to type messages that could be projected onto the smart glass display. These annotation features provide the remote consultant with more channels (in addition to audio and video) to direct and guide local medical practitioners to perform critical procedures.

Augmented reality (AR)—a technique that can enhance an individual’s visual experience of the real world through the integration of digital visual elements—was also tested in several studies. In Ponce et al [29], for example, AR enabled a remote surgeon to insert their hands or instruments virtually into the visual field of the local surgeon who wore smart glasses for real-time guidance, training, and assistance as needed. In another study [41], a remote specialist used AR-based markers to guide the execution of an echocardiographic examination performed by a local operator. The markers were overlaid on the ultrasound device and could be seen through the screen of the local operator’s smart glasses.

Other features of smart glasses reported in the studies included zooming in and out of the live stream video [35]; using voice commands [27,28,30,31] or head movements [27] to control, and interact with, the smart glass device; taking photographs [19,30,31,35]; automatically detecting the geographic location of on-site medical teams with the built-in GPS [31]; and presenting prehospital triage algorithm on the glass screen for decision support during mass casualty incidents [39].

**Textbox 4. Summary of smart glass features as described in the reviewed studies.**

<table>
<thead>
<tr>
<th>System features</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Real-time synchronous video and audio streaming</td>
</tr>
<tr>
<td>- Record and forward video recordings</td>
</tr>
<tr>
<td>- Imagery and text-based remote guidance and annotation</td>
</tr>
<tr>
<td>- Augmented reality</td>
</tr>
<tr>
<td>- Zooming in and out of the live stream video</td>
</tr>
<tr>
<td>- Hands-free interaction with smart glasses</td>
</tr>
<tr>
<td>- Taking photographs</td>
</tr>
<tr>
<td>- GPS-based tracking of the geographic location of on-site medical teams</td>
</tr>
<tr>
<td>- Presenting prehospital triage algorithm on the glass screen for decision support</td>
</tr>
</tbody>
</table>

**TRLs of the Systems Tested in the Reviewed Studies**

On the basis of our analysis, we found that the TRLs of all the systems used or tested in the reviewed studies ranged between 7 and 9. Our TRL assessment for each system is visualized in Figure 6 [18,19,27-45]. The reasoning for our assessment is summarized in Multimedia Appendix 2 [18,19,27-45].
The systems in 24% (5/21) of the studies [18,34,40-42] have a TRL of 7, which indicates that the technology is in the form of a high-fidelity prototype and has all key functionality available for demonstration and test; for example, in the study by Widmer and Müller [42], the Google Glass device on the local site was set up to connect with a computer application on the remote site for teleconsultation. This integrated system was only preliminarily tested by the research team but not in a simulated or real environment (a criterion for TRL 8); thus, its TRL was set to 7. It is worth mentioning that of these 5 studies, 3 (60%) [18,40,41] tested smart glasses in simulated scenarios; however, there were several reasons for their failure to meet the criteria for TRL 8, such as using non–HIPAA-compliant videoconferencing software, testing the technology with only 1 volunteer, or not fully integrating smart glasses with the network and remote devices.

The majority of the studies (15/21, 71%) [19,27-33,35-39,43,44] tested or used systems that met the criteria for TRL 8, indicating that they are actual systems in their final configuration and have been fully developed and tested in either simulated or real operational scenarios. However, these studies provided limited information regarding some criteria for TRL 9, such as whether the system had been fully integrated with other operational hardware and software systems (eg, database and hospital IT infrastructure), whether all system documentation had been completed, whether training on system use was available, and whether engineering support team was in place. Without such information, it is difficult to assess the readiness of these systems for large-scale deployment.

In comparison, only the system in the study by Diaka et al [45] was assessed to have a TRL of 9 because the system had been successfully operated on actual missions and tasks in the operational environment for a relatively long time (ie, more than a year). Furthermore, the system was fully integrated with other operational software, hardware, and network devices, as well as care delivery services (eg, moto-ambulances to facilitate patient referrals after teleconsultation).

### System Evaluations

**Overview**

The reviewed studies evaluated different dimensions of the smart glass system, including technical feasibility, effectiveness, and user experience. The details regarding the aspects of the smart glass system that were evaluated as well as the evaluation methods used in the reviewed studies are summarized in Table 1 and then elaborated on in the following sections.

<table>
<thead>
<tr>
<th>Evaluated dimensions</th>
<th>Specific evaluated aspects</th>
<th>Evaluation methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical feasibility [27,34-36,44]</td>
<td>• Success rate of established video teleconsultations between local and remote medical practitioners [27,36]</td>
<td>• Researchers’ observations of the successfullness of teleconsultations [27,36]</td>
</tr>
<tr>
<td>Effectiveness [18,19,27,28,30-33,36,39,40,43]</td>
<td>• Compared with in-person patient evaluation, whether the use of smart glasses could achieve similar performance and accuracy regarding patient evaluation and diagnosis [19,28,32,43]</td>
<td>• Comparison study between control (without smart glass support) and treatment (with smart glass support) groups [18,19,28,32,39,40,43]</td>
</tr>
<tr>
<td></td>
<td>• Compared with either mobile phone–based or no remote patient consultation, whether the use of smart glasses could lead to changes in clinical management and remote consultant’s confidence regarding diagnosis [18,27,30,39,40]</td>
<td>• Questionnaire [33]</td>
</tr>
<tr>
<td></td>
<td>• Whether the use of smart glasses could improve medical training (eg, surgical operation) [33,36]</td>
<td>• Exit interview [36]</td>
</tr>
<tr>
<td></td>
<td>• Opinions regarding using and adopting smart glasses in practice [18,27,30,31,35,37-41,43-45]</td>
<td>• Interviews and observations [18,31,35,45]</td>
</tr>
</tbody>
</table>

### Technical Feasibility

Several studies assessed whether the smart glass technology was a practical means to support care coordination and communication in different contexts, such as teletoxicology consults [27] and remote surgical teleproctoring [34-36,44]. The main measurements included the success rate of established video teleconsultations between local and remote medical practitioners and whether the quality of video streaming was acceptable and good enough to allow for real-time, seamless guidance and assistance. The technical feasibility was primarily determined by the researchers’ observations and the users’ ratings via questionnaire; for example, in a study evaluating the feasibility and acceptability of Google Glass for teletoxicology consults [27], questionnaires were administered immediately after the study to elicit remote consultants’ opinions regarding whether consults through smart glasses were considered successful and the technical feasibility of using smart glasses for teleconsultation.

### Effectiveness

Of the 21 reviewed studies, 10 (48%) evaluated the effectiveness of smart glasses, that is, whether this novel technology could improve patient care and decision-making compared with current approaches (eg, no remote consultation, in-person patient evaluation, or consultation via telephone).
[18,19,27,28,30,32,33,39,40,43]; for example, in some settings where remote consultations were usually accomplished via telephone or radio, which typically do not support visual communications [27,30], researchers compared using such traditional communication mechanisms with using smart glasses to determine whether the use of smart glasses could lead to changes in clinical management and the remote consultant’s confidence regarding diagnosis.

Of these 10 studies, 7 (70%) [18,19,28,32,39,40,43] conducted an experiment with a control group (no smart glasses and either in-person consultation or no remote consultation at all) and an intervention group (with smart glasses) to measure whether using smart glasses could increase the quality and accuracy of patient diagnosis while reducing the time needed to perform patient care; for example, in the scenario of patient triage during mass casualty incidents [19], researchers asked 2 emergency medicine (EM) physicians (control group) to make triage decisions after examining the simulated patients in person as 2 other EM physicians (intervention group) simultaneously evaluated the same group of patients via real-time point-of-view video stream from a paramedic wearing Google Glass. They then used the agreement within and among the groups of EM physicians on the need for immediate trauma evaluation to determine the effectiveness of smart glasses for supporting patient triage.

**User Experience**

Of the 21 studies, 15 (71%) examined end users’ experience and perceptions to some extent with regard to using smart glasses in their work [18,19,27,29-31,35,37-41,43-45]. The primary methodology used for eliciting user experience was a survey, which was adopted by 80% (12/15) of these studies [18,19,27,29-31,35,37-40,43,44]; for example, in a recent study [43], a survey was sent to the participants on completion of the study to assess acceptance, satisfaction, overall impact, efficacy, and potential of adopting smart glasses as an alternative method of teleconsultation in neurosurgery. Among these 12 studies that administered a survey, 9 (75%) specifically reported the number of participants, which ranged between 2 and 276. Other methods such as interviews and observations were also used to gather more qualitative, in-depth insights from end users [18,31,35,45]. In particular, of these 4 studies, 2 (50%) [31,35] conducted interviews in conjunction with a survey.

It is also worth mentioning that of the 15 studies, 2 (13%) [19,31] specifically focused on evaluating the usability of smart glasses, that is, whether smart glass technology is perceived as easily usable by, and acceptable to, medical professionals. Another study [30] also examined patient perceptions of medical providers wearing smart glasses with recording capability. Finally, of the 15 studies, 5 (33%) [29,37,38,41,42] mentioned that they collected end users’ opinions and experiences but did not specify the methods they used.

**Benefits and Challenges of Using and Adopting Smart Glasses for Teleconsultation**

**Benefits**

Our reviewed work highlights the advantages of smart glasses in improving communication and care coordination among distributed medical teams because this technology enables local medical providers to share visual information and perform teleconsultation in a hands-free manner. Regarding the effects on clinical care and patient outcome, the studies reported that smart glasses could shape clinical management and boost remote consultants’ confidence in clinical care [27,30], achieve diagnostic accuracy comparable with that achieved in in-person patient examination [19,28,32,43], improve proficiency and performance of the clinical tasks [31,33,35,38-40], and lower the medical service cost and improve quality of life for people in rural areas or LMICs [36,38]. Finally, many studies reported positive user perceptions, acceptance, and satisfaction with the use of smart glasses [19,27,29-31,35,38,39,41,43,45].

Notwithstanding these reported benefits, the reviewed studies also highlight a set of challenges and user concerns regarding the adoption of smart glasses in practice. We grouped them into 4 main categories: technical challenges, human factors and ergonomics, privacy and security concerns, and organizational challenges (Textbox 5).
Textbox 5. Challenges to using and adopting smart glasses in practice.

Technical challenges
- Unstable or low-bandwidth internet connections [18,19,29,33,35,36,39,44]
- Battery drain becomes higher during video streaming [18,29,39]
- The microphone is unable to filter out background noise [18,29]
- Screen contrast and readability issues in bright or dark environments [18]
- Image distortion owing to overexposure to room light [18,29,35]
- Smart glass see-through screen is too small for easy interaction [41]
- Difficulty controlling video streaming software [18,35,38]
- Lack of a lock function to prevent the possibility of inadvertently halting the video streaming and ability to opt out of frequent software updates [18]

Human factors and ergonomics
- Compatibility issues with wearer’s glasses or personal protective equipment [27,29,35,37,39-41]
- Misalignment between the direction of gaze and range of smart glass camera [29,35,37,40,41,43]
- Voice control function could be problematic [18,30]
- Added distractions for medical professionals [31]

Privacy and security concerns
- Concerns regarding violations of patient privacy and data breach [28-30,43]

Organizational challenges
- Added workload for medical professionals [39]
- Costly device and software [35]
- End users have limited experience with, and prior knowledge of, smart glasses; need extensive equipment and software training [27,37,41,43]

Technical Challenges
The reviewed studies reported a variety of technical challenges that may impede the effective use of smart glasses in teleconsultation. These challenges are mainly related to internet connections, hardware limitations, and software reliability. More specifically, because smart glasses require a high-speed network to transmit visual media (eg, video streaming, audio, and pictures), unstable or low-bandwidth internet connections were seen as a major technical barrier because this issue would compromise video and audio quality, leading to breakdowns in communication and loss of patient information [18,19,29,33,35,36,39,44]. This is more evident in low-resource or out-of-hospital settings where medical practitioners have limited access to the internet; for example, because Wi-Fi is not steadily available in the prehospital environment, the problem with internet connections was commonly reported in this domain [18,19,39]. One practical and successful solution used by a study in prehospital communication [31] was using a mobile router to provide a fault-tolerant network that ran independent of Wi-Fi and other external networks, allowing for deployment at any location.

Regarding hardware limitations, medical professionals were concerned about battery life (eg, the battery could get drained quickly during video streaming) [18,29,39], microphone sensibility (eg, not being able to filter out background noise) [18,29], screen contrast and readability (eg, hard to read the screen in extremely bright or dark environment) [18], image quality (eg, the image could be distorted because of overexposure to room light) [18,29,35], and small screen for interaction [33,41,44].

Issues regarding software were primarily related to controlling and interacting with the video streaming software; for example, 14% (3/21) of the studies [18,35,38] mentioned difficulties regarding zooming in or out during video streaming; as such, the smart glass wearer needs to bring their face close to the patient. Other software issues included the lack of a lock function to prevent the possibility of inadvertently halting the video streaming and the inability to opt out of frequent software updates [18].

Human Factors and Ergonomics
Many issues related to the interactions between users and the smart glass system were also reported. First, 38% (8/21) of the studies [27,29,33,35,37,39-41] highlighted the compatibility issue with users’ spectacles or personal protective equipment. In particular, fitting the smart glass headset onto surgical loupes was problematic, interfering with the surgeon’s ability to wear such devices [35]. Some users had to remove their spectacles to wear the smart glass headset or tie up their hair to prevent the glass camera from being hidden [40]. Second, the difference in line of sight—misalignment between what the glass wearer sees (eg, the direction of gaze) and what the camera captures (eg, range and angle of the camera)—was also cited as a major
This issue was often attributed to the limited FOV of smart glasses [33,44]. This misalignment problem could be worsened owing to sudden head movements and frequent relocation of the smart glass wearer or the patient’s unpredictable movements because these could cause motion blur for remote experts or consultants and make it difficult for them to identify the clinical situation [44]. Third, although the reviewed studies reported that their participants perceived that the smart glass was easy to use overall, usability issues still exist; for example, the voice control function did not work perfectly and thus required the user to remove their gloves to use the built-in touchpad or buttons to operate the device, such as starting or stopping the video call [18,30]. In another study, smart glasses were reported to be a distraction for medical practitioners [31].

Privacy and Security Concerns

Patient privacy and data security issues were perceived as important to address because smart glasses can transfer or even store sensitive patient data [28-30,43]. These studies stated that any implementation of smart glasses must not only comply with HIPAA requirements but also alleviate patient concerns about any potential privacy violation or misuse of their data [30,43].

Organizational Challenges

As medical professionals have limited prior knowledge of using the novel smart glass technology (compared with their experience of using smartphones or tablet devices), a few studies mentioned that user training is necessary to increase efficiency and reduce human errors in system operation [27,33,37,41,43]. In addition, the smart glass technology is costly; for example, as McCullough et al [35] reported, the cost of a yearly contract for a piece of wearable hardware and the videoconferencing platform is approximately US $7000. Such high costs could become a critical barrier to adopting this technology at scale, especially for those health care providers who have limited resources. Finally, integrating smart glasses into the current workflow is a prominent challenge; for example, Follmann et al [39] reported that adopting smart glasses in prehospital triage and communication added more workload to emergency care providers in the field and took markedly more time compared with not using smart glasses.

Discussion

Methodological Implications

In this work, we conducted a systematic review of studies focused on the use and application of smart glasses in supporting care coordination and communication among distributed medical teams. Of the 5862 papers included for screening, only 21 (0.36%) met our criteria, highlighting the paucity of studies examining the feasibility, effectiveness, and user experience of using smart glasses as a telemedicine tool. Furthermore, the studies were mostly conducted in the United States and a few other high-income countries (eg, Italy, Germany, and France). One possible explanation is that smart glass technology is costly, hindering its adoption in LMICs and low-resource settings. However, 14% (3/21) of the reviewed studies [35,36,45] revealed the substantial benefits that smart glasses could bring to LMICs and rural areas, such as providing remote training and mentoring and more accurate instructions to the field medical practitioners in low-resource settings who otherwise have limited access to remote experts. Given such benefits, more future work is needed to expand the research of smart glasses to LMICs.

Another interesting observation is that all the reviewed studies (21/21, 100%) only used off-the-shelf hardware and software without involving users in the system design process. Prior work has suggested that it is critical to involve users and understand user requirements in the early phase of system development to identify and address potential usability and technical issues [6,46,47]. In addition, regarding the methodology for eliciting user opinions, out of 15 studies conducted user evaluation, 33% (5/15) of them did not specify what questions they asked, how the questionnaire was developed, and what procedure was followed. Despite the user-friendliness of health care information technology being a determinant factor for user adoption and acceptance [48,49], the usability of smart glasses was neglected by most of the studies (19/21, 90%), with only the studies by Broach et al [19] and Demir et al [31] specifically examining this aspect. These facts highlight the need to adopt a user-centered design approach in the development of smart glass technology by placing users at the center of the system design process from inception to implementation and deployment.

A similar concern is that a few of the reviewed studies (4/21, 19%) only recruited a small number of study participants (eg, 2 health care professionals) to participate in their user studies (eg, survey or interview). In addition, some of the studies (5/21, 24%) did not report the details of their user research, including the number of participants. These findings may suggest that the important role of user research was not recognized in some of the reviewed studies (9/21, 43%), and their results might not be generalizable because of the limited number of study participants. Given these study limitations, we argue that involving human-computer interaction researchers in such type of research and establishing close collaborations between these researchers and health care domain experts are critical and much needed, as demonstrated in the study by Schlosser et al [50].

Finally, almost all of the reviewed studies (20/21, 95%) focused on evaluating the smart glass technology either from a technical perspective or a clinical perspective, while neglecting other important factors that could substantially affect the use and adoption of this technology, such as workflow, teamwork, policies, and organizational cultures. As prior work has argued [51], an ongoing challenge to the successful implementation and deployment of health IT (HIT) interventions is to operationalize their use within the workflow of a complex health care system; for example, a new technology could disrupt current clinical work, causing not only frustrations for medical providers but also patient safety issues [52-54]. When this problem occurs, not surprisingly, medical practitioners are left with no choice but to bypass the technology or adopt informal, low-tech, potentially unsafe workarounds that deviate from the formal protocol [55,56]. As such, researchers have highlighted the importance of examining the design, use, and application of HIT interventions through the lens of a sociotechnical
perspective [55-57]. This approach allows researchers and practitioners to understand the complex interrelations between various social and technical elements of systems that are equally important in determining the success of HIT adoption in a health care organization. In line with this argument, we believe that more research adopting a sociotechnical model [51,58] is needed to investigate the factors (eg, human-computer interaction, workflow and communication, internal organizational features, and external rules) that contribute to the uptake of smart glasses in routine use.

**Design Implications**

The reviewed studies revealed a set of challenges and barriers to adopting and using smart glasses in practice; for example, a commonly cited technical challenge is internet connection quality—smart glasses rely on a high-bandwidth internet network for streaming videos and transmitting other visual media data (eg, high-resolution pictures, texts, and augmented objects). However, this technical requirement could be challenging to fulfill, especially in low-resource or out-of-hospital settings [59]. With the rapid development of 5G technology, this technical barrier might be overcome in the near future; for example, a study [60] showed that 5G technology could not only enable safe and efficient complex surgical procedures during telementored surgery but also lead to a very high degree of surgical team satisfaction. In addition to internet connections, other technical improvements suggested by the reviewed studies include increasing the memory space of smart glasses to store more information, adding autofocus and stabilization features to the smart glass camera, and improving the camera resolution [35].

Human factors and usability issues make up another set of important considerations for smart glass designers and developers; for example, the difference in line of sight between the local medical practitioner and remote consultant impeded the remote consultant from seeing exactly what the smart glass wearer’s eyes were fixed on. In addition, the limited FOV further complicated the video transmission to the remote experts. One reviewed study [44] experimented by attaching a mirror to the smart glass to increase the FOV of the local practitioner by stabilizing features to the smart glass camera, and improving the camera resolution [35].

Defining the search keywords was difficult. To generate a comprehensive and relevant list of keywords, we iteratively discussed and selected the keywords for the search based on suggestions from the health librarian and a review of systematic review articles regarding smart glasses. Another limitation is that we did not assess the quality or impact of the results from the included articles. A meta-analysis was not feasible because of the heterogeneity of the study designs and results.

**Conclusions**

Smart glasses were found to be an acceptable and feasible tool in enabling visual communication and information sharing among distributed medical teams. Despite the high potential of this novel technology, the reviewed articles pointed out a set of challenges that need to be addressed before the wide deployment of this technology in complex health care systems. Thoughtful system design involving end users from the beginning and improved hardware and software reliability are needed to improve the usefulness and usability of smart glasses for medical practitioners [11,59]. We suggest that more user-centered design and evaluation research is needed to examine and evaluate medical professionals’ needs and perceptions and determine how to design smart glass technology to meet their needs. In addition, more research is required to elucidate how smart glasses affect the workflow of medical professionals in complex care environments.

**Acknowledgments**

The authors thank the health librarians Lilian Hoffecker and Ben Harnke at the University of Colorado for performing the literature search and documenting the search process and results. This study was supported by the National Science Foundation (grant 1948292) and the Agency for Healthcare Research and Quality (grant 1R21HS028104-01A1).

**Conflicts of Interest**

None declared.
Multimedia Appendix 1
Summary of study objectives and major findings.
[DOCX File, 28 KB - medinform_v11i1e44161_app1.docx]

Multimedia Appendix 2
Technology readiness levels of the systems reported in the reviewed studies.
[DOCX File, 27 KB - medinform_v11i1e44161_app2.docx]

References


Review

Mining Sensor Data to Assess Changes in Physical Activity Behaviors in Health Interventions: Systematic Review

Claudio Diaz¹, MBE; Corinne Caillaud², PhD; Kalina Yacef¹, PhD

¹School of Computer Science, The University of Sydney, Sydney, Australia
²Charles Perkins Centre, School of Medical Sciences, The University of Sydney, Sydney, Australia

Corresponding Author:
Kalina Yacef, PhD
School of Computer Science
The University of Sydney
Building J12/1 Cleveland Street
Camperdown NSW
Sydney, 2006
Australia
Phone: 61 (02) 9351 2222
Email: kalina.yacef@sydney.edu.au

Abstract

Background: Sensors are increasingly used in health interventions to unobtrusively and continuously capture participants’ physical activity in free-living conditions. The rich granularity of sensor data offers great potential for analyzing patterns and changes in physical activity behaviors. The use of specialized machine learning and data mining techniques to detect, extract, and analyze these patterns has increased, helping to better understand how participants’ physical activity evolves.

Objective: The aim of this systematic review was to identify and present the various data mining techniques employed to analyze changes in physical activity behaviors from sensors-derived data in health education and health promotion intervention studies. We addressed two main research questions: (1) What are the current techniques used for mining physical activity sensor data to detect behavior changes in health education or health promotion contexts? (2) What are the challenges and opportunities in mining physical activity sensor data for detecting physical activity behavior changes?

Methods: The systematic review was performed in May 2021 using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. We queried the Association for Computing Machinery (ACM), IEEE Xplore, ProQuest, Scopus, Web of Science, Education Resources Information Center (ERIC), and Springer literature databases for peer-reviewed references related to wearable machine learning to detect physical activity changes in health education. A total of 4388 references were initially retrieved from the databases. After removing duplicates and screening titles and abstracts, 285 references were subjected to full-text review, resulting in 19 articles included for analysis.

Results: All studies used accelerometers, sometimes in combination with another sensor (37%). Data were collected over a period ranging from 4 days to 1 year (median 10 weeks) from a cohort size ranging between 10 and 11615 (median 74). Data preprocessing was mainly carried out using proprietary software, generally resulting in step counts and time spent in physical activity aggregated predominantly at the daily or minute level. The main features used as input for the data mining models were descriptive statistics of the preprocessed data. The most common data mining methods were classifiers, clusters, and decision-making algorithms, and these focused on personalization (58%) and analysis of physical activity behaviors (42%).

Conclusions: Mining sensor data offers great opportunities to analyze physical activity behavior changes, build models to better detect and interpret behavior changes, and allow for personalized feedback and support for participants, especially where larger sample sizes and longer recording times are available. Exploring different data aggregation levels can help detect subtle and sustained behavior changes. However, the literature suggests that there is still work remaining to improve the transparency, explicitness, and standardization of the data preprocessing and mining processes to establish best practices and make the detection methods easier to understand, scrutinize, and reproduce.

(JMIR Med Inform 2023;11:e41153) doi:10.2196/41153

https://medinform.jmir.org/2023/1/e41153

JMIR Med Inform 2023 | vol. 11 | e41153 | p.52
(page number not for citation purposes)
KEYWORDS
activity tracker; wearable electronic devices; fitness trackers; data mining; artificial intelligence; health; education; behavior change; physical activity; wearable devices; trackers; health education; sensor data

Introduction
Wearable sensors are increasingly employed in health interventions because of their ability to track participants’ physical activity (PA) in an unobtrusive, continuous, and precise manner under free-living conditions [1]. In the context of health promotion, sensor data are commonly used to objectively assess interventions by monitoring PA changes and progress toward compliance with public health PA guidelines [2].

The rich data captured by activity sensors contain information about the participants’ PA, potentially unlocking valuable insights into PA behaviors and patterns [3]. These insights can help to advance the understanding of how interventions affect PA behaviors and how behaviors change, thereby scaffolding the design of future interventions, and enhancing their outcomes, efficacy, and adherence.

In the last decade, a growing number of artificial intelligence and data mining models and techniques have been developed to detect and extract these latent PA patterns beyond the typical summaries of pre- and postintervention daily steps or time spent in various PA levels. In this systematic review, we aimed to describe the data mining models and techniques currently used to detect PA with a focus on behavior changes. We discuss their value, identify gaps or challenges, and highlight opportunities. The following research questions (RQs) guided this review:

RQ1: What are the current techniques used for mining PA sensor data to detect behavior changes in health education or health promotion contexts?
RQ1.1 What are the types of sensors used and what data are collected?
RQ1.2 How are data preprocessed?
RQ1.3 What features are used to detect behavior changes?
RQ1.4 What are the data mining models and techniques used to detect behavior changes?
RQ1.5 What are the interpretation of data mining models used for?

RQ2: What are the challenges and opportunities in mining PA sensor data for detecting PA behavior changes?

The RQ1 subquestions were established following the reasoning and order of the process of knowledge discovery in databases [4]. Figure 1 summarizes this process and maps each step with the relevant RQ1 subquestion.

Methods
Design
For this systematic review, we followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [5] and used the Rayyan QCRI web application [6] to manage the review process. We identified studies by searching the Association for Computing Machinery (ACM), IEEE Xplore, ProQuest, Scopus, Web of Science, Education Resources Information Center (ERIC), and Springer digital libraries. We also searched Google Scholar to identify grey literature and extracted the first 100 results. For this scholarly reference search, we used the following query: (education OR promotion OR “behaviour change”) AND (“data mining” OR “machine learning” OR “artificial intelligence”) AND (sensor OR accelerometer OR tracker OR wearable) AND “physical activity” AND health. All extracted scholarly references had been added to the database at the latest on the search day (May 28, 2021). The inclusion and exclusion criteria are presented in Textbox 1.

Figure 1. Knowledge discovery in database steps (in grey) and research question 1 (RQ1) subquestions (in blue).

Data Capture  →  Data Pre-processing  →  Data Transformation  →  Data Mining  →  Interpretation

RQ1.1  ↓  RQ1.2  ↓  RQ1.3  ↓  RQ1.4  ↓  RQ1.5
Textbox 1. Inclusion and exclusion criteria for article selection in the review.

**Inclusion criteria**
- Full-length articles
- Peer-reviewed articles in journals or conference papers
- Articles that used data mining techniques for data from physical activity (PA) wearable sensors
- Articles that included PA data
- Articles on applied health education/promotion or on behavior change scenarios
- Articles that used well-known data mining techniques such as classification, regression, clustering, association, and sequence algorithms, as well as specific algorithms to model PA data

**Exclusion criteria**
- Use of analytics without data mining
- Studies on animals (e.g., accelerometers on dogs)
- Self-quantification without a health education or health motivation component
- Dissertations and theses, due to lack of a peer review process
- Systematic reviews, reviews, and meta-analyses
- Health care applications without a health education or motivation for behavior change component
- Specific movement detection (abnormal gait, falls)
- Aid for sport training (e.g., maintaining heart rate, postures, specific movements)

**Search Outcome**
The number of references extracted from each electronic database is summarized in Table 1.

Following the PRISMA methodology, we retrieved 4388 references from the sources listed in Table 1. We then removed 415 duplicates, leaving 3973 unique references that were screened by reading their titles and abstracts. Using the inclusion/exclusion criteria (Textbox 1), we excluded 3688 references and selected 285 publications. After full-text reading, we excluded 266 references: 33 on activity recognition, 5 on data mining, 24 on systems, 31 on rehabilitation, 39 not on behavior changes, 54 without data mining, 51 not on health education/promotion, 13 not on PA, and 16 reviews. At the end of the selection process (summarized in Figure 2), we retained 19 references for this systematic review.

**Table 1.** Number of references extracted from each database.

<table>
<thead>
<tr>
<th>Database</th>
<th>Query result, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACM(^a)</td>
<td>584</td>
</tr>
<tr>
<td>IEEE Xplore</td>
<td>12</td>
</tr>
<tr>
<td>ProQuest</td>
<td>1678</td>
</tr>
<tr>
<td>Scopus</td>
<td>44</td>
</tr>
<tr>
<td>Web of Science</td>
<td>16</td>
</tr>
<tr>
<td>Scopus</td>
<td>2</td>
</tr>
<tr>
<td>Springer</td>
<td>1952</td>
</tr>
<tr>
<td>Google Scholar</td>
<td>100</td>
</tr>
</tbody>
</table>

\(^a\) ACM: Association for Computing Machinery.  
\(^b\) ERIC: Education Resources Information Center.
Results

Overview

The 19 included articles were published between 2013 and 2021. Their number per year increased from 1 in 2013 to 2 in 2017 and up to 5 in 2018. Subsequently, the number of publications decreased to a mean of 3 per year.
Table 2. Conference proceedings and journals in which the included articles were published (N=19).

<table>
<thead>
<tr>
<th>Conference or journal</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Medical and public health</strong></td>
<td></td>
</tr>
<tr>
<td>BMJ Open</td>
<td>Aguilera et al [7]</td>
</tr>
<tr>
<td>Public Health Nutrition</td>
<td>Lee et al [8]</td>
</tr>
<tr>
<td><strong>Medical and health informatics</strong></td>
<td></td>
</tr>
<tr>
<td>JMIR mHealth and uHealth</td>
<td>Zhou et al [9], Rabbi et al [10], Galy et al [11]</td>
</tr>
<tr>
<td>JMIR Public Health and Surveillance</td>
<td>Fukuoka et al [12]</td>
</tr>
<tr>
<td>Journal of Biomedical Informatics</td>
<td>Sprint et al [13]</td>
</tr>
<tr>
<td><strong>Human-computer interactions</strong></td>
<td></td>
</tr>
<tr>
<td>Proceedings of the ACM on Human-Computer Interaction</td>
<td>Zhu et al [14]</td>
</tr>
<tr>
<td>User Modeling and User-Adapted Interaction</td>
<td>Gasparetii et al [15]</td>
</tr>
<tr>
<td>Journal of Ambient Intelligence and Humanized Computing</td>
<td>Batool et al [16]</td>
</tr>
<tr>
<td>Multimedia Tools and Applications</td>
<td>Angelides et al [17]</td>
</tr>
<tr>
<td>Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization</td>
<td>Schäfer et al [18]</td>
</tr>
<tr>
<td><strong>Physical human behavior</strong></td>
<td></td>
</tr>
<tr>
<td>Journal of Behavioral Medicine</td>
<td>Forman et al [19]</td>
</tr>
<tr>
<td>Journal of Electromyography and Kinesiology</td>
<td>Hermens et al [20]</td>
</tr>
<tr>
<td><strong>Engineering and science</strong></td>
<td></td>
</tr>
<tr>
<td>Applied Sciences</td>
<td>Chen et al [21]</td>
</tr>
<tr>
<td>Sensors</td>
<td>Dijkhuis et al [22]</td>
</tr>
<tr>
<td>Springer Proceedings in Complexity</td>
<td>Mollee et al [23]</td>
</tr>
<tr>
<td>IEEE Access</td>
<td>Diaz et al [24]</td>
</tr>
<tr>
<td>International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems</td>
<td>Mollee and Klein [25]</td>
</tr>
</tbody>
</table>

Sensor Types and Data Capture

The characteristics of the sensors (eg, number and type) used to capture PA behaviors and of the collected raw data are summarized in Table 3.

The length of data recordings varied between 4 days and 1 year, with a median of 70 days. Recording lasted ≤7 days in two studies, between 3 and 5 weeks in six studies, between 10 and 16 weeks in eight studies, and ≥6 months in three studies.

The number of participants varied between 10 and 11,615, with <30 in five studies, between 30 and 299 in 10 studies, and ≥300 participants in four studies.

All included studies used accelerometer sensors. We could categorize these devices into three groups: (1) commercial wrist-worn wearable accelerometers that are consumer-grade devices with a sample rate between 30 Hz and 60 Hz, such as Fitbits [13,14,19,22,25], Samsung Gear [17], and Nokia [15]; (2) smartphone accelerometers with a sample rate usually set to 50 Hz and up to 100 Hz, in which data were collected via an app installed in the smartphone [7,9,10,16,18]; and (3) scientifically validated wearable accelerometers with a sample rate up to 100 Hz, such as ActiGraph [8], GENEActiv [11,24], and other devices developed for health care [12,20].

In 7 out of the 19 (37%) selected studies, accelerometers were used with other sensors such as GPS tracking [10,16,17], compass position tracking [17,20], heart rate trackers [17,21], and smart scales [15,19].

The recorded raw data varied in function of the sensor characteristics, including sampling frequency, accuracy, and axis number. Moreover, other sensor features such as battery duration and storage capacity affected the recording length. For instance, a long battery life and high storage capacity enable longer recording without interruptions. Table 4 summarizes the number of participants and data recording duration for the included studies.
<table>
<thead>
<tr>
<th>Sensor type</th>
<th>Device and model</th>
<th>Raw data</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accelerometer</td>
<td>ActiGraph GT1M uniaxial</td>
<td>Uniaxial accelerometry</td>
<td>Lee et al [8]</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>GENEActiv triaxial accelerometer</td>
<td>Gravity-subtracted signal vector magnitudes (SVMgs) per second</td>
<td>Galy et al [11], Diaz et al [24]</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>Generic device from the mobile phone</td>
<td>Acceleration (sample rate not specified)</td>
<td>Aguilera et al [7], Zhou et al [9]</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>Triaxial accelerometer (HJA-350JT, Active Style Pro, Omron Healthcare Co, Ltd)</td>
<td>Triaxial acceleration (6 Hz)</td>
<td>Fukuoka et al [12]</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>Fitbit</td>
<td>Triaxial acceleration (sample rate not specified)</td>
<td>Zhu et al [14]</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>Fitbit Flex</td>
<td>Triaxial acceleration (sample rate not specified)</td>
<td>Dijkhuis et al [22]</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>Fitbit Charge HR and Fitbit Flex</td>
<td>Triaxial acceleration (sample rate not specified)</td>
<td>Sprint et al [13]</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>Fitbit One</td>
<td>Triaxial acceleration (sample rate not specified)</td>
<td>Mollee and Klein [25]</td>
</tr>
<tr>
<td>Accelerometer and heart rate monitor</td>
<td>Not specified</td>
<td>Not specified</td>
<td>Mollee et al [23]</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>Smartphone and Actigraph (GT3X model)</td>
<td>Triaxial acceleration (sample rate not specified)</td>
<td>Schäfer et al [18]</td>
</tr>
<tr>
<td>Accelerometer, GPS, self-log PA, and food</td>
<td>Mix of devices and models</td>
<td>Accelerometry, heart rate monitor, PA(^a) information, and user information (sample rate not specified)</td>
<td>Chen et al [21]</td>
</tr>
<tr>
<td>Activity tracker, smart scale, and smartphone (what they ate and drank in the Fitbit app)</td>
<td>Smartphone</td>
<td>Smartphone accelerometry, GPS data, PA and food logs with sample rate specified</td>
<td>Rabbi et al [10]</td>
</tr>
<tr>
<td>Accelerometer, gyroscope, and magnetic compass</td>
<td>Fitbit Flex 2 activity tracker, Yunmai smart scale, smartphone</td>
<td>Accelerometry, weight and food logs (sample rate not specified)</td>
<td>Forman et al [19]</td>
</tr>
<tr>
<td>Accelerometer and GPS</td>
<td>ProMove-3D (developed by Inertia Technology)</td>
<td>Accelerometry (sample rate not specified)</td>
<td>Hermens et al [20]</td>
</tr>
<tr>
<td>Triaxial accelerometer, heart rate monitor, GPS, 3-axis gyroscope, digital compass, altimeter, light sensor</td>
<td>Smartphone</td>
<td>Accelerometry and GPS (sample rate not specified)</td>
<td>Batool et al [16]</td>
</tr>
<tr>
<td>Accelerometer, heart rate, and smart scale</td>
<td>Samsung Gear Fit and Fitbit Surge</td>
<td>Accelerometer, heart rate data, GPS, 3-axis gyroscopes, digital compass, altimeter, light sensor (sample rate not specified)</td>
<td>Angelides et al [17]</td>
</tr>
<tr>
<td>Accelerometer, heart rate, and smart scale</td>
<td>Nokia; models not specified</td>
<td>Accelerometer, heart rate data, and smart scale (sample rate not specified)</td>
<td>Gasparetti et al [15]</td>
</tr>
</tbody>
</table>

\(^a\)PA: physical activity.
Table 4. Length of data recording and number of participants among the included studies.

<table>
<thead>
<tr>
<th>Length of recording</th>
<th>Participants, n</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 7 days</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 days</td>
<td>1714</td>
<td>Lee et al [8]</td>
</tr>
<tr>
<td>7 days</td>
<td>215 (women)</td>
<td>Fukuoka et al [12]</td>
</tr>
<tr>
<td>1 to 5 weeks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 weeks</td>
<td>17</td>
<td>Rabbi et al [10]</td>
</tr>
<tr>
<td>3 weeks</td>
<td>48</td>
<td>Zhu et al [14]</td>
</tr>
<tr>
<td>1 month</td>
<td>14</td>
<td>Angelides et al [17]</td>
</tr>
<tr>
<td>4 weeks</td>
<td>24 (adolescents)</td>
<td>Galy et al [11]</td>
</tr>
<tr>
<td>4 weeks</td>
<td>74 (children)</td>
<td>Schäfer et al [18]</td>
</tr>
<tr>
<td>5 weeks</td>
<td>87 (children)</td>
<td>Diaz et al [24]</td>
</tr>
<tr>
<td>6 to 20 weeks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 weeks</td>
<td>11</td>
<td>Schäfer et al [18]</td>
</tr>
<tr>
<td>10 weeks</td>
<td>64</td>
<td>Zhou et al [9]</td>
</tr>
<tr>
<td>3 months</td>
<td>10</td>
<td>Hermens et al [20]</td>
</tr>
<tr>
<td>12 weeks</td>
<td>48</td>
<td>Dijkhuis et al [22]</td>
</tr>
<tr>
<td>12 weeks</td>
<td>108</td>
<td>Mollee and Klein [25]</td>
</tr>
<tr>
<td>3 months</td>
<td>269</td>
<td>Chen et al [21]</td>
</tr>
<tr>
<td>12 weeks</td>
<td>2472</td>
<td>Mollee et al [23]</td>
</tr>
<tr>
<td>16 weeks</td>
<td>52</td>
<td>Forman et al [19]</td>
</tr>
<tr>
<td>21 weeks to 1 year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 months</td>
<td>276</td>
<td>Aguilera et al [7]</td>
</tr>
<tr>
<td>6 months</td>
<td>500</td>
<td>Batool et al [16]</td>
</tr>
<tr>
<td>1 year</td>
<td>11,615</td>
<td>Gasparetti et al [15]</td>
</tr>
</tbody>
</table>

Data Preprocessing

Raw data extracted from sensors need to be transformed into variables that will contribute to generating the input features for data mining models to detect PA behavior changes. Table 5 provides a summary of the initial transformation and the resulting preprocessed data.

The preprocessing of the raw data from sensors was carried out in two ways. The first approach was to use proprietary programs to transform the sensors’ data directly into the resulting preprocessed data, without specifying whether there was an initial preprocessing stage such as that used to generate steps, metabolic equivalents (METs), calories, heart rate, or exercise characteristics (type, duration, distance, or frequency). The second approach was to produce intermediate data that were then transformed in the resulting preprocessed data using a custom preprocessing tool. For instance, to generate PA levels (PALs), raw data were first transformed into MET, activity classes, or signal vector magnitudes.

The resulting preprocessed data were mainly activity characteristics (step count, PAL, integrals of the moduli of acceleration, activity types, duration, distance travelled, and frequency) and energy expenditure (MET and calories). Step count from smartphones and commercial wrist-worn devices was the most frequent, followed by PAL from research-grade devices.

The resulting preprocessed data were aggregated at different time levels (Table 6). Day and minutes were the most frequent time levels of aggregation. Generally, PAL and MET were aggregated per minute. Calories and step counts were calculated per day.
Table 5. Summary of data preprocessing variables.

<table>
<thead>
<tr>
<th>Resulting/initial preprocessing</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steps: unknown (proprietary program)</td>
<td>[7,9,13-15,17,19,22,25]</td>
</tr>
<tr>
<td>Metabolic equivalents: unknown (proprietary program)</td>
<td>[21]</td>
</tr>
<tr>
<td>Calories: unknown (proprietary program)</td>
<td>[10,17,19]</td>
</tr>
<tr>
<td>Exercise characteristics: unknown (proprietary program)(^a)</td>
<td>[10,17,21]</td>
</tr>
<tr>
<td>Sleeping time: unknown (proprietary program)</td>
<td>[15,17]</td>
</tr>
<tr>
<td>Weight: unknown (proprietary program)</td>
<td>[15,19]</td>
</tr>
<tr>
<td>Heart rate: unknown (proprietary program)</td>
<td>[17,21]</td>
</tr>
</tbody>
</table>

**Physical activity (PA) levels**

- Signal vector magnitudes                                           | [11,24]                                                                   |
- PA counts                                                          | [8]                                                                       |
- Metabolic equivalents                                              | [12]                                                                      |
- Activity classes                                                   | [18]                                                                      |
- Not specified                                                      | [23]                                                                      |
- Integrals of the moduli of acceleration signals                    | [20]                                                                      |
- Actual activity level; not specified\(^b\)                          | [16]                                                                      |

\(^a\)Type, duration, distance, frequency.

\(^b\)Definition of activity level was not specified.

Table 6. Aggregation level of the resulting preprocessed data.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Month</th>
<th>Week</th>
<th>Day</th>
<th>Hour</th>
<th>Minute</th>
<th>Seconds</th>
<th>Not specified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angelides et al [17]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhou et al [9]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aguilera et al [7]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhu et al [14]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mollee and Klein [25]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forman et al [19]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gasparetti et al [15]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chen et al [21]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dijkhuis et al [22]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lee et al [8]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fukuoka et al [12]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sprint et al [13]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schäfer et al [18]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diaz et al [24]</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Galy et al [11]</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Mollee et al [23]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hermens et al [20]</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Rabbi et al [10]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Batool et al [16]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Features Used to Detect and Extract Behavior Changes

The features of the data mining models were mostly generated from the sensors’ preprocessed data and, in some cases, from other sources (nonsensor data). Table 7 provides the features categorized with respect to the function of their source: accelerometers, other sensors, and nonsensor devices.

Most of the included articles used descriptive statistics to present the preprocessed data as features, for instance total number of steps per day [9,11,14,17,25], mean number of steps per day [17], or PA count per hour [8]. Other studies created windows or segments of time to calculate PA characteristics, including segments of steps or sleep [15] and PA bouts [13,24]. Other articles used the preprocessed data to calculate the participants’ step achievements such as whether they reached their step goal [9,11,19,23]. Zhu et al [13] used more complex features such as the ratio between the most active and least active period or the circadian rhythm strength.

In addition to the features derived from sensors, others were created from measurements carried out during the intervention by scientists, such as the number of days that a person participated in the intervention [19] and anthropometric [7,21] or psychological [14,16,25] characteristics. Data were collected through surveys/questionnaires or interviews with participants.
Table 7. Features used for data mining to detect behavior changes.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Features derived from accelerometers</th>
<th>Features derived from other sensors</th>
<th>Features derived from nonsensor devices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aguilera et al [7]</td>
<td>Number of minutes of activity in the last day, cumulative number of minutes of activity this week, fraction of activity goal, fraction versus expected activity goal at this point in the week</td>
<td>Number of days since each feedback message was sent</td>
<td>Age, gender, language, 8-item Patient Health Questionnaire (depression) score</td>
</tr>
<tr>
<td>Hermens et al [20]</td>
<td>Not specified</td>
<td>Monthly mean exercise and resting heart rate</td>
<td>Not specified</td>
</tr>
<tr>
<td>Chen et al [21]</td>
<td>Monthly mean metabolic equivalent of task, effective exercise time, type, frequency</td>
<td>Not specified</td>
<td>Gender, height, weight, age</td>
</tr>
<tr>
<td>Forman et al [19]</td>
<td>Days where PA(^a) goal is met</td>
<td>Sum of days with self-monitored weight, days where calorie goal is met, weight loss in pounds</td>
<td>Number of days in the intervention period</td>
</tr>
<tr>
<td>Gasparetti et al [15]</td>
<td>Consecutive daily segments of steps, consecutive daily segments of sleep</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Batool et al [16]</td>
<td>Actual activity level</td>
<td>—</td>
<td>Desired activity level, intention (attitude, subjective norms, perceived behavioral control), habit, and 16 demographic features (eg, age, gender, marital status)</td>
</tr>
<tr>
<td>Dijkhuis et al [22]</td>
<td>Hour of the workday, number of steps for that hour, number of steps in the past hour, total number of steps up to that hour, mean number of steps of workdays</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Rabbi et al [10]</td>
<td>PA frequency and calories</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Zhou et al [9]</td>
<td>Daily steps and goal</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Angelides et al [17]</td>
<td>Total and mean hourly, daily, weekly, and monthly sleep duration; sleep calories; exercise duration; exercise distance; exercise calories; step count; step distance; step calories; BMI; and basal metabolic rate</td>
<td>—</td>
<td>Height (cm), weight (kg), age, gender</td>
</tr>
<tr>
<td>Diaz et al [24]</td>
<td>Hourly and daily frequency, and mean time spent in moderate to vigorous PA bouts of at least 3, 10, and 30 seconds, and in sedentary bouts of at least 60, 120, and 300 seconds</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Galy et al [11]</td>
<td>Total daily time spent in light/moderate/vigorous PA, total daily number of steps, and a binary goal achievement feature</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Fukuoka et al [12]</td>
<td>Mean metabolic equivalent of tasks per minute, mean moderate-to-vigorous PA per minute</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Lee et al [8]</td>
<td>24-hour mean PA count on weekdays and 24-hour mean PA count on weekends</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Sprint et al [13]</td>
<td>Steps, PAL(^c) and bouts count, mean, percentages, ratios and SD. Circadian rhythm time-series statistics and texture features from an image-processing technique</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Mollee et al [23]</td>
<td>Impact of online community (sharing my PAL with peers), target PAL and goal achievement</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Schäfer [18]</td>
<td>PAL per minute</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Reference</td>
<td>Features derived from accelerometers</td>
<td>Features derived from other sensors</td>
<td>Features derived from nonsensor devices</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------------------------------------</td>
<td>------------------------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>Zhu et al [14]</td>
<td>Daily steps</td>
<td>Motivation to exercise (Likert scale)</td>
<td>Iowa-Netherlands Comparison Orientation Measure-23 (INCOM-23) for social comparison (psychometrics)</td>
</tr>
</tbody>
</table>

*aPA: physical activity.
*bNot applicable.
*cPAL: physical activity level.

### Data Mining

#### Algorithm Overview

Table 8 summarizes the data mining methods and specific algorithms used in the selected articles.

Clustering was the most used method, particularly the K-means algorithm. Indeed, in health interventions, the PA performed by each participant varies in duration, form, and intensity. Therefore, an algorithm that clusters PA behaviors is required to analyze them. The unsupervised K-means algorithm is suitable for this task. Indeed, due to its simplicity and ease of use, this is one of the most popular options for data mining [26]. Decision-making algorithms and classifiers were the second most used methods. Both rely on supervised algorithms that use PA characteristics as a method for predicting when and/or what information must be delivered to individual participants for increasing their PA. Other algorithms were also tested to extract PA behaviors, such as social cognitive and contagion models, PA windows permutations, and recommendation algorithms.

### Table 8. Data mining methods and algorithms.

<table>
<thead>
<tr>
<th>Data mining method and algorithm</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classifiers</strong></td>
<td></td>
</tr>
<tr>
<td>K-nearest neighbor and support vector machine</td>
<td>[20]</td>
</tr>
<tr>
<td>Random forest</td>
<td>[22]</td>
</tr>
<tr>
<td>Random forest and weighted score</td>
<td>[18]</td>
</tr>
<tr>
<td>Shallow neural networks</td>
<td>[16]</td>
</tr>
<tr>
<td><strong>Clustering techniques</strong></td>
<td></td>
</tr>
<tr>
<td>K-means</td>
<td>[8,11,12,24]</td>
</tr>
<tr>
<td>Agglomerative</td>
<td>[21]</td>
</tr>
<tr>
<td>Partitioning around medoids and reinforcement learning</td>
<td>[15]</td>
</tr>
<tr>
<td><strong>Decision-making algorithms</strong></td>
<td></td>
</tr>
<tr>
<td>Multiarmed bandit</td>
<td>[10]</td>
</tr>
<tr>
<td>Multiarmed bandit upper confidence bound</td>
<td>[19]</td>
</tr>
<tr>
<td>Reinforcement learning multiarmed bandit</td>
<td>[7,9]</td>
</tr>
<tr>
<td>Behavioral analytics algorithm</td>
<td></td>
</tr>
<tr>
<td>MAB&lt;sup&gt;a&lt;/sup&gt;</td>
<td>[14]</td>
</tr>
<tr>
<td>Social cognitive model for predicting exercise behavior change</td>
<td>[25]</td>
</tr>
<tr>
<td>Social contagion model combined with a linear model</td>
<td>[23]</td>
</tr>
<tr>
<td>Physical activity change detection: small window permutation-based change detection in activity routine</td>
<td>[13]</td>
</tr>
<tr>
<td>Recommendation: genetic algorithms and Pareto optimality</td>
<td>[17]</td>
</tr>
</tbody>
</table>

<sup>a</sup>MAB: multiarmed bandit.

#### Classifiers

Hermens et al [20] used a k-nearest neighbor model and a support vector machine to determine whether a specific time of the day was suitable for sending a motivational message to optimize adherence to the intervention. Dijkhuis et al [22] used a tree and tree-based ensemble algorithm classifiers to predict whether users will achieve their daily PA goal. On the basis of this prediction, a personalized PA coaching program was proposed. Forman et al [18] developed gamified personalized...
feedback using a score model depending on the PA change detected from accelerometer data. Batool et al [16] predicted the likelihood that the PA level of a given patient was too low. They also predicted which patients were at higher risk of not adhering to the prescribed therapy to optimize their PA.

**Clustering Techniques**

Lee et al [8] grouped participants in two clusters on the basis of their step counts (one more active than the other), and analyzed them to better understand these PA patterns. Diaz et al [24] used a clustering-based approach for a more insightful analysis of the participants’ PA behavior and of the nature of the PA behavior changes, if present. Galy et al [11] clustered PA levels and daily step goal achievement to assess the adherence to a health program. Fukuoka et al [12] identified PA clusters to analyze and compare sociodemographic features and cardiometabolic risks among participants belonging to these clusters. Chen et al [21] clustered the participants’ PA, and then established a system to adapt the exercise program for the next week as a function of the individual PA behavior change. Gasparetti et al [15] clustered the participants’ PA to generate groups of habits recommended by a system to the participants with the objective of changing their PA to obtain weight loss effects.

**Decision-Making Algorithms**

Rabbi et al [10] generated personalized suggestions in which users were asked to continue, avoid, or make small changes to their existing PA behaviors in order to help them reach their PA goals. Forman et al [19] developed an algorithm that could personalize and optimize the PAL during the intervention as a function of the amount of PA performed. Aguilera et al [7] generated personalized messages for participants in the intervention to increase their PA and consequently the intervention effectiveness. Zhou et al [9] adapted the step goal settings of the intervention depending on the PA behavior change. Zhu et al [14] personalized social comparison among participants to motivate them toward improving their PA behavior.

**Social Cognitive Model**

Mollee and Klein [25] developed a model that simulates changes in PALs over 2 to 12 weeks to optimize the participants’ health outcome.

**Social Contagion Model**

Mollee et al [23] used a social contagion model to explain the PAL dynamics in a community.

**PA Windows Permutations**

Sprint et al [13] proposed a window-based algorithm to detect changes in segments of users’ PA behavior to motivate progress toward their goals.

**Recommendation Algorithms**

Angelides et al [17] used genetic algorithms and Pareto optimality to compare the participants’ and peer community’s data to help participants interpret the PA data and to generate personal lifestyle improvement recommendations.

**Interpretation of the Data Mining Models**

**Overview of Models**

The resulting data mining models detecting PA behavior changes were used for several purposes, as summarized in Table 9 and below.

Table 9. Main uses of the resulting data mining models.

<table>
<thead>
<tr>
<th>Main use</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personalized feedback</td>
<td>[7,10,15,16,18,20]</td>
</tr>
<tr>
<td>Personalized program</td>
<td>[9,19,21,22]</td>
</tr>
<tr>
<td>Support for self-reflection</td>
<td>[17]</td>
</tr>
<tr>
<td>Cohort analysis of the intervention impact on PA*</td>
<td>[8,11-13,24]</td>
</tr>
<tr>
<td>Analysis of the social component effects on PA</td>
<td>[14,23,25]</td>
</tr>
</tbody>
</table>

*PA: physical activity.

**Personalized Feedback**

The PA behavior changes extracted from participants’ data were used to promote PA by creating and sending personalized messages that reported the behaviors and gave suggestions for achieving the previously established PA goals. For instance, Aguilera et al [7] built a system that detects the participants’ PA behavior changes and generates personalized daily text messages with custom timing, frequency, and feedback about their step count/goal and motivational content. Hermens et al [20] built a system that chooses the best suitable time to send a message with personalized intention, content, and representation. Schäfer et al [18] created an app with gamified feedback where different avatars are awarded based on the participant’s daily PA behavior. Gasparetti et al [15] suggested personalized PA patterns based on the participants’ PA patterns. Batool et al [16] detected the participants’ PA behavior while commuting and suggested how to increase it. Rabbi et al [10] generated personalized simple PA suggestions (continue, avoid, or make small changes).

**Personalized Programs**

The PA intervention program and objectives are adapted to each participant’s needs. For instance, Chen et al [21] created a guided exercise prescription system that adapts as the participants’ PA behavior changes. Similarly, Forman et al [19] changed the participant’s exercise intensity suggestion depending on their PA behavior achievements. On the basis of

**Support for Self-Reflection**

Algorithms can help participants to interpret their PA behavior changes. For example, Angelides et al [17] used an algorithm to assist in the interpretation of the participant’s PA data by comparing them with those of the peer community and to generate personalized recommendations to achieve their daily goals.

**Cohort Analysis of the Intervention Impact on PA**

These algorithms detect PA behavior changes in participants that allow analyzing the intervention impact. For example, Fukuoka et al [12] determined PA patterns in women throughout the day that could help to develop more personalized interventions and guidelines. Diaz et al [24] analyzed the changes in PA behavior (bouts and frequency) during an intervention. Galy et al [11] tracked the participants’ adherence to the international recommendations during an intervention. Lee et al [8] identified PA patterns associated with specific subgroups of people who participated in an intervention. Sprint et al [13] analyzed the participants’ PA changes during an intervention by comparing multiple time windows.

**Analysis of the Social Component Effect on PA**

These algorithms analyze the psychosocial influences on the participants’ PA. For example, Mollee et al [23] analyzed the PA dynamics in a community using a social contagion model. Mollee and Klein [25] analyzed the PA dynamics in a networked community using social cognitive theories, and Zhu et al [14] personalized social comparison during an intervention to increase the participants’ PA.

The main uses can be classified in two groups. The first group, composed of 11 out of the 19 (58%) selected studies, aimed to generate personalized feedback/PA programs to scaffold and support PA behavior changes among participants. Indeed, researchers seem inclined to generate greater personalization because it increases the intervention efficiency, effectiveness, enjoyment, and reliability [27]. The second group, composed of 8 out of the 19 (42%) selected studies, sought to analyze the impact of interventions on the participants’ PA. Specifically, these studies analyzed the intervention impact on PA at the cohort level to assess health education interventions, and analyzed participants’ PA to show their behaviors and help to understand them. The main objective of both groups was to explore how PA behavior patterns relate to the intervention effectiveness, which can add new evidence on how to create more effective interventions [28].

**Discussion**

**Principal Findings**

**Summary**

We found 19 articles about data mining models and techniques to detect PA behavior changes in health education or promotion studies, and their number has progressively increased over time. We here discuss the principal findings, identify opportunities and challenges for future research directions, and present the limitations of this systematic review. The Discussion is structured according to the RQs as a guide.

**Opportunities and Challenges**

**Sensor Types and Data Capture**

All selected studies used accelerometer sensors to capture PA behaviors. While 7 out of the 19 (37%) studies utilized accelerometers exclusively, the rest employed them with other sensors. Nonaccelerometer sensors capture additional information that may be relevant to PA (such as work/school schedule, itineraries, and sleep patterns [29]) and could yield auxiliary features for the data mining models. For instance, GPS sensors provide the number of kilometers and location of PA performed.

The median number of participants in the selected studies was 74, and participants were mainly young or middle-aged adults. This low number of participants and the skew toward adults may have generated biased data mining models that can detect and find behavior changes only in a specific population. Different population groups behave differently and should be studied independently. For instance, PA behaviors are different in children and adults [2]. Some of the studies focused on groups with specific PA behaviors, such as children [18,24], adolescents [11], and women [12]. However, some population groups with distinctive PA patterns, such as pregnant women [30] and people with health conditions or disabilities [31], may need custom detection models.

In 15 out of the 19 (79%) included studies, data were recorded for less than 3 months. Therefore, the current methods for detecting PA behavior changes have been developed mostly for capturing short-term patterns, making the conclusions valid only for short periods. To detect medium- and long-term PA behavior changes, studies with more extended recording periods are needed, such as the study by Gasparetti et al [15] based on data collected during 1 year. Moreover, new methods to detect extended (eg, annual or seasonal) PA patterns are required to study how the participants’ behavior and habits change over time. An increase in the participants’ number and recording length will lead to new challenges related to big data analysis, such as efficient data management and data mining processing speeds.

**Data Preprocessing**

Many of the selected studies used commercial accelerometers that allow only the retrieval of aggregated preprocessed data using proprietary software (ie, number of steps per minute), without being transparent on how data were preprocessed (ie, how steps were calculated from the accelerometry data). This data preprocessing black box makes it impossible to determine the quality of the captured PA data and makes the data mining results scientifically irreproducible. Conversely, in studies that used medical-grade accelerometers, the accelerometry data were explained in detail and the preprocessing steps were documented and referenced.
We found a lack of standard procedures for data preprocessing that made it challenging to compare the study results and conclusions. Indeed, if data are not preprocessed correctly, this could cause the transfer of incorrect information to the features and then to the data mining models. This could lead to the creation of inaccurate models, thus limiting the study validity. Data cleaning is a good example of this issue. Indeed, the best procedure to eliminate the nonwearing time remains unclear along with the impact on the accuracy of the resulting models. If nonwearing time is poorly removed, features can generate a PA underestimation by recognizing nonwearing time as sedentary behavior when it is not. Moreover, if sensor data concerning changes in accelerations while commuting by car or bus are not completely removed, they will be erroneously classified as steps, thereby overestimating PA in the model and in the conclusions. Similarly, sedentary activities could be overestimated if sleep time is not correctly removed.

Most of the selected studies aggregated information by day or minute. Although data aggregation is useful when comparing general features of PA behaviors, such as daily steps, this procedure may overlook subtle behavioral changes that can be crucial for detecting major PA behavior changes. For instance, if a person who walks every morning decides to change their behavior and starts to walk at night, the sum of daily steps will be the same, but this new behavior will not be detected. Conversely, it could be detected if the aggregation level is changed to the hour. To detect these and other subtle behavior changes, PA should be analyzed simultaneously at different aggregation levels, and new time frames should be created to match daily habits and behaviors, such as periods of the day (eg, morning, afternoon) or participants’ office hours.

Features Used to Detect and Extract Behavior Changes

Most of the preprocessed data were transformed into features that are simple descriptive statistics, such as the total time spent at a specific PAL or the mean number of steps. These features are valuable to detect behavior changes, but they mainly capture the PA intensity and the PA presence or absence. Yet, PA has more valuable characteristics that vary during PA behavior changes and that can help to detect such behavior changes, such as the length of PAL bouts or the amount of time spent doing PA. These PA characteristics can be extracted from current sensor data. For instance, Galy et al [11] explored different moderate-to-vigorous PA bout lengths and Sprint et al [13] assessed the circadian rhythm. International PA guidelines can serve as inspiration to identify new PA features. For instance, according to World Health Organization recommendations, adults should perform muscle-strengthening activities (involving all major muscle groups) at moderate or higher intensity at least twice per week [2]. This calls for the creation of features that capture the muscle activity type, intensity, and frequency. Moreover, most of the included studies used only PA-derived features to detect behavior changes, and did not consider relevant non-PA data associated with PA changes, such as the participants’ weight and quality of sleep. Some studies captured non-PA data, but they did not use them to detect PA changes. For instance, Rabbi et al [10] used only PA-derived data (PA frequency and calories burned) to detect behavior changes, although they also recorded the participants’ food intake, thus excluding their caloric intake that is closely related to weight and the amount of PA participants are likely to perform.

The use of simple descriptive statistics as features and the exclusion of non-PA data associated with behavior changes indicate that sensor data were underexploited and that the features used to detect PA behavior changes are still underdeveloped. Including new PA characteristics and new non-PA features could help to better understand the nature of PA changes and how these features influence PA behavior changes, ultimately increasing the model detection accuracy.

Data Mining Methods and Techniques

Most studies used off-the-shelf classifiers, clusters, and decision-making algorithms to detect PA behavior changes. We expected to find tailor-made algorithms because in health education settings, it is important to find specific PA patterns in participants of different classes who follow learning modules with different contents and with different PA goals. Moreover, we noticed that most authors did not explain how they chose the algorithms and did not specify the efficiency and accuracy of the models used for detecting PA behavior changes, raising uncertainty about how good they are at this task. This suggests that more efficient and accurate algorithms could be created and calls for more transparency in the algorithm choice process. Therefore, authors should explicitly describe the steps and methodology of new algorithms, and share their source codes to be scrutinized and to compare their detection accuracy. The creation of open accelerometry databases is also needed to enable benchmarking.

Interpretation of the Resulting Data Mining Models

The main uses of the data mining models focused on personalization, support for self-reflection, and analysis of PA behaviors. Model interpretation focused on generating personalization and support for promoting behavior changes. Personalized feedback and intervention programs were based mostly on the participants’ PA data. The inclusion of additional information that may influence behavior changes (eg, contexts, schedules, social constraints, motivation, and weather) would allow for better interpretation and use of the detected behavior changes. Systems could exploit these additional data to improve the feedback delivery time and content, with positive effects on the effectiveness of health education programs and interventions. For instance, with the current models, a participant could receive an automatized personalized behavior change message that suggests taking a short walk, although it is snowing outside. This would decrease the likelihood of following the suggestion. However, if the system could be aware of the weather, the participant would receive this suggestion only after the weather conditions have improved, or a different suggestion that is more likely to trigger a behavior change at that point in time. Moreover, as the models relied mainly on PA features to model and interpret the behavior changes, only the physical dimension of the learning process in health education was incorporated in the models and their interpretation, leaving aside the knowledge dimension of the learning process. Learning management systems and intelligent tutoring systems already capture the knowledge dimension. Their integration would help to understand, in a comprehensive way, how participants learn,
and would enable the real-time monitoring of how PA behavior changes align with the intervention purpose. This would allow adapting each participant’s content and learning objectives in real time, thereby improving instructions and learning, ultimately increasing the program or intervention effectiveness.

Most of the included studies generated complex output models that require detailed knowledge of how they were created to interpret the resulting patterns, making them difficult to understand for health scientists and any other scientist not familiar with machine learning. This is a common problem in interdisciplinary teams; however, an effort can be made to create more readable, intuitive, and easy-to-understand algorithms and methods, a goal that exists in related machine learning areas such as explainable artificial intelligence [32].

**Limitations**

Studies on wearable machine learning devices to detect changes in PA in health education have only started to be published in the last decade. As research is advancing, keywords are changing and new terms are created. Although we used a wide range of keywords in our query to include sensors, PA, and health education, we may have left some keywords out, and thus we may have missed some references. This may also have affected the initial reference screening process by title and abstract. We minimized this issue by testing several queries before starting our systematic review until we found the one we ultimately used. Another possible limitation in our search is that we might have omitted references listed only in other peer-reviewed databases (we searched only the most popular databases in engineering and computer science), such as medical databases (ie, PubMed). We mitigated this risk by including grey literature in our systematic review (see the Methods section).

Regarding the research subquestions and the review structure, we created research subquestions in line with the usual data mining process steps, but we certainly left some topics unaddressed. For instance, we did not address ethics, privacy, and security issues, or how data are filtered during preprocessing (eg, sleeping time or sensor nonuse). Although these are common substeps during the data mining process and including them would have made this systematic review more comprehensive, we preferred to limit this review only to the critical steps.

**Conclusions**

In the last 10 years, different methods have been developed to detect behavior changes in health education or health promotion contexts. These methods have been tested in small populations, are based on short data-recording periods, and rely mainly on accelerometry data. Incorporating information that is complementary to the participants’ PA data would allow for creating more precise detection models, better interpreting these models, and understanding how participants learn and what triggers new behaviors. Exploring other data aggregation levels, in addition to days and minutes, could help to detect more subtle and long-term behavior changes. Fully describing the data preprocessing methods and the efficiency and accuracy of the behavior change detection models would help to better understand, scrutinize, and compare studies. Detection models were mainly used to generate personalized feedback and to provide support for promoting or maintaining behavior changes, but did not integrate the knowledge dimension of the learning process. Adding the knowledge dimension and creating easier-to-understand models could facilitate the interpretation of participants’ behavior changes in a more comprehensive way, opening the way toward better and deeper analyses and personalization.

**Acknowledgments**

CD acknowledges the support of the Universidad Adolfo Ibáñez and Comisión Nacional de Investigación Científica y Tecnológica/Chilean National Commission for Scientific and Technological Research (CONICYT) “Becas Chile” Doctoral Fellowship program (grant number 72200111).

**Conflicts of Interest**

None declared.

**References**


**Abbreviations**

**ACM:** Association for Computing Machinery  
**ERIC:** Education Resources Information Center  
**MET:** metabolic equivalent  
**PA:** physical activity  
**PAL:** physical activity level  
**PRISMA:** Preferred Reporting Items for Systematic Reviews and Meta-Analyses  
**RQ:** research question
Methods Used in the Development of Common Data Models for Health Data: Scoping Review

Najia Ahmadi¹, MSc; Michele Zoch¹, Dipl-Wi-Inf; Patricia Kelbert², MSc; Richard Noll³, MSc; Jannik Schaal³, Dr rer med; Markus Wolfien¹, Dr-Ing; Martin Sedlmayr¹, Prof Dr

¹Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany
² Fraunhofer Institute for Experimental Software Engineering IESE, Kaiserslautern, Germany
³ Institute of Medical Informatics, Goethe University Frankfurt, University Hospital, Frankfurt, Germany
⁴Center for Scalable Data Analytics and Artificial Intelligence, Dresden/Leipzig, Germany

Corresponding Author:
Najia Ahmadi, MSc
Institute for Medical Informatics and Biometry
Carl Gustav Carus Faculty of Medicine
Technische Universität Dresden
Fetscherstr 74
Dresden, 01307
Germany
Phone: 49 351458 87 7704
Email: najia.ahmadi@tu-dresden.de

Abstract

Background: Common data models (CDMs) are essential tools for data harmonization, which can lead to significant improvements in the health domain. CDMs unite data from disparate sources and ease collaborations across institutions, resulting in the generation of large standardized data repositories across different entities. An overview of existing CDMs and methods used to develop these data sets may assist in the development process of future models for the health domain, such as for decision support systems.

Objective: This scoping review investigates methods used in the development of CDMs for health data. We aim to provide a broad overview of approaches and guidelines that are used in the development of CDMs (ie, common data elements or common data sets) for different health domains on an international level.

Methods: This scoping review followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist. We conducted the literature search in prominent databases, namely, PubMed, Web of Science, Science Direct, and Scopus, starting from January 2000 until March 2022. We identified and screened 1309 articles. The included articles were evaluated based on the type of adopted method, which was used in the conception, users’ needs collection, implementation, and evaluation phases of CDMs, and whether stakeholders (such as medical experts, patients’ representatives, and IT staff) were involved during the process. Moreover, the models were grouped into iterative or linear types based on the imperativeness of the stages during development.

Results: We finally identified 59 articles that fit our eligibility criteria. Of these articles, 45 specifically focused on common medical conditions, 10 focused on rare medical conditions, and the remaining 4 focused on both conditions. The development process usually involved stakeholders but in different ways (eg, working group meetings, Delphi approaches, interviews, and questionnaires). Twenty-two models followed an iterative process.

Conclusions: The included articles showed the diversity of methods used to develop a CDM in different domains of health. We highlight the need for more specialized CDM development methods in the health domain and propose a suggestive development process that might ease the development of CDMs in the health domain in the future.

(JMIR Med Inform 2023;11:e45116) doi:10.2196/45116

KEYWORDS
common data model; common data elements; health data; electronic health record; Observational Medical Outcomes Partnership; stakeholder involvement; Data harmonisation; Interoperability; Standardized Data Repositories; Suggestive Development Process; Healthcare; Medical Informatics;
**Introduction**

**Rationale**
Integration of heterogeneous data is a ubiquitous topic in modern medicine. The arising large variety of data has the potential to provide in-depth insights about different aspects of clinical care and can lead to improvements in health care [1,2]. Yet, challenges, such as the identification and access of relevant data, the association between different data sources, and the assurance of data quality given the structural variations among data sources, still pose major barriers [3,4]. Common data models (CDMs) provide the possibility of harmonizing data from disparate sources, storing information in a standard structure by defining the syntax and semantics of data, and enabling operations on data using standard analysis methods [5]. In particular, a CDM contains a unified set of metadata, allowing data and its information content to be shared across applications and institutional borders, and thus enabling harmonized data integration and analysis on an international scale [6].

In the health domain, there are different types of CDMs (eg, CDMs for harmonization and storage of electronic health record–based patient data). An example is the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) developed by the Observational Health Data Science and Informatics (OHDSI) community, which ensures homogeneous storage of observational health care data across different databases with similar formats and terminologies [7]. There are also further CDMs for clinical data, like Sentinel CDM, Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM), and National Patient-Centered Clinical Research Network (PCORnet) [8], and data warehouse models, like Informatics for Integrating Biology and the Bedside (i2b2) [9]. Moreover, some CDMs define the data from patient cohorts and describe a medical specialty or a group of diseases. For example, there are specific CDMs for the domain of rare diseases [10,11] or radiology [12]. Overall, there is a large variety of CDMs in the literature for common, rare, and context-specific medical examinations, and each of them follows a more self-defined development process.

As described by Melles et al [13], a practical design meets the users’ needs. While designing a CDM in the health domain, in addition to the developers (ie, IT staff and computer scientists), the primary stakeholders (ie, patients and clinicians) are particularly interested in the outcome. It is therefore recommended to include them in the design process as early as possible [13,14]. In addition to the stakeholders, the medical context is also quite complex and requires extensive medical and technical expertise to ensure the usefulness of the model after its development. This is why the development process of a CDM is critical and a comprehensive development method or guideline is necessary.

Studies, such as those by Gericke and Blessing [15] and Bobbe et al [16], have already tried to determine the commonalities and differences in development processes across disciplines. Bobbe et al [16] performed a comparison of design models from academic theory and professional practice, and discussed 8 types of design processes. In particular, the basic design cycle, V design process, human-centered design, hypercyclic design, Munich procedural model, double diamond model, frog model, and IDEO model were presented. Additionally, Melles et al [13] introduced categories for models, namely, whether a model is activity-based or stage-based, solution-oriented or problem-oriented, and design-focused or project-focused.

However, given the complexity of the health domain and the importance of many stakeholders taking part in the process, it might be difficult to transfer models from other disciplines. This is why we aim to derive such a process and review the available CDM instances in the domain. Exemplarily, the results of this scoping review will be integrated into the design and development of a CDM for the SATURN (“Smartes Arztportal für Betroffene mit unklarer Erkrankung” [“Smart physicians’ platform for patients with unclear diseases”]) Project in the future [17]. This project aims to develop an artificial intelligence–based diagnosis support tool for primary care physicians. With the help of user-centered design, the requirements of a decision support tool, especially for noncharacteristic symptoms, will be studied. The medical focus is on the diagnosis of unclear and rare medical conditions. This is why, in this review, we focus on the similarities between the CDM development methods in rare medical conditions and common medical conditions in order to determine whether the methods for common medical conditions can be adopted for rare medical conditions as well. On a technical level, rule-based systems, machine learning, and case-based reasoning will be implemented. As part of this project, CDMs for 3 groups of rare diseases, namely, endocrinology, gastroenterology, and pneumology, will be developed.

Our review contributes to the analysis of CDM development methods in the health domain on an international scale and aims to explore the actual involvement of stakeholders, especially medical experts, in the development process. To the best of our knowledge, this is the first scoping review focusing on CDM development methods in the health domain.

**Objectives and Research Questions**
This scoping review has been conducted to provide an overview of the methods used for the initial and further development of CDMs in the health domain. We divided the overall development process into conception, users’ needs collection (eg, collection of evidence, review of the literature, and guidelines), and implementation, as well as individual evaluations within the phases. We consider the conception phase as an initial step, where the CDM is theoretically designed along with stakeholders. Subsequently, the essential elements previously identified are gathered in the “users’ needs collection” phase. The finalized process, in which the conceptualized model is implemented and ready-to-use, is termed the implementation phase.

According to the rationale and objective explained above, this scoping review examines the following questions:
1. How are CDMs methodically developed in the health domain? What requirement analysis methods, design processes, and validation methods were used?
2. How or when do stakeholders, especially medical experts, get involved in the development process?

3. How can the CDM development methods be classified based on their requirement analysis methods, design processes, validation methods, and model type?

**Methods**

**Protocol and Registration**

To ensure methodological quality, this scoping review has followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist [18]. According to this checklist, we published and registered the review protocol [19]. Out of the 22 items of the PRISMA checklist, 20 have been considered in this review (Multimedia Appendix 1).

**Search Strategy**

To achieve a comprehensive query, an initial search was performed in PubMed with the term "common data model." Six randomly chosen articles matching the topic were analyzed [10-12,20-22]. The keywords associated with the articles listed in Table 1 were considered and subsequently tested in the query.

<table>
<thead>
<tr>
<th>Article title</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>The EPIRARE proposal of a set of indicators and common data elements for the European platform for rare disease registration [10]</td>
<td>Registries, common data elements, European platform, rare diseases, patient registration, and EPIRARE</td>
</tr>
<tr>
<td>A methodology for a minimum data set for rare diseases to support national centers of excellence for healthcare and research [11]</td>
<td>Common data elements, interoperability, metadata, minimum data set, national health program, and rare diseases</td>
</tr>
<tr>
<td>Development and validation of the Radiology Common Data Model (R-CDM) for the international standardization of medical imaging data [12]</td>
<td>Metadata, standardization, and radiology information system</td>
</tr>
<tr>
<td>Common data model for natural language processing based on two existing standard information models: CDA+GrAF [20]</td>
<td>Natural language processing, medical informatics, data model, information model, HL7 clinical document architecture, and ISO graph annotation format</td>
</tr>
<tr>
<td>Genomic common data model for biomedical data in clinical practice [21]</td>
<td>High-throughput nucleotide sequencing, data analysis, and observational study</td>
</tr>
<tr>
<td>Towards a newborn screening common data model: The Utah Newborn Screening Data Model [22]</td>
<td>Newborn screening, newborn screening laboratory information management system, common data model, interoperability, electronic data exchange, NBS, LIMS, and standards</td>
</tr>
</tbody>
</table>

**Table 2.** Search strings used to identify articles from PubMed.

<table>
<thead>
<tr>
<th>Search aspects</th>
<th>Variations</th>
<th>Search string&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common data model</td>
<td>Common data model (CDM), common data element (CDE), and common data sets (CDS)</td>
<td>(“common data model” AND CDM) OR (“common data element*” AND CDE) OR “Common Data Elements”[Mesh] OR “common dataset*” OR “common data set*”</td>
</tr>
<tr>
<td>Health care</td>
<td>Medical, medicine, health, healthcare, health care, electronic health, and disease</td>
<td>medical OR medicine OR “Medicine”[Mesh] OR health OR “Health”[Mesh] OR healthcare OR “health care” OR “electronic health” OR clinical OR disease OR “Disease”[Mesh]</td>
</tr>
</tbody>
</table>

<sup>a</sup>The common data model and health care search terms were combined with “AND.”

In particular, literature from 2000 to 2022 was considered, which is an extension of the previously published study protocol [19]. It is also noteworthy that the MeSH terms were only available in PubMed. The language of the articles was limited to English. Using the Boolean operators “AND” and “OR,” the systematic search was carried out in the following electronic databases: PubMed, Web of Science, Science Direct, and Scopus. The search was performed in March 2022. The publication date tag in PubMed and Web of Science was set to January 1, 2000, to March 15, 2022, and that in Science Direct and Scopus was set to 2000 to 2022 (it is not possible to specify the month and day in Science Direct and Scopus).
Inclusion and Exclusion Criteria

The inclusion and exclusion criteria are summarized in Multimedia Appendix 3 and are visualized along with the number of outcome articles in Figure 1.

**Figure 1.** PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart showing the paper selection process and the inclusion and exclusion criteria. CDM: common data model.

**Selection and Review of Articles**

Duplicates were removed using the built-in function in Rayyan [24]. The process of deletion was monitored by the author NA. After eliminating duplicates, the selection of studies was performed in 2 steps. The title and abstract screening steps were performed by the authors in groups of two. The articles were tagged as “include,” “exclude,” or “maybe.” Tagged articles were decided upon based on the tags described in Table 3. Disagreements were resolved by a third author. This process was initially carried out on 10% of the articles to confirm the accuracy of our inclusion and exclusion criteria, and clarify ambiguities. After the title and abstract screening, the full text of the included articles was screened by the authors, again in groups of two. The selected articles were included in the data extraction step.

**Table 3.** Description of tags used by the authors in the article screening process.

<table>
<thead>
<tr>
<th>Decision</th>
<th>Author 1</th>
<th>Author 2</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Include</td>
<td>Include</td>
<td>Include</td>
<td>Included</td>
</tr>
<tr>
<td>Include</td>
<td>Include</td>
<td>Exclude</td>
<td>Discuss and decide together</td>
</tr>
<tr>
<td>Include</td>
<td>Maybe</td>
<td>Include</td>
<td>Include</td>
</tr>
<tr>
<td>Exclude</td>
<td>Maybe</td>
<td>Exclude</td>
<td>Exclude</td>
</tr>
<tr>
<td>Maybe</td>
<td>Maybe</td>
<td></td>
<td>Discuss and decide together</td>
</tr>
</tbody>
</table>

**Data Charting and Extraction Process**

A data charting table was developed and refined throughout the study, with several iterations. This table contained a list of items that were extracted from all included publications. All authors examined 10% of the articles for the defined data items and refined the data charting table, if necessary. The data charting table, including the extracted information from articles, is included in Multimedia Appendix 4.

For each article, we focused on 4 major aspects: (1) the meta information, such as DOI, authors, year, country, and project name, if applicable; (2) the medical condition for which the CDM was built, whether the condition is rare or common, the organ affected by the condition, and whether the condition is long term (longer than a year) or short term; (3) methodological information, such as requirement analysis, design, and validation process; whether the design process was linear or iterative; and advantages and disadvantages of the method, as stated in the respective article; and (4) information about stakeholder involvement. The extracted data elements, their categories, and their definitions are shown in Table 4.
Table 4. Data extraction sheet with specified elements, categories, and subcategories, including their definitions.

<table>
<thead>
<tr>
<th>Category and subcategory</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Meta information</strong></td>
<td></td>
</tr>
<tr>
<td>DOI</td>
<td>A link to the article</td>
</tr>
<tr>
<td>Author</td>
<td>First author’s name</td>
</tr>
<tr>
<td>Publication year</td>
<td>Year of the publication date of the article</td>
</tr>
<tr>
<td>Country of study</td>
<td>Country of the leading author’s affiliation</td>
</tr>
<tr>
<td>Project name</td>
<td>If applicable; when the CDM study was part of a project/consortium</td>
</tr>
<tr>
<td><strong>Medical background</strong></td>
<td></td>
</tr>
<tr>
<td>Medical condition</td>
<td>Name of the medical condition for which the CDM was built</td>
</tr>
<tr>
<td>Organ function</td>
<td>Organ affected by the medical condition</td>
</tr>
<tr>
<td>Short-term/long-term condition</td>
<td>Short term: less than a year; long term: longer than a year</td>
</tr>
<tr>
<td>Is the condition rare or common?</td>
<td>Is the medical condition considered rare or common based on its occurrence? Available answers: common medical condition, rare medical condition, and conditions that can be rare and common.</td>
</tr>
<tr>
<td><strong>Requirement analysis method</strong></td>
<td></td>
</tr>
<tr>
<td>Literature analysis</td>
<td>It includes searching in a variety of literature, such as extraction of frequent CDEs from real-world data, data harmonization across studies, multicenter longitudinal and observational studies, consensus documents and guidelines, primary outcome data of trials, review of instruments, and forms like report forms, users’ needs collection forms, etc.</td>
</tr>
<tr>
<td>Interview/questionnaire</td>
<td>It includes expert interviews, focus group meetings, working group meetings, consensus meetings, workshops and discussions, and online surveys.</td>
</tr>
<tr>
<td>Delphi</td>
<td>Delphi or modified Delphi was used. Delphi techniques involve experts evaluating complex issues iteratively, where knowledge is incomplete or uncertain. Typically, the response from the previous questionnaire is appended to the next questionnaire [25].</td>
</tr>
<tr>
<td>Review of existing CDEs</td>
<td>When an existing CDE was validated/reviewed.</td>
</tr>
<tr>
<td><strong>Design</strong></td>
<td></td>
</tr>
<tr>
<td>Creation of new CDEs</td>
<td>If there were no CDEs in the domain and the experts tried to come up with some CDEs using literature in the field.</td>
</tr>
<tr>
<td>Modification of existing CDEs</td>
<td>If existing CDEs in a disease domain were modified.</td>
</tr>
<tr>
<td>Reuse of existing CDEs (without modification)</td>
<td>If existing CDEs in the domain were used without any modification.</td>
</tr>
<tr>
<td><strong>Validation</strong></td>
<td></td>
</tr>
<tr>
<td>External experts</td>
<td>It includes only external validation of any sort, such as public reviews on a website from experts or nonexperts in the field. Excluded are experts that were part of the conception process of the model.</td>
</tr>
<tr>
<td>Others</td>
<td>Any other type of validation, such as internal reviews, working group consensus, etc.</td>
</tr>
<tr>
<td><strong>Model type</strong></td>
<td></td>
</tr>
<tr>
<td>Iterative</td>
<td>When at least one iterative process was performed during development of the CDM.</td>
</tr>
<tr>
<td>Linear</td>
<td>When there was no iteration in the process.</td>
</tr>
<tr>
<td><strong>Stakeholder information</strong></td>
<td></td>
</tr>
<tr>
<td>Were stakeholders involved in the design process?</td>
<td>Yes/no</td>
</tr>
<tr>
<td>Which stakeholders were involved?</td>
<td>Patients’ representatives, clinicians, domain experts, computer scientists, IT personnel, and registry staff</td>
</tr>
<tr>
<td>When did they get involved in the process?</td>
<td>In users’ needs collection (when experts were involved in the preanalysis step, eg, collection of evidence, review of literature, guidelines, etc), in conception (when experts were involved in conception of the CDEs), in evaluation (when the model was evaluated via experts), and in implementation (when experts were involved in the implementation of the model).</td>
</tr>
<tr>
<td>What was the nature of stakeholder involvement?</td>
<td>Through expert workshops, semistructured interviews, questionnaires, etc</td>
</tr>
</tbody>
</table>

**Pros and cons of methods as mentioned in the article**
Table 4

<table>
<thead>
<tr>
<th>Category and subcategory</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pros</td>
<td>Advantages of the method as stated in the article</td>
</tr>
<tr>
<td>Cons</td>
<td>Disadvantages of the method as stated in the article</td>
</tr>
</tbody>
</table>

aCDM: common data model.
bCDE: common data element.

Visualization and Summarization of Results

At the end of the data extraction, the data items collected in Table 4 were summarized and visualized. A flowchart according to the PRISMA-ScR guidelines was designed to show the article processing approach (Figure 1). Tables, timeline plots, histogram charts, pie plots, and scatter histograms were used to display the extracted data items. The graphics and the required analysis were performed using Python version 3.9.12 (Python Software Foundation), with matplotlib, pandas, and NumPy packages. The script used for the plots is publicly available [19,26].

First, we aimed for a broad overview of available CDMs and whether original CDMs were developed or existing CDMs were modified, as well as whether they focused on common or rare diseases and addressed a specific organ function. Second, to answer our first research question, we documented the medical domain of each article, whether the medical condition was considered as long term (more than a year) or short term (less than a year), and the affected organ as stated in the respective original article. To classify the development process of the CDMs, we documented 4 categories of data information for each article: requirement analysis, design, validation, and model type (Table 4). We categorized the methodology that was used for the requirement analysis (ie, why a CDM was needed), as well as the context to design a set of common data elements (CDEs). For validation, we distinguished between external evaluation and any other type of evaluation. The “other” category included the evaluations performed by the same clinical experts who were involved in the conception process, such as working group consensus, user evaluations, reviews performed via the members of the project, statistical tests, and pilot tests conducted within the project. Additionally, we investigated stakeholder involvement in the development stages in those studies and whether the studies followed an iterative or linear method of development. We used the advantages and disadvantages of the methods as stated in the articles (Table 4) and formulated them into a list of constraints in the area of CDM development to further highlight the need for streamlined methods. Finally, after analyzing the included CDMs, we summarized the most frequent methods used in the included literature in a suggestive development process that could be a reasonable basis to start with when developing a novel model.

Results

Selection of Articles

In total, we identified 1309 articles from PubMed, Web of Science, Science Direct, and Scopus search engines. From the identified articles, after duplicate removal, 695 articles were included in the title and abstract screening. Finally, 465 articles underwent full-text screening, and of these, 59 matched the full-text screening criteria of this review and were finally included. We excluded articles that did not describe the development or evaluation of a CDM in the health domain. Additionally, articles that were not publicly available and those in a language other than English were excluded. The article identification process along with the inclusion and exclusion criteria are shown in Figure 1.

The selected articles defined CDMs, common data sets, or CDEs for common or rare medical conditions. All included articles were published between 2000 and 2022. As shown in Figure 2, the number of articles that focused on CDM development increased after 2011 and continued to increase in the last years.
Country of Publication

We categorized the articles into countries based on the affiliation of the first author. Among the 59 articles, 26 (44%) were published in the United States, 8 (14%) were published in Canada, and 6 (10%) were published in Germany. The number of articles according to country is as follows: Belgium, 2 [27,28]; Canada, 8 [29-36]; China, 1 [37]; Denmark, 2 [38,39]; France, 2 [11,40]; Germany, 6 [41-46]; Italy, 1 [10]; Spain, 1 [47]; Republic of Korea, 1 [48]; Norway, 3 [49-51]; Switzerland, 1 [52]; Taiwan, 1 [53]; the Netherlands, 1 [54]; United Kingdom, 3 [55-57]; and United States, 26 [58-83].

Medical Conditions and Their Domains

According to our research, CDMs were developed for a variety of medical domains in the past 22 years; however, we divided them into 3 categories, namely, rare, common, and rare and common (both). An aggregated list of the medical conditions and their domains is shown in Figure 3. A full list of the medical conditions extracted during this scoping review is shown in Multimedia Appendix 4. An organ function overview and the long- and short-term conditions are shown in Multimedia Appendix 5. Among these, 10 (17%) CDMs were designed for rare medical conditions, such as myeloid leukemia and rare lung diseases, and mitochondrial diseases [41,44-46,59]. Moreover, 1 CDM, namely, the CDM in the study by Berger et al [44], was designed for undiagnosed diseases in general.

Among the 59 articles, 45 involved the development of a CDM for common medical conditions. These included traumatic brain injury [27,28,30], spinal cord injury in children and youth [67], dental caries [68], sport-related concussion [65], cerebral palsy [29], degenerative cervical myelopathy [55], unruptured intracranial aneurysms and subarachnoid hemorrhage [32,42,55,60], Chiari malformation type I [63], breast implant [43], stroke [37], venous thromboembolism [33], pediatric epilepsy [61], pediatric critical illness [62], pregnancy drugs and treatments [49], sepsis [31], medication use in pregnancy and breastfeeding [40], degenerative cervical myelopathy [55], Gulf War illness [58], neuroinflammatory demyelinating disease [43], traumatic brain injury [27], and neurologic disorder and stroke [69]. Wandner et al [66] focused on clinical pain management, and Jaboyedoff et al [52] focused on pediatric diseases in general.
**Figure 3.** Characteristics of the included studies. A Venn diagram showing the proportions of identified common data models (CDMs) for common medical conditions (76.3%; blue), rare medical conditions (16.9%; golden yellow), and medical conditions that could fit into both categories (6.8%). Additionally, an aggregated list of medical conditions that CDMs were developed for in the studies is shown in 3 different colors according to their categories.

### Stakeholder Involvement

To investigate the involvement of stakeholders, we summarized at which particular stage they were involved in the CDM development process. Out of the 59 included articles, 54 (92%) mentioned at least one stakeholder in the design process. Additionally, we were interested in the different types of stakeholders that were involved, how they were involved, and at what stage of the process they typically got involved. As shown in Figure S1 in Multimedia Appendix 6, stakeholders were mostly involved in the initial stage, namely, the conception phase. Domain experts and clinicians were the most common stakeholders involved in the studies (Figure S2 in Multimedia Appendix 6). Additionally, while many different methods were used to involve the stakeholders, such as expert groups, surveys, consensus meetings, interviews, teleconferences, questionnaires, and workshops, “working group” was the most frequent method used (Figure S3 in Multimedia Appendix 6).

### Design Process

The methods used in the articles for designing a CDM were literature analysis, interview, Delphi, and review of existing CDEs. From our extraction table (Multimedia Appendix 4), we noted that 39 articles involved the definition of an original model/set of CDEs, 13 involved the modification of an existing set of CDEs, and 29 involved the use of an existing set of CDEs without any modifications. The external evaluation included web-based feedback, public review and comments, and feedback in a conference, among others. Finally, we found that 26 articles involved a rather linear design method and 22 others involved an iterative process. The list of articles that involved the use of each of these categories is shown in Figure 4. Detailed information is presented in Multimedia Appendix 4.
Figure 4. Methodological information on the articles [10,11,27-83]. The y-axis shows the list of articles by publication year. The x-axis shows the methodological categories. The scatter plot includes a cross mark when the Boolean is true for a specific article, for example, if the authors have used literature analysis as a preanalysis method, a cross (x) is added. The sum of cross marks in each column contributes to the bar size of the bar plot positioned on the x-axis. To improve visibility, each subcategory is shown with a different color. The subcategories of the same category are grouped via the same family of colors. CDE: common data element.

Methodological Constraints Highlighted in Previous Studies

The included articles presented a range of constraints in the development process from the methods used in the different stages of the process to the applicability of the outcome elements. For example, Thurin et al [40] performed interviews with a single data access provider per data source and mentioned that other data access providers might conceptualize the data source differently. Additionally, they tested the applicability of the developed model only on the included data sources in the project. The model might require modification to use it with...
other data sources. The limited sample size used to test the developed model is a common problem in rare conditions [44] given the rarity of the disease. One of the limitations mentioned by Broglio et al [65] is that some of their developed CDEs require special expertise that might not be implementable in certain settings. Grinspan et al [61] mentioned that some subcategories of epilepsy syndrome were merged at a level higher into a single category, which might have led to reduced data resolution, although uphill mapping is often used, especially in the OMOP context [5]. Additionally, the elements considered do not cover every possible influencing element, and the source was limited to only US-based patients, which means the elements can differ once an international data level is considered. They also included CDEs that were documented as free text, and processing of such elements might require natural language processing applications. The authors also highlighted the possible bias caused by the methodology used for consensus and discussion, and the Delphi approach, focus groups, and interviews might have also influenced the outcome of the study.

Figure 5. Summary of a basic common data model development process.

Discussion

Overview

One of the major challenges faced by CDM developers in the health domain is the lack of a comprehensive methodology or workflow to follow, which is also reflected in this review. The general models from industrial design and even academia (eg, the model introduced by Bobbe et al [16]) do not generally translate one-to-one to the health domain. The medical context is usually complex, and the involvement of stakeholders, such as clinicians, patients’ representatives, and IT staff, is of utmost importance to ensure the applicability of a to-be-developed CDM. In addition, user-friendly, adaptable, and straightforward models are preferred in health care as one can start working with them without requiring a substantial amount of time [84].

This scoping review provides a summary of the development methods for CDMs and categorizes them based on the requirement analysis method, design process, validation approach, and model type. A variety of methods were used in the requirement analysis step in the articles, starting from searching in different types of literature and medical guidelines [43,44] to interviews [29], the Delphi approach [31], and a review of existing CDEs. A full list of these articles is shown in Figure 4 and Multimedia Appendix 4.

The majority of the developed CDMs have been designed for common medical conditions, and only 10 articles involved the design of a particular CDM for rare diseases. However, we did not find a significant difference in the development process of a CDM for rare and common conditions. Interestingly, based on our analysis, we can conclude that common medical conditions were the focus of CDM studies from early 2000, whereas the first CDM for rare conditions was developed in 2014. Despite methodological similarities, every article usually mentioned following a more individualistic method of development. This may arise because rare conditions occur rarely and the number of patients included in studies is limited [44]. Moreover, finding an expert for each rare or unclear disease is a challenging task. Additionally, most of the information crucial in the diagnosis of such diseases (like symptoms or phenotypes and genotypes) is currently stored in unstructured forms (eg, clinical notes). Extraction of such information requires a lot of time and effort from technical and clinical stakeholders [41].
Thus, given the variety of studies, the methods used for common conditions might be adaptable for rare conditions. Considering that a CDM is an essential part of data harmonization (a necessity in the health domain), we see highly emphasized development models as essential. Therefore, after analyzing the included CDMs, we summarized a suggestive development process that is shown in Figure 5, which could be the starting point for conceptualizing and implementing novel CDMs.

Limitations
The findings of our study are subject to certain limitations. First, our analysis is restricted to the selected databases, namely, PubMed, Web of Science, Science Direct, and Scopus. Additionally, the scope of our investigation is confined to articles published within a specific time frame and written in English. Moreover, we did not conduct any assessment of the quality of the included articles. In addition, it may also be worth noting that the authors of this review have varying interdisciplinary backgrounds, expertise levels, and experiences in the CDM field. However, to optimize the screening and analyzing processes, we performed them in pairs and first tested the method on a subset of 10% of the articles, resulting in a minimal number of conflicts.

Conclusion
We considered 4 steps in the development of a CDM: conception, users’ needs collection, implementation, and evaluation. We could identify 4 groups of methods that were most often used in the articles as part of the requirement analysis of the CDM development process. These were literature analysis, interviews, Delphi approaches, and review of existing CDEs. The articles considered in this review either developed a new CDE or made use of an existing set of CDEs with or without modification.

Most of the articles involved at least one stakeholder from among domain experts, clinicians, IT staff, registry staff, and patients’ representatives, and mostly from the initial step, which was conception. The methods used to involve the stakeholders were expert groups, surveys, consensus meetings, interviews, working groups, teleconferences, questionnaires, and workshops, and among these, working groups were most often used.

We conclude that the methods used in the development of CDMs in the health domain are heterogeneous and this field is lacking solid guidelines that may ease up this process, especially in terms of the reusability and adaptability of a CDM. This is why the proposed outline (Figure 5) could be a reasonable basis to start with. In our future work, we plan to test and improve the proposed outline for developing a CDM.

Acknowledgments
This work was accomplished as part of the SATURN (Smartes Arztportal für Betroffene mit unklarer Erkrankung; Smart physicians’ platform for patients with unclear diseases) Project funded by the German Federal Ministry of Health as part of the research focus “Digital Innovation,” Module 3: “Smart Algorithms and Expert Systems” (funding codes: 2520DAT02C, 2520DAT02B, and 2520DAT02D).

Data Availability
The script used for analysis and visualization in this review is available at GitHub [26], and the study protocol can be accessed on the Open Science Framework (OSF) [19].

Authors' Contributions
NA, MZ, PK, RN, MW, JS, and MS contributed to conceptualization and methodology. NA contributed to data acquisition. NA, MZ, PK, RN, and MW contributed to the literature screening. NA contributed to data analysis and interpretation. NA contributed to writing and preparing the original draft. NA, MZ, PK, RN, MW, and JS contributed to reviewing and editing the manuscript. NA and MW contributed to the visualization. MS contributed to resources. All authors take responsibility for the scientific integrity of the work. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 940 KB - medinform_v11i1e45116_app1.pdf ]

Multimedia Appendix 2
Search strings used in the PubMed, Web of Science, Science Direct, and Scopus databases to search for articles.
[DOCX File, 17 KB - medinform_v11i1e45116_app2.docx ]

Multimedia Appendix 3
References


17. SATURN Projekt. URL: https://www.saturn-projekt.de/ [accessed 2023-07-06]


Review

Designing Interoperable Health Care Services Based on Fast Healthcare Interoperability Resources: Literature Review

Jingwen Nan¹, MA; Li-Qun Xu¹, PhD
Health IT Research, China Mobile (Chengdu) Industrial Research Institute, Chengdu, China

Corresponding Author:
Li-Qun Xu, PhD
Health IT Research
China Mobile (Chengdu) Industrial Research Institute
Unit 2, Block C1, AI Innovation Center
Hele Second Street, Gaoxin District
Chengdu, 610213
China
Phone: 86 (28) 60103585
Email: xuliqun@chinamobile.com

Abstract

Background: With the advent of the digital economy and the aging population, the demand for diversified health care services and innovative care delivery models has been overwhelming. This trend has accelerated the urgency to implement effective and efficient data exchange and service interoperability, which underpins coordinated care services among tiered health care institutions, improves the quality of oversight of regulators, and provides vast and comprehensive data collection to support clinical medicine and health economics research, thus improving the overall service quality and patient satisfaction. To meet this demand and facilitate the interoperability of IT systems of stakeholders, after years of preparation, Health Level 7 formally introduced, in 2014, the Fast Healthcare Interoperability Resources (FHIR) standard. It has since continued to evolve. FHIR depends on the Implementation Guide (IG) to ensure feasibility and consistency while developing an interoperable health care service. The IG defines rules with associated documentation on how FHIR resources are used to tackle a particular problem. However, a gap remains between IGs and the process of building actual services because IGs are rules without specifying concrete methods, procedures, or tools. Thus, stakeholders may feel it nontrivial to participate in the ecosystem, giving rise to the need for a more actionable practice guideline (PG) for promoting FHIR’s fast adoption.

Objective: This study aimed to propose a general FHIR PG to facilitate stakeholders in the health care ecosystem to understand FHIR and quickly develop interoperable health care services.

Methods: We selected a collection of FHIR-related papers about the latest studies or use cases on designing and building FHIR-based interoperable health care services and tagged each use case as belonging to 1 of the 3 dominant innovation feature groups that are also associated with practice stages, that is, data standardization, data management, and data integration. Next, we reviewed each group’s detailed process and key techniques to build respective care services and collate a complete FHIR PG. Finally, as an example, we arbitrarily selected a use case outside the scope of the reviewed papers and mapped it back to the FHIR PG to demonstrate the effectiveness and generalizability of the PG.

Results: The FHIR PG includes 2 core elements: one is a practice design that defines the responsibilities of stakeholders and outlines the complete procedure from data to services, and the other is a development architecture for practice design, which lists the available tools for each practice step and provides direct and actionable recommendations.

Conclusions: The FHIR PG can bridge the gap between IGs and the process of building actual services by proposing actionable methods, procedures, and tools. It assists stakeholders in identifying participants’ roles, managing the scope of responsibilities, and developing relevant modules, thus helping promote FHIR-based interoperable health care services.

(JMIR Med Inform 2023;11:e44842) doi:10.2196/44842

KEYWORDS
Health level 7 Fast Healthcare Interoperability Resources; HL7 FHIR; interoperability; literature review; practice guideline; mobile phone
Introduction

Background

The development and innovation of health care service models have accelerated the demand for data exchange and service interoperability. In the United States, the Health Information Technology for Economic and Clinical Health Act took effect in 2009, specifying health IT–based systems as an integrated part of the country’s health care reform. It has spurred the electronic health record (EHR) adoption rate through reward and punishment measures [1]. In addition, the US Department of Health and Human Services established a specific agency, the Office of the National Coordinator for Health Information Technology, to accelerate the implementation of advanced medical IT standards, promote the exchange of electronic health care information, and improve the quality of health care services throughout the country. In Canada, the federal government funded an independent, not-for-profit organization called Canada Health Infoway, tasked with accelerating the adoption of digital health solutions, such as EHR, across the country. The government has set a 10-year implementation strategy for EHR in cooperation with the Canadian Institute for Health Information [2]. Japan has made great efforts to develop remote health care technology and has established a communication system among regional institutions by implementing electronic medical records (EMRs) in the form of an app or software as a service [3]. In China’s state health system, major public hospitals administered by national, provincial, and local health authorities are the pioneers in reforms. Over the years, the government has issued a series of policies promoting coordinated care among health care institutions at different levels of the health system [4,5], together with many qualitative or quantitative assessment criteria that guide the establishment of high-standard EMR system, regional information interoperability, and intelligent service and management in hospitals. In summary, the demand for tiered and coordinated care delivery among health care institutions worldwide is increasing rapidly, and the requirement for health care data exchange continues unabated.

The enhancement of interoperability is required by transforming health care service models and tackling the challenges of societal problems. According to a United Nations report [6], the share of the population aged ≥65 years is expected to increase from 9.3% in 2020 to approximately 16% in 2050. The rapid aging of the population unavoidably increases the burden of chronic disease care, bringing about the requirements for people-centered and continuous care delivery built on the foundation of a robust primary health care system. Therefore, it is necessary to enhance health IT system interoperability to bridge the gap between uneven health care resource distribution, remove the barrier of isolated data islands, and comprehensively improve the quality of health care services.

Health Level 7 Fast Healthcare Interoperability Resources

Health Level 7 (HL7), founded in 1987, is a not-for-profit, standards-developing organization dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery, and evaluation of health services. It has successively released many standards, including HL7 version 2, HL7 version 3, and Clinical Document Architecture (CDA). However, with the constant evolution of the internet and the thriving of the application programming interface (API) economy, digital services or assets of health organizations tend to be exposed even more widely in the form of APIs. In this context, HL7 formally introduced Fast Healthcare Interoperability Resources (FHIR) in 2014, highlighting the core concept of resources, and thus, creating a new era for health care service interoperability. A resource is the smallest exchangeable logical unit in FHIR. Resources are independent of each other but can be linked or assembled through specific rules to meet diverse service requirements. FHIR combines web standards to support resource operations through RESTful API in XML or JavaScript Object Notation format. Compared with other alternative standards, FHIR has more advantages and potential, such as comprehensive coverage of data definitions, substantial flexibility of data exchange, explicit semantics, and many available open-source tools, among others. Therefore, it has attracted constant and favorable attention from health care stakeholders since its first release, as shown in Figure 1.

We investigated the literature from the Web of Science and plotted 2 statistical charts in Figure 1. Figure 1A shows the promotion trends of different health data standards. By using the search term “HL7 v2,” “HL7 v3,” “HL7 CDA,” and “FHIR,” we identified the corresponding papers in the Web of Science database from 2010 to 2022. The results show that the attention paid to FHIR has increased rapidly within a short time, far exceeding the HL7 version 2, HL7 version 3, and CDA standards. Figure 1B compares FHIR-relevant literature among different countries. We used the search term “FHIR” to find the corresponding papers in the Web of Science database from 2014 to 2022. By reading each paper’s abstract and the corresponding author’s information, we identified the country to which the work belongs. Countries that record <5 papers fall into the “others” category. The chart shows that the United States, Germany, and Canada were the top 3 countries that published the most studies on FHIR, accounting for 28.39% (197/694), 11.67% (81/694), and 4.18% (29/694), respectively.

In addition to the dissemination activities of enthusiastic researchers and pioneering health IT ecosystem players, national health policy makers also play a pivotal role in FHIR adoption, as evidenced by the actions in the United Kingdom, United States, and Canada [7]. Overall, FHIR has gradually gained worldwide recognition and acceptance, and it has the most potential for future large-scale promotion in the health care ecosystem.
Figure 1. Works of literature that focus on health data standards. (A) The attention to Fast Healthcare Interoperability Resources (FHIR) has risen rapidly within a short time of its first release, far exceeding HL7 version 2, HL7 version 3, and Clinical Document Architecture (CDA) standards. (B) The United States, Germany, and Canada are the top 3 countries that published the most literature on FHIR. HL: Health Level.

Objectives

Owing to the growing popularity of FHIR, some academic researchers have authored review papers from their perspectives in the last few years. Ayaz et al [8] searched for FHIR-related papers published between 2012 and 2019 in 6 databases (ACM, IEEE, Springer, Google Scholar, PubMed, and ScienceDirect) and selected 80 papers for review. They found that FHIR is identical in supporting intelligent technologies, such as smartphones, tablets, mobile health apps, smartwatches, and fitness trackers, which could solve numerous health care problems that were impossible for the previous standards. Lehne et al [9] searched for FHIR-related papers in 2 databases (Web of Science and PubMed) up to 2019 and selected 131 papers for review. The statistical results revealed that data model–related topics mainly focusing on constructing profiles to implement FHIR in specific scenarios were the most attractive direction. At the same time, analytics-related topics concerning data analysis, modeling, machine learning, and more were less attractive because most FHIR projects were still in the initial development phase, dealing with implementation and data definitions rather than large-scale data analysis. Barker and Johnson[10] surveyed 734 apps released up to December 2020 in 5 digital health care application libraries (hosted by Cerner, Epic, Allscripts, Athenahealth, and Substitutable Medical Applications Reusable Technologies [SMART]) and measured their support for FHIR. They found that the number of apps that support the FHIR standard had increased from 19% in 2019 to 22% in 2020.

However, to our knowledge, there is a lack of systematic reviews that focus on the FHIR practice. A gap remains between the FHIR Implementation Guide (IG) and building actual services because IGs are rules specifying no methods, procedures, or tools. Thus, stakeholders may feel it nontrivial to participate in the ecosystem, giving rise to the need for a more actionable practice guideline (PG) for promoting FHIR’s fast adoption. Therefore, this study proposed a general FHIR PG to facilitate stakeholders in the health care ecosystem to understand FHIR and quickly develop interoperable health care services.

Methods

Article Selection

Figure 2 presents the paper selection flowchart used in this review. Initially, we identified a total of 487 papers in the Web of Science and IEEE databases by using the search term “FHIR” or “Fast Healthcare Interoperability Resources.” The time range of publications was set from January 1, 2020, to July 1, 2022, and we finalized 205 articles. After excluding those that merely mentioned the term FHIR but did not elaborate on it, 65 articles were retained. A check of duplications from this batch removed a further 3 articles. Finally, from the references of the remaining 62 articles, we found an additional 23 relevant articles, ending up with a total of 85 articles as the research materials of this study.
Analysis Process
By carefully analyzing and collating the recent studies on the design of FHIR-based interoperable health care services, we derived the details of the FHIR PG.

We selected 85 FHIR-related articles and found that building FHIR-based health care services contains typically 3 stages, that is, data standardization, data management, and data integration. Each stage may use different practice methods, depending on the targeted scenarios and types of services.

The way to categorize these 85 articles is as follows: if an article’s main innovation feature focused on 1 of the 3 stages, we assigned it to the corresponding group. Specifically, we assigned those articles emphasizing the design process of FHIR profiles or proposing methods for migrating data from specific clinical data models (CDMs) to FHIR to the data standardization group, articles discussing the management of RESTful APIs to the data management group, and articles presenting approaches for integrating data with specific apps or platforms to the data integration group.

After categorizing the articles, we reviewed the key techniques used by each group to build their respective health care services. We compiled a general FHIR PG through this review. The workflow of the FHIR PG was derived by linking the stages, each consisting of multiple steps. It is important to note that alternative solutions might be identified for certain steps in the workflow based on different conditions. In addition, we leveraged the collective experience of our team working on health care IT projects to further refine and optimize the FHIR PG.

Finally, as an example, we arbitrarily selected a use case outside the scope of the reviewed articles and mapped it back to the FHIR PG to demonstrate the effectiveness and generalizability of the PG.

Results

Article Classification

Data Standardization
Data standardization typically involves two main steps: (1) defining profiles based on the data exchange requirements of interoperable services and (2) filling these profiles with the corresponding exchange data.

The base FHIR specification provides foundational resources applicable to various health care contexts. However, health care services often exhibit significant variability across different jurisdictions. Therefore, the base FHIR specification typically requires further adaptation, known as profile definition, to suit specific application contexts. Profile definition mainly encompasses three aspects: (1) rules about which resource elements to use and what additional elements to add to the base specification, (2) rules about which terminologies to use in
particular elements, and (3) the restricted value range and cardinality of the elements.

Table 1 lists the typical profile definitions and the corresponding FHIR foundational resources discussed in the reviewed articles. As shown, these articles cover a wide range of categories, including genomics [10-14], imaging [15-17], cancer [18-20], diabetes [21,22], COVID-19 [23,24], infections [25], electrocardiography [26], screening [27], and allergy [28].

There are typically 2 approaches to filling the profiles with exchange data. One is redesigning the database to align with the FHIR resource structure, and the other is mapping data from an existing CDM-based legacy system to the FHIR-based system. Table 2 lists relevant articles discussing the latter approach. These articles could roughly fall into 7 groups based on the types of source CDMs. The groups include informatics for integrating biology and the bedside [29,30], Observational Medical Outcomes Partnership (OMOP) [31,32], OpenEHR [33,34], HL7 version 2 [35], variant call format [36], free text or arbitrary proprietary data [37,38], and multisource [39-42]. Multisource refers to cases where multiple CDMs are involved. For example, the study by Lenert et al [40] focused on transforming data from the OMOP and Patient-Centered Outcomes Research Network to FHIR. The study by Pfaff et al [39] aimed to transform data from informatics for integrating biology and the bedside, OMOP, and Patient-Centered Outcomes Research Network to FHIR. The study by Prud’hommeaux et al [41] compared 3 methods for transforming data from various source CDMs into FHIR. The study by Kiourtis et al [42] proposed a resource description framework transformation toolkit to combine FHIR and non-FHIR data.

The studies in Table 2 indicate that the transformation from a specific CDM type to FHIR typically involves a 2-step mapping process: model mapping and element mapping. Model mapping establishes a relationship between the original data model and the FHIR resource. Element mapping comprises 2 parts, key mapping and value mapping, which define how to map the data fields from the source CDM to the corresponding fields in the FHIR resources. The mapping rules observe the consensus-mapping relationships established by domain experts. These experts analyzed the semantic and structural differences between the source CDMs and FHIR and determined the appropriate mappings to ensure accurate and meaningful data transformation. Although current data transformation approaches intend to support specific source data and target FHIR resource types, it is worth noting that ongoing research and advancements in domain-based applied artificial intelligence, including natural language processing and deep learning, hold great potential for developing more generalized data transformation algorithms.

As highlighted in previous studies, the granularity of data plays a crucial role in data standardization. When the granularity of the source data is finer than that of the target data, there is potential for information loss during the transformation process: the severity of information loss increases with the extent of the granularity gap.
<table>
<thead>
<tr>
<th>Theme and study, year</th>
<th>Involved Fast Healthcare Interoperability Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genomics</strong></td>
<td></td>
</tr>
<tr>
<td>Murugan et al [10], 2021</td>
<td>DiagnosticReport, Specimen, ServiceRequest, Observation, and Task</td>
</tr>
<tr>
<td>Seong et al [11], 2021</td>
<td>MolecularSequence</td>
</tr>
<tr>
<td>Alterovitz et al [12], 2020</td>
<td>DiagnosticReport, ServiceRequest, and Observation</td>
</tr>
<tr>
<td>Klopfenstein et al [13], 2021</td>
<td>Questionnaire and Document</td>
</tr>
<tr>
<td>Khalifa et al [14], 2021</td>
<td>Patient, PractitionerRole, Organization, Specimen, ServiceRequest, Media, RiskAssessment, Task, MedicationRequest, CarePlan, DeviceRequest, NutritionOrder, SupplyRequest, and RequestGroup</td>
</tr>
<tr>
<td><strong>Imaging</strong></td>
<td></td>
</tr>
<tr>
<td>Kohli et al [15], 2018</td>
<td>Patient, DiagnosticReport, ImagingStudy, AllergyIntolerance, Condition, MedicationOrder, Specimen, Organization, Practitioner, and Medication</td>
</tr>
<tr>
<td>Madrigal and Le [16], 2021</td>
<td>Media</td>
</tr>
<tr>
<td>Boufahja et al [17], 2021</td>
<td>Observation</td>
</tr>
<tr>
<td><strong>Cancer</strong></td>
<td></td>
</tr>
<tr>
<td>Zong et al [18], 2021</td>
<td>Observation and DiagnosticReport</td>
</tr>
<tr>
<td>Gonzalez-Castro et al [19], 2021</td>
<td>Observation, Device, FamilyMemberHistory, AllergyIntolerance, Condition, Patient, MedicationStatement, Encounter, Questionnaire, QuestionnaireResponse, and Procedure</td>
</tr>
<tr>
<td>Zong et al [20], 2020</td>
<td>QuestionnaireResponse</td>
</tr>
<tr>
<td><strong>Diabetes</strong></td>
<td></td>
</tr>
<tr>
<td>Ludmann et al [21], 2020</td>
<td>Observation</td>
</tr>
<tr>
<td>Glachs et al [22], 2020</td>
<td>Procedure, ProcedureRequest, Communication, Appointment, Observation, Condition, CommunicationRequest, Device, Encounter, Composition, Goal, Order, OrderResponse, MedicationAdministration, MedicationOrder, Organization, Patient, Practitioner, RiskAssessment, QuestionnaireResponse, Basic, and Parameters</td>
</tr>
<tr>
<td><strong>COVID-19</strong></td>
<td></td>
</tr>
<tr>
<td>Bauer et al [23], 2021</td>
<td>Questionnaire</td>
</tr>
<tr>
<td>Sass et al [24], 2020</td>
<td>Procedure, Observation, Condition, DiagnosticReport, Procedure, Consent, Immunization, MedicationStatement</td>
</tr>
<tr>
<td><strong>Infections</strong></td>
<td></td>
</tr>
<tr>
<td>Shivers et al [25], 2021</td>
<td>Consent, Coverage, DeviceUseStatement, Encounter, HealthcareService, Medication, MedicationAdministration, MedicationStatement, Observation, Patient, Practitioner, Procedure, ServiceRequest, and Specimen</td>
</tr>
<tr>
<td><strong>Electrocardiogram</strong></td>
<td></td>
</tr>
<tr>
<td>Benhamida et al [26], 2020</td>
<td>Observation</td>
</tr>
<tr>
<td><strong>Neonatal screening</strong></td>
<td></td>
</tr>
<tr>
<td>Bathelt et al [27], 2020</td>
<td>Patient, ServiceRequest, DiagnosticReport, Contract, Organization, and Practitioner</td>
</tr>
<tr>
<td><strong>Allergy</strong></td>
<td></td>
</tr>
<tr>
<td>Lenivtceva and Kopanitsa [28], 2021</td>
<td>AllergyIntolerance</td>
</tr>
</tbody>
</table>
Table 2. Data migration from the existing clinical data model to Fast Healthcare Interoperability Resources.

<table>
<thead>
<tr>
<th>Study, year</th>
<th>Clinical data model of the source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boussadi and Zapletal [29], 2017; Wagholikar et al [30], 2017</td>
<td>Informatics for integrating biology and the bedside</td>
</tr>
<tr>
<td>Jiang et al [31], 2017; Fischer et al [32], 2020</td>
<td>Observational Medical Outcomes Partnership</td>
</tr>
<tr>
<td>Ladas et al [33], 2022; Fette et al [34], 2020</td>
<td>OpenEHR</td>
</tr>
<tr>
<td>Xiao et al [35], 2021</td>
<td>HL7® version 2</td>
</tr>
<tr>
<td>Dolin et al [36], 2021</td>
<td>Variant call format</td>
</tr>
<tr>
<td>Peterson et al [37], 2020; Wang et al [38], 2020</td>
<td>Free text or arbitrary proprietary</td>
</tr>
<tr>
<td>Lenert et al [40], 2021; Pfaff et al [39], 2019; Prud’hommeaux et al [41], 2021; Kiourtis et al [42], 2020</td>
<td>Multisource</td>
</tr>
</tbody>
</table>

*HL7: Health Level 7.*

**Data Management**

Data management includes data storage and data exposure. Although FHIR defines 5 approaches for data exposure, including RESTful API, messaging, documents, services, and persistent store, recent articles predominantly chose to expose data in the form of APIs because of the rapid growth of the APIs economy. There are typically 2 methods for data management: developing a customized FHIR warehouse to store and manage FHIR data or selecting a mature third-party warehouse to handle the task.

Table 3 shows various data management choices and their corresponding targets. It reveals that developing a customized FHIR warehouse to maintain FHIR data often requires meeting some special service requirements. For instance, the customized FHIR warehouse developed by Demurjian et al [43] aimed to enable sensitivity and multilevel security controls. The one developed by Chatterjee et al [44] and Saripalle et al [45] served to integrate with specific terminology. The one developed by Ruminski et al [46], Saripalle [47], and Yu et al [48] intended to support multiple Internet of Things protocols. Finally, the one discussed in the studies by Khvastova et al [49], Drudi et al [50], Lee et al [51], Tanaka and Yamamoto [52], Cheng et al [53], Semenov et al [54], and Gruendner et al [55] was used to support data preprocessor plug-ins.

On the other hand, several mature third-party platforms are available for managing FHIR data. In 2018, a total of 6 technology giants, including Amazon, Microsoft, Google, IBM, Oracle, and Salesforce, jointly announced that they would be committed to removing the barriers to adopting health care interoperability technologies, particularly those enabled through the cloud [56]. All these companies have launched FHIR data management platforms, providing FHIR data APIs for resource operations. Users of these platforms can store their data as FHIR resources and use the data APIs offered by the cloud platform for service development. For instance, the studies by Shi et al [57], Zampognaro et al [58], Ploner and Prokosch [59], and Kamel and Nagy [60] chose cloud warehouses, and the study by Mandl et al [61] chose an on-premises warehouse to rapidly deploy an FHIR development environment.

The abovementioned analysis highlights that choosing between proprietary and third-party warehouses involves trade-off considerations. Maintaining FHIR data through a proprietary warehouse offers 2 advantages: better privacy and greater flexibility for functional expansion. However, developing a proprietary warehouse requires extensive knowledge of FHIR standards and software development skills, resulting in higher costs. On the other hand, relying on third-party platforms offers the advantages of lower cost and higher implementation efficiency. However, storing sensitive data in a third-party warehouse, with the service provider not being the data owner, raises security and privacy concerns.
Table 3. Fast Healthcare Interoperability Resources (FHIR) data management methods and their corresponding targets.

<table>
<thead>
<tr>
<th>Method and study, year</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Develop FHIR warehouse</strong></td>
<td>Support lattice-based access control</td>
</tr>
<tr>
<td>Demurjian et al [43], 2020</td>
<td>Integrate with specific terminologies</td>
</tr>
<tr>
<td>Chatterjee et al [44], 2022; Saripalle et al [45], 2020</td>
<td>Support multiple IoT protocols</td>
</tr>
<tr>
<td>Ruminski et al [46], 2016; Saripalle [47], 2019; Yu et al [48], 2021</td>
<td>Support data preprocess plug-ins</td>
</tr>
<tr>
<td>Khvastova et al [49], 2020; Dridi et al [50], 2020; Lee et al [51], 2020; Tanaka and Yamamoto [52], 2020; Cheng et al [53], 2021; Semenov et al [54], 2019; Gruendner et al [55], 2021</td>
<td><strong>Use third-party FHIR warehouse</strong></td>
</tr>
<tr>
<td>Shi et al [57], 2021; Zampognaro et al [58], 2021</td>
<td>Rapidly deploy a development environment through a cloud FHIR warehouse</td>
</tr>
<tr>
<td>Ploner and Prokosch [59], 2020; Kamel and Nagy [60], 2018</td>
<td>Rapidly deploy a development environment through an on-premises FHIR warehouse</td>
</tr>
<tr>
<td>Mandl et al [61], 2020</td>
<td></td>
</tr>
</tbody>
</table>

IoT: Internet of Things.

**Data Integration**

Data integration plays a vital role in health care across various domains, including service delivery, public health management, and clinical medicine or health care economics research, enabling better decision-making and improving overall health care outcomes. In service delivery, data integration is crucial for coordinating multiple IT systems, including the hospital information system (HIS), laboratory information system, picture archiving and communication system, EMR, and EHR. In public health, local governments need to collect health-related data within their jurisdictions to monitor regional health status and effectively address public health issues. In clinical medicine or health care economics research, it is essential to obtain data from diverse domains to conduct comprehensive studies and analyses.

There are 2 typical modes of FHIR data integration, as listed in Table 4.

The first mode of data integration is using an integrated service platform (ISPF). The ISPF is an orchestrating platform offering a series of API management functions such as API registration, API calling authorization, and API routing forward. Organizations wishing to exchange data through the ISPF must register their APIs on the platform. Other organizations can search for the appropriate APIs on the ISPF and make API calls. The ISPF performs API calling authorization to verify the calling rights and then routes the API calls to the respective organization to which the API belongs. This process facilitates data exchange among multiple organizations [62-75]. An example of this mode is the efficient transfer of medical records when a patient referral occurs.

The second mode of data integration is by way of interoperable apps. Different architectures can be selected for different application scenarios. In the case of apps with specific functions, such as statistics and analysis, SMART on FHIR would be a more efficient option [76-85]. In the case of apps with customized functions, such as supporting microservice architecture or blockchain architecture, customized architecture apps would be a more suitable option [86-94].

Table 4. Fast Healthcare Interoperability Resources data integration modes and their corresponding application scenarios.

<table>
<thead>
<tr>
<th>Interoperable modes and study, year</th>
<th>Applied scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Integrated service platform</strong></td>
<td>Control exchange data through APIs for service coordination among multiple organizations.</td>
</tr>
<tr>
<td>Nan et al [62], 2021; Taechoyotin et al [63], 2021; Maxi and Morocho [64], 2022; Rosenau et al [65], 2022; Corici et al [66], 2020; Papaioannou et al [67], 2021; Hidayat and Hermanto [68], 2020; Sloane et al [69], 2021; Mukhiya and Lamo [70], 2021; Gruendner et al [71], 2022; Gruendner et al [72], 2020; Park et al [73], 2022; Ziminski et al [74], 2021; De et al [75], 2021</td>
<td></td>
</tr>
<tr>
<td><strong>App</strong></td>
<td>Substitutable Medical Applications and Reusable Technologies app: apps with specific functions, such as statistics and analysis.</td>
</tr>
<tr>
<td>Suraj et al [76], 2022; Michaels et al [77], 2021; Curran et al [78], 2020; Thayer et al [79], 2021; Karhade et al [80], 2021; Wesley et al [81], 2021; Burkhardt et al [82], 2021; Hoffman et al [83], 2017; Stoldt and Weber [84], 2020; Stoldt and Weber [85], 2021</td>
<td>Other architecture app: apps with customized functions, such as supporting microservice and blockchain architecture.</td>
</tr>
<tr>
<td>Alamri et al [86], 2021; George and Chacko [87], 2022; Gulden et al [88], 2021; Chaves et al [89], 2021; Bae and Yi [90], 2022; Bettoni et al [91], 2021; Weber et al [92], 2020; Sfat et al [93], 2021; Mohammed et al [94], 2021</td>
<td></td>
</tr>
</tbody>
</table>

API: application programming interface.
We present an FHIR practice design in Figure 3, which defines the responsibilities of stakeholders and outlines the complete practice process from data to services.

Figure 3. The general Fast Healthcare Interoperability Resources (FHIR) practice guideline—practice design. API: application programming interface; IG: Implementation Guide; ISPf: integrated service platform.

**IGs Editing Group**

The first stakeholder involved in the process is the IGs editing group, usually coordinated by a government agency or an institution with significant influence in the ecosystem. The primary responsibility of this group is to define the data and service models and release the IGs. The detailed processes are as follows. First, select necessary FHIR resources based on the service requirements. Second, for specific requirements beyond the scope of the original FHIR resources, the group needs to customize resource structure by FHIR profile. Profile generally involves 3 aspects: extending the data field by FHIR extension, linking the local CodeSystem to the CodeableConcept field of FHIR resources, and restricting the cardinality and ValueSet of FHIR foundational resource. The customized resources created by the profile enable better alignment with the data requirements in various scenarios. After completing the data unification task, the IGs editing group moves on to the unification of services workflow, which involves specifying the implementation steps in the workflow and standardizing the corresponding APIs. Ultimately, the abovementioned data and workflow specifications are integrated to form the comprehensive FHIR IGs that health care IT system vendors can adopt.

**Health Care IT System Vendor**

The second type of stakeholder is the health care IT system vendor, responsible for developing and maintaining systems, such as the HIS, laboratory information system, and picture archiving and communication system. First, the vendor must implement the IGs published by the IGs editing group, which involves standardizing data by redesigning the database according to the FHIR resource structure and mapping data from existing CDM-based legacy systems to FHIR-based systems. Second, with RESTful APIs, the vendor has 2 options for data exposure: either maintaining the FHIR data and APIs themselves or selecting a mature third-party platform. FHIR APIs must be exposed to support resource-level operations regardless of the chosen option.

It is worth pointing out that in terms of data exposure, FHIR defines 5 different approaches, and each data exposure approach has a different data integration method; it would be a lengthy discussion if all approaches are considered. To make FHIR PG more compatible with current technology stacks, we chose to focus on RESTful API rather than on other approaches in this study.

**Health Care Application Developer**

The third stakeholder involved in this process is the health care application developer, responsible for developing interoperable services using open FHIR APIs. As described in the Data Integration section, there are 2 typical modes. The first is to develop an ISPf, that is, an orchestrating platform, for service interoperability. The ISPf manages open APIs registered by each organization and enforces access specifications such as IGs, profiles, and workflows. Any IT systems accessing the ISPf and exchanging data must comply with these specifications. When an IT system needs to access multiple ISPfs, it must support multiple specifications. In such cases, the IT system can deploy an adapter above its native database to comply with various specifications. When the IT system acts as a producer, it reads the corresponding specifications from the adapter to expose the data. When it acts as a consumer, it reads the corresponding specifications from the adapter to parse data. The second mode is to develop specific apps that cater to specific requirements. For example, an app built with SMART on FHIR
architecture supports a flexible and switchable application ecosystem.

**Beneficiary**

Beneficiaries such as hospitals, patients, public health institutions, and research institutions can benefit from high-quality FHIR-based health care services. For instance, if there is a need to exchange data through APIs to facilitate service coordination among multiple organizations, they can easily access the ISPf to fulfill this objective. Alternatively, they can choose a suitable app from the application gallery that caters to their needs and functions.

**The Development Architecture for the Practice Design**

**Overview**

We presented a 3-stage development architecture for the practice design, as shown in Figure 4. In addition, we compiled a list of commonly used tools in Table 5 to support the development process.

*Figure 4. The general Fast Healthcare Interoperability Resources (FHIR) practice guideline—the development architecture for the practice design. IG: Implementation Guide; ISPf: integrated service platform; SMART: Substitutable Medical Applications Reusable Technologies.*
### Data Standardization

<table>
<thead>
<tr>
<th>Tool and description</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data standardization</strong></td>
<td></td>
</tr>
<tr>
<td>HAPI(^b) FHIR(^b)</td>
<td>This tool provides Java API(^c) for HL7(^d) FHIR clients and servers [95]</td>
</tr>
<tr>
<td>IG(^e) Auto-Builder</td>
<td>An IG publishing tool that makes your IGs to be visible on the internet [96] [97]</td>
</tr>
<tr>
<td>Firely Forge</td>
<td>The official FHIR tool for managing FHIR profiles [98]</td>
</tr>
<tr>
<td>Firely Terminal</td>
<td>A cross-platform command line tool with a range of commands for working with FHIR resources and installing and publishing FHIR packages [99]</td>
</tr>
<tr>
<td><strong>Data management</strong></td>
<td></td>
</tr>
<tr>
<td>Firely Facade</td>
<td>A special type of plug-in that registers services to access the existing data repository. It speaks FHIR in the front-end and talks directly to native data in the back-end [100]</td>
</tr>
<tr>
<td>FHIR Works on AWS(^f)</td>
<td>A framework used to deploy an FHIR server on AWS [101]</td>
</tr>
<tr>
<td>FHIR server for Azure</td>
<td>An open-source implementation of FHIR specification designed for the Microsoft cloud [102]</td>
</tr>
<tr>
<td>GCP(^g) Healthcare API</td>
<td>A cloud application that accelerates health care solution development with fully managed, enterprise-scale HL7 FHIR, HL7 version 2, and DICOM(^b) APIs [103]</td>
</tr>
<tr>
<td>IBM FHIR server</td>
<td>An open-source Java solution that supports the processing, validation, and storage of health care data according to the HL7 FHIR specification [104]</td>
</tr>
<tr>
<td>Oracle Healthcare Data Repository</td>
<td>The foundation of a health care information exchange platform that makes health care data more useful by supporting the integration and operation of a full spectrum of health care applications [105]</td>
</tr>
<tr>
<td>Health Cloud</td>
<td>A tool that combines clinical and nonclinical customer data to drive efficiencies in health [106]</td>
</tr>
<tr>
<td><strong>Data integration</strong></td>
<td></td>
</tr>
<tr>
<td>Spring Cloud Gateway</td>
<td>This tool provides an API Gateway built on top of the Spring Ecosystem [107]</td>
</tr>
<tr>
<td>Redis</td>
<td>This tool provides access to mutable data structures via a set of commands sent using a server-client model with TCP(^i) sockets and a simple protocol [108]</td>
</tr>
<tr>
<td>Validator</td>
<td>The HAPI FHIR Validator API is a simple REST(^j) API to validate the structure and content of an FHIR object [109]</td>
</tr>
<tr>
<td>Elasticsearch</td>
<td>A distributed, RESTful search and analytics engine is at the heart of the Elastic Stack [110]</td>
</tr>
<tr>
<td>OpenID</td>
<td>An open standard and decentralized authentication protocol promoted by the nonprofit OpenID Foundation [111]</td>
</tr>
<tr>
<td>OAuth</td>
<td>An open protocol to allow secure authorization in a simple and standard method from web, mobile, and desktop applications [112]</td>
</tr>
<tr>
<td>SMART(^k)</td>
<td>Define a workflow that an application can use to securely request access to data and then receive and use that data [113]</td>
</tr>
</tbody>
</table>

\(^{\text{a}}\)HAPI: Health Level 7 application programming interface.
\(^{\text{b}}\)FHIR: Fast Healthcare Interoperability Resources.
\(^{\text{c}}\)API: application programming interface.
\(^{\text{d}}\)HL7: Health Level 7.
\(^{\text{e}}\)IG: Implementation Guide.
\(^{\text{f}}\)AWS: Amazon Web Services.
\(^{\text{g}}\)GCP: Google Cloud Platform.
\(^{\text{h}}\)DICOM: Digital Imaging and Communications in Medicine.
\(^{\text{i}}\)TCP: transmission control protocol.
\(^{\text{j}}\)REST: representational state transfer.
\(^{\text{k}}\)SMART: Substitutable Medical Applications Reusable Technologies.

### Data Standardization

In the data standardization development stage, several components are defined to ensure the consistent use of codes within a specific context. The terminology system comprises essential resources such as CodeSystem, ValueSet, and ConceptMaps. These resources establish a framework for determining which codes can be used. Furthermore, the
conformance system includes resources such as StructureDefinition, OperationDefinition, CapabilityStatement, and ImplementationGuide. These resources are crucial in creating profiles and IGs that adhere to a specific exchange framework. As mentioned in the **Data Standardization** section, the granularity of the data plays a crucial role in information loss. Pfaff et al [39] pointed out that information loss can be avoided by defining custom values or extensions during the data standardization stage. By incorporating custom values or extensions defined in this stage, it is possible to capture and preserve the finer-grained information that is likely to be lost during the transformation process.

During this process, developers can use various tools to facilitate efficient data standardization. The HL7 API (HAPI) FHIR offers a Java API for developing HL7 FHIR clients and servers. Forge serves as a management tool for FHIR profiles. The Firely Terminal, a cross-platform command line tool, provides a wide array of commands for working with FHIR resources and installing and publishing FHIR packages. IG Auto-Builder is another helpful tool that simplifies the creation and publication of IGs, available on the internet [96].

Ultimately, the data standardization stage would generate a set of IGs to ensure consistency and conformity in implementing higher-level services.

**Data Management**

Various situations can arise in the data management development stage, each bringing different challenges. These situations can fall into 3 options.

The first is to develop an FHIR-native warehouse that the health care IT system vendor manages. In this scenario, the vendor assumes responsibility for designing, implementing, and maintaining the warehouse.

The second is to select a well-established third-party warehouse, such as FHIR Works on Amazon Web Services, IBM FHIR Server, Google Cloud Platform Healthcare API, FHIR Server for Azure, Health Cloud, and Oracle Healthcare Data Repository, to store and explore the FHIR APIs. This approach allows vendors to leverage the capabilities of mature third-party warehouses for FHIR API functionality.

The third is to provide FHIR data using plug-ins. In this scenario, vendors retain their existing data infrastructure and use plug-ins to facilitate data transformation from its native format to the FHIR format. A tool called Facade is available to facilitate this mapping process.

As discussed in the **Data Standardization** section, the discrepancy in granularity between different systems can lead to potential information loss. To mitigate this issue, developers can incorporate a mapping log within the transformer component. When encountering a granularity gap during the mapping process, the mapping log captures and records the lost information, associating it with the corresponding target resource ID. This mapping log serves as a reference for any subsequent services or systems requiring detailed information about the mapping process. If the overlying services need to retrieve the lost information, they can make a request based on the resource ID recorded in the mapping log. This measure allows them to access the details lost during the initial mapping, ensuring that the required information is preserved and available for further analysis or processing.

Ultimately, the data management stage generates a series of FHIR APIs. These APIs serve as a foundation for data exploration and form the backbone of the infrastructure required for high-level services.

**Data Integration**

Two types of interoperable services are commonly used in the data integration development stage.

The first type is the IS Pf, which enables interoperability among multiple organizations. The IS Pf comprises 4 key components: gateway, validator, flow control, and log system. The gateway, built by the Spring Cloud Gateway, is responsible for API authorization and forwarding API requests between organizations. The validator ensures that the structure and content of the API data comply with the FHIR object defined in IGs. The HAPI FHIR Validator can build this functionality. The flow control component is designed to limit the number of simultaneous API calls to ensure a stable operation. Redis can effectively fulfill the flow control requirements. As IS Pf manages multiple organizations and facilitates data exchange, maintaining a comprehensive log system is crucial for history tracking and auditing. Elasticsearch, a powerful search and analytics engine, can be used to develop the log system within the IS Pf, enabling efficient storage and retrieval of API call records.

The second type of interoperable service is represented by apps built by the SMART on FHIR architecture [114]. This architecture consists of 3 key components: the resource server, authorization server, and the SMART on FHIR apps. The resource server is an access layer between the data management layer and the SMART on FHIR apps. The authorization server manages multiple organizations and facilitates data exchange, maintaining a comprehensive log system is crucial for history tracking and auditing. Elasticsearch, a powerful search and analytics engine, can be used to develop the log system within the IS Pf, enabling efficient storage and retrieval of API call records.

**Use Case**

We arbitrarily selected a use case that was in addition to the reviewed articles. Portugal et al [115] designed a smart bed infrastructure with an HIS using FHIR. We mapped it back to the FHIR PG to demonstrate PG’s effectiveness and generalizability. In this case, the roles and responsibilities can be mapped to the FHIR PG–practice design. The authors and their research partners formed an IGs editing group to define IGs consisting of profiles and workflows. The profiles were derived from foundational FHIR resources such as Observation, Device, and ServiceRequest. The workflows defined the frequency at which the smart bed would collect vital signs from the smart bed. Subsequently, the authors’ team, acting as a health care IT system vendor, developed a gateway that gathers raw data from sensors and converts it into FHIR for transmission. Although they did not discuss the final applications in detail, it can be...
inferred that health care application developers can build a better smart bed monitor based on their infrastructure.

The development architecture described in this paper can also be mapped back to the FHIR PG–development architecture. In the data standardization stage, the authors used the HAPI FHIR for HTTP processing, parsing and serialization, and FHIR REST semantics. It provided a bare-bones structure to build the API. In the data management stage, the authors developed a fog server as a gateway between the smart bed and HIS. This fog server is responsible for collecting raw data from the HIS, transforming it into the FHIR format, and facilitating its integration into the FHIR ecosystem. Finally, in the data integration stage, the authors enabled the HIS software to monitor patient procedures and flows, accompanied by the OAuth2 protocol for secure API communication.

**Discussion**

**Principal Findings**

FHIR has shown significant advantages in facilitating interoperability among health IT systems compared with established international standards. However, there are challenges in large-scale implementation and promotion, particularly in different countries. First, countries without incentive policies to encourage FHIR research and implementation may exhibit less enthusiasm for adopting FHIR standards. Second, the lack of a suitable infrastructure to support the implementation process can result in high costs associated with FHIR adoption. Third, the foundational resources provided by FHIR may not directly align with the specific service requirements in different regions, necessitating additional customization processes.

The following steps must be taken to address these challenges. First, it is crucial to have government policies that encourage the evolution and adoption of health care data standards. These policies can stimulate the enthusiasm and investment of stakeholders in the health care ecosystem to promote FHIR implementation on a larger scale. Second, strengthening the infrastructure helps reduce the cost and complexity associated with FHIR adoption, which includes developing services such as FHIR data storage, data standard quality control, and managed services for data operations. Third, FHIR profiles and workflows should be defined to address the specific requirements and characteristics of local health systems. By tailoring FHIR IGs to match the needs of different regions, the gap between FHIR foundational resources and specific service requirements can be bridged.

FHIR holds significant potential in standardizing health care data and promoting service interoperability among health care institutions. Its adoption can drive the transformation of the health care service model and enhance the overall quality of health care services. With the growing recognition of the benefits of FHIR and its demonstrated impact on health care interoperability, more stakeholders are expected to actively participate in enriching its implementation. This collective effort would lead to the emergence of extensive health care service innovations, further enhancing the delivery of high-quality health care services.

**Limitations**

There are a few current limitations when applying the FHIR PG: (1) PG is derived from the waterfall model that follows a sequential and linear approach. Each step must be completed before proceeding to the next step. Therefore, it is time-consuming and costly to return and modify the previous steps if changes are necessary during the development process. (2) Although PG emphasizes the achievement of interoperability, it leaves out the security discussion. Developers must incorporate additional security mechanisms into PG–development architecture to ensure secure interoperation among multiple organizations.

**Conclusions**

Owing to the unique characteristics of FHIR, including comprehensive coverage of data definitions, substantial flexibility of data exchange, explicit semantics, and many available open-source tools, FHIR-based services have attracted strong interest from stakeholders in the health care ecosystem. Current studies reveal that many institutions, such as hospitals, regulators, and researchers, have already begun collaborations in actively building FHIR foundational frameworks or application use cases. After conducting the latest literature review, we proposed a general FHIR PG to bridge the gap between FHIR IGs and the practice of building usable services. This PG helps stakeholders identify their participant roles, manage the scope of responsibilities, and develop relevant modules, which we believe would effectively facilitate the application and promotion of HL7 FHIR standards across the health care ecosystem.

**Conflicts of Interest**

None declared.

**References**

2. Protti D. Integrated care, information management and information technology in Canada: have we made any progress in the past 12 years? Healthc Q 2013;16(1):54-59. [Medline: 24863308]


7. FHIR®: it is time to shine for the interoperability standard. Enovacom. URL: https://www.enovacom.com/resource/fhir-it-is\_time-to-shine-for-the-interoperability-standard [accessed 2022-04-01]


https://medinform.jmir.org/2023/1/e44842

JMIR Med Inform 2023 | vol. 11 | e44842 | p.102

(page number not for citation purposes)


95. HAPI FHIR - Java API for HL7 FHIR. GitHub. URL: https://github.com/hapifhir/hapi-fhir [accessed 2023-08-01]

96. Welcome to FHIR®. Health Level Seven International and Fast Healthcare Interoperability Resources. URL: https://build.fhir.org/ [accessed 2023-08-01]


98. Forge. SIMPLIFIER.NET. URL: https://simplifier.net/forge [accessed 2023-08-01]

99. Firely terminal. SIMPLIFIER.NET. URL: https://simplifier.net/firely-terminal [accessed 2023-08-01]

100. FirelyTeam. GitHub. URL: https://github.com/FirelyTeam/VonkFacadeStarter [accessed 2023-08-01]

101. FHIR Works on AWS deployment. GitHub. URL: https://github.com/aws-labs/fhir-works-on-aws-deployment [accessed 2023-08-01]

102. FHIR server for Azure. GitHub. URL: https://github.com/microsoft/thir-server [accessed 2023-08-01]

103. Cloud healthcare API. Google Cloud. URL: https://cloud.google.com/healthcare-api [accessed 2023-08-01]

104. IBM Watson Health is now Merative. IBM. URL: https://ibm.com/products/thir-server [accessed 2023-08-01]


106. Health cloud. Salesforce. URL: https://www.salesforce.com/products/health-cloud/overview/ [accessed 2023-08-01]


110. Elasticsearch. GitHub. URL: https://github.com/elastic/elasticsearch [accessed 2023-08-01]
111. OpenID homepage. OpenID. URL: https://openid.net/ [accessed 2023-08-01]
112. OAuth 2.0 homepage. OAuth 2.0. URL: https://oauth.net/ [accessed 2023-08-01]
113. SMART on FHIR. GitHub. URL: https://github.com/smart-on-fhir [accessed 2023-08-01]


Abbreviations

API: application programming interface
CDA: Clinical Document Architecture
CDM: clinical data model
EHR: electronic health record
EMR: electronic medical record
FHIR: Fast Healthcare Interoperability Resources
HAPI: Health Level 7 application programming interface
HIS: hospital information system
HL7: Health Level 7
IG: Implementation Guide
ISPF: integrated service platform
OMOP: Observational Medical Outcomes Partnership
PG: practice guideline
SMART: Substitutable Medical Applications Reusable Technologies

©Jingwen Nan, Li-Qun Xu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 21.08.2023. 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.
Viewpoint

Developing a Capsule Clinic—A 24-Hour Institution for Improving Primary Health Care Accessibility: Evidence From China

Dongliang Li1*, PhD; Rujia Zhang2*, BPA; Chun Chen2*, PhD; Yunyun Huang2, BPA; Xiaoyi Wang2, BPA; Qingren Yang3, BPA; Xuebo Zhu3, PhD; Xiangyang Zhang3, MPA; Mo Hao1*, PhD; Liming Shui4*, MPH

1School of Public Health, Fudan University, Shanghai, China
2School of Public Health and Management, Wenzhou Medical University, Wenzhou, China
3Engineering Research Center of Intelligent Medicine (2016E10011), The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China
4Yinzhou District Health Bureau, Ningbo, China
*these authors contributed equally

Corresponding Author:
Liming Shui, MPH
Yinzhou District Health Bureau
No 1221, Bachelor Road, Shounan Street, Yinzhou District
Ningbo, 315199
China
Phone: 86 13967820698
Email: 707761065@qq.com

Abstract

Telehealth is an effective combination of medical service and intelligent technology. It can improve the problem of remote access to medical care. However, an imbalance in the allocation of health resources still occurs. People spend more time and money to access higher-quality services, which results in inequitable access to primary health care (PHC). At the same time, patients’ usage of telehealth services is limited by the equipment and their own knowledge, and the PHC service suffers from low usage efficiency and lack of service supply. Therefore, improving PHC accessibility is crucial to narrowing the global health care coverage gap and maintaining health equity. In recent years, China has explored several new approaches to improve PHC accessibility. One such approach is the capsule clinic, an emerging institution that represents an upgraded version of the internet hospital. In coordination with the United Nations, the Yinzhou district of Ningbo city in Zhejiang, China, has been testing this new model since 2020. As of October 2022, the number of applications in Ningbo was 15, and the number of users reached 12,219. Unlike internet hospitals, the entire process—from diagnosis to prescription services—can be completed at the capsule clinic. The 24-hour telehealth service could also solve transportation problems and save time for users. Big data analysis can accurately identify regional populations’ PHC service needs and improve efficiency in health resource allocation. The user-friendly, low-cost, and easily accessible telehealth model is of great significance. Installation of capsule clinics would improve PHC accessibility and resolve the uneven distribution of health resources to promote health equity.

(JMIR Med Inform 2023;11:e41212) doi:10.2196/41212

KEYWORDS
primary health care; accessibility; capsule clinic; 24-hour clinic; big-data; China; United Nations; internet clinic

Background

Combining intelligent technology with health care has become increasingly popular [1,2]. Telehealth is the remote provision of clinical health care and health administration services via information and telecommunication technologies [3]. In exploring suitable methods to meet people’s increasing demands, countries are establishing intelligent medical models in telehealth to provide convenient health services [4] and solve geographical, temporal, and economic problems of accessibility to health care. Doctors use telehealth to transmit digital imaging, conduct video consultations, and make medical diagnoses. Telehealth began in the 20th century [5], with the advent of television, and has developed rapidly in the 21st century along with advancements in technology. Most people currently have access to basic devices, such as mobile phones and computers, which can be used to obtain telehealth services. With improved accessibility of health care through telehealth, individuals in rural areas and busy urban areas can often connect more easily with a provider. In the United States and the United Kingdom
telehealth is a popular trend. According to the National Health Service Long Term Plan in the United Kingdom, “digitally enabled care will go mainstream” [7].

Compared with other countries, China is a late starter in the field of telehealth. In recent years, China has been committed to applying internet technology to medical services. The internet hospital is a rapidly developing, new telehealth model that is gaining wide popularity [8]. The establishment of internet hospitals could help some citizens overcome temporal and geographical barriers to accessing traditional medical services [9]. However, the issue of accessibility to primary health care (PHC) services has not been completely resolved. Owing to urban-rural differences and unbalanced allocation of regional health human resources [10], the phenomenon of inequality in health resource allocation still exists [11]. Furthermore, as the population continues to age [12] and the number of chronic diseases grows [13], the demand for health services increases every year. Researchers have found that accessibility of PHC services has been challenged by insufficient funds [14], limited-service locations [15], poverty-stricken areas, and inconvenient transportation [16-18]. The use of PHC is inefficient because people rely on doctors in higher-tier hospitals [19]. Although the internet hospital is a promising public health tool because it could significantly increase access to health care for medically underserved populations, some obstacles remain. For example, people with low digital information literacy, especially older people and children, cannot easily use internet health care. Low digital literacy remains the key barrier to accessing intelligent health systems [20,21]. In order to use telehealth, such as internet hospitals, mobile devices have high requirements [22,23]. Inadequate communication facilities hinder telehealth services [24,25] for many people in remote areas, and many older people experience barriers to using smartphones. Internet hospitals must improve applications for older adults. In addition, patients also face the problem of waiting for drug delivery after receiving remote consultation services via the internet hospital, which could cause a loss of time [26,27]. Moreover, big data is not being used well in the allocation of drug resources; the advantages of big data analysis of population characteristics and rational allocation of drug resources cannot be realized in internet hospitals.

In many cases, although internet hospitals address the problem of unbalanced distribution of health resources, the shortcomings mentioned above must be addressed to provide access to PHC services. In the near future, the application of telehealth will not only increase convenience and access to health care but also reduce costs. Telehealth services should also be more user-friendly and implement low-cost devices to facilitate remote health care services. Intelligent medical service systems should be better integrated with the traditional medical models, and internet hospitals must be upgraded to meet the needs of all members of society.

Capsule Clinics in China

The accessibility of PHC services in China is usually considered in terms of geographical accessibility [28-30], economic accessibility, and temporal accessibility [31,32]. To that end, we propose the development of the Internet Hospital 2.0 in the era of big data, in the context of Ningbo, a city in China with an estimated population of over 9 million individuals [33]. Ningbo residents face challenges in accessing PHC services. They remain disappointed with the inconvenient primary care model and are anxious about the time and economic costs of daily medical visits, particularly during the COVID-19 pandemic. Ningbo is a coastal city that includes many islands and other remote areas. People living in remote villages or islands experience greater challenges in accessing and paying for PHC services.

In an attempt to combine the use of available public health data with the provision of easy access to health care, and toward the achievement of universal health coverage, the World Health Organization implemented a pilot project for developing a higher-quality and efficient medical and health service e-system. The Yinzhou district in Ningbo is one of the pilot locations. Yinzhou has created a specialized health service model in the Chinese context—the capsule clinic, an upgraded version of the internet hospital. The capsule clinic (20 m² in area) is a novel type of health service facility that began operating in 2020. It incorporates traditional medical models, including diagnosis and prescription. Capsule clinic services comprise 3 parts: health examination, consultation, and an intelligent pharmacy.

The special clinic, which relies on Zhejiang province’s excellence in digital reform and Yinzhou’s superior health care resources, is an emerging form of medical treatment facility that can provide residents with more convenient PHC services than those currently available. One advantage of advanced intelligent technology–based models is that they increase geographical accessibility to PHC. Capsule clinics are placed within communities to enable residents’ convenient access to PHC without leaving the community. The capsule clinic offers some important advantages, such as 24-hour services, comprehensive medical services that include the entire process from diagnosis to prescription services, and efficient allocation of health resource services. The capsule clinic overcomes some of the administrative shortcomings of the internet hospital.

In recent years, Zhejiang has attached great importance to the combination of intelligent technology and medical services, allowing grassroots residents to enjoy higher-quality and inclusive PHC services. Figure 1 shows that as of October 2022, there were 15 applications for capsule clinics in Yinzhou (Figure 1). According to China’s “one village, one health care room” principle, Yinzhou plans to establish over 190 capsule clinics across communities or villages and expand throughout China.
The Function of the Capsule Clinic

Figure 2 shows that several critical hardware and software components are included in the capsule clinic to ensure the necessary functions to meet the medical demands of residents living nearby. The capsule clinic has continuous medical services, including diagnosis and prescription, which fit the traditional medical model. Doctors can communicate with patients for remote health care consultations around the clock. Patients can access services on their own any time they need them and even gain access to essential medicine. In addition, patients can use the telehealth service of capsule clinics to consult doctors from higher-tier hospitals; this feature provides an opportunity for PHC service resources to be equitably deployed, especially for people in remote areas.

The capsule clinic, a product of the digital reform, has developed rapidly and is devoted to improving geographic accessibility to PHC. For each community, the capsule clinic will be conveniently located within 15 minutes of the community’s health service center, which is responsible for the full scope of the capsule clinic. At nearby capsule clinics, people located in remote areas (mountainous areas and islands) can access health resources equal to those available to urban residents. This measure is intended to benefit the public and solve issues of not only geographic accessibility but also temporal accessibility; with the capsule clinic, people do not have to miss work or travel long distances for medical appointments.

Although some internet hospitals and telehealth centers are also available 24 hours a day, capsule clinics are more helpful to residents because they provide comprehensive services. In the past, patients received their medications within a few days after receiving a web-based diagnosis. The capsule clinic takes advantage of intelligent technology to help residents save time in receiving their medications. After completing a remote consultation with the doctor, patients can easily access the capsule clinic’s intelligent pharmacy to pick up their medication. They can also refill prescriptions from offline doctors. Moreover, if further examination or tests are needed, patients can visit the higher-tier hospitals within the period determined by the capsule clinic doctors. If hospitalization is required, treatment can be transferred to the hospital, as determined by the capsule clinic doctors. The capsule system embodies the advantages of an integrated health care delivery system. It facilitates convenient access to health care and makes good use of PHC resources.

Big data analysis and artificial intelligence can assist doctors in a variety of ways, such as by detecting lesions and improving diagnostic efficiency. They also play a role in improving medical services and easing constraints on medical resources. With the deep integration of digital technology and medical measures, as well as comprehensive improvements in patients’ health literacy, diversified medical and health service models can provide convenience for doctors and patients, thus improving the efficiency of diagnosis and treatment. Big data allows for systematic analysis of the relevant characteristics and medical needs of people in each region and the subsequent reasonable allocation of drug resources. For instance, in a community with a wide distribution of older or chronic patient populations, the capsule clinic would be equipped with more drugs to treat geriatric issues and chronic diseases. One benefit of a big data–based system is the reduction in medical processing time and distance, which significantly improves the efficiency and quality of the medical care provided and facilitates optimum resource allocation.
Comparing the Capsule Clinic With the Traditional Hospital and Internet Hospital

Overview

The capsule clinic has the physical appearance and essential equipment of a traditional medical institution. It also includes 24-hour remote consultation, integrated consultation and pharmacy services, optimal health resource allocation, and health management functions. The clinic has the strong support and oversight of a traditional hospital, and offers web-based services for relatively simple problems that do not necessitate a visit to the hospital.

Digital technology will soon be deeply integrated with medical treatment and will provide important support for medical professionals’ diagnosis and treatment decisions. However, the internet hospital model \cite{34,35} still has much room for improvement, for example, in terms of medical insurance and the allocation of medical resources. With the support of big data technology, capsule clinics analyze the regional group characteristics, realize the reasonable allocation of essential drug resources, and meet the health needs of different populations. Even more importantly, capsule clinics may alleviate shortages in human resources chronically faced by primary-level medical institutions. The capsule clinic allows community residents to conveniently purchase medication 24 hours a day, taking pressure off in-person facilities. Remote consultation can also reduce pressures on grassroots medical staff and, thus, improve work efficiency. Figure 3 provides comparison of 3 medical models.

In order to understand the user experience of the capsule clinic during the implementation process, we conducted the research during November and December 2021. The research team visited communities and villages where capsule clinics were established, such as Haichuang Community, Lijia Village in Yunlong Town, and Dongfu Community in Qianhu Street, to conduct qualitative interviews with users (Table 1). Qualitative interviews collected basic user characteristics, home addresses, user acceptance of the capsule clinic, and feelings about using it. The interviews revealed that internet hospitals are different from capsule clinics, which have physical clinics and equipment, and rely on medical resources from offline hospitals. Users can not only experience the same diagnosis and prescription services in capsule clinics as in offline hospitals, but also use intelligent equipment that improves the accessibility of PHC services. As capsule clinics are physical entities, patients view them as more reliable than intangible internet hospitals. Moreover, this physicality makes them more user-friendly and greatly reduces use disparities caused by the digital divide \cite{36-38}.

![Diagram](image-url)
**Figure 3.** Comparison of the 3 medical models. PHC: primary health care.

**Table 1.** Qualitative interviews of the user experience of capsule clinics.

<table>
<thead>
<tr>
<th>Interviewee</th>
<th>Personal information</th>
<th>Address of capsule clinic</th>
<th>Job</th>
<th>Interview content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviewee 1</td>
<td>45</td>
<td>Female</td>
<td>Haichuang Community Health Service Station, Yinzhou District, Ningbo</td>
<td>Employee of a state-owned enterprise</td>
</tr>
<tr>
<td>Interviewee 2</td>
<td>28</td>
<td>Male</td>
<td>Hefeng Community Health Service Station, Hefeng Creative Plaza, Yinzhou District, Ningbo</td>
<td>Employee of a foreign company</td>
</tr>
<tr>
<td>Interviewee 3</td>
<td>50</td>
<td>Male</td>
<td>Shanglijia Village Committee, Yunlong Town, Yinzhou District, Ningbo</td>
<td>Factory worker</td>
</tr>
<tr>
<td>Interviewee 4</td>
<td>68</td>
<td>Female</td>
<td>Zhongxing Community Health Service Station, Dongliu Street, Yinzhou District, Ningbo</td>
<td>Retiree</td>
</tr>
<tr>
<td>Interviewee 5</td>
<td>38</td>
<td>Female</td>
<td>Party Mass Service Center of Junrui Community, Xiaying Street, Yinzhou District, Ningbo</td>
<td>Full-time housewife</td>
</tr>
<tr>
<td>Interviewee 6</td>
<td>35</td>
<td>Male</td>
<td>Guanying Village Committee, Yunlong Town, Yinzhou District, Ningbo</td>
<td>Private entrepreneur</td>
</tr>
</tbody>
</table>
Operation Process of the Capsule Clinic

The patient arrives at the capsule clinic and enters. They scan their identification or insurance card to gain access to the web-based site, where they can review their medical records, refill prescriptions, and schedule a remote appointment with a doctor. Figure 4 illustrates this flowchart of capsule clinic health care services for patients. After receiving the signal that a patient needs a consultation, a doctor from a higher-tier medical institution provides web-based medical services, including face-to-face consultation via a remote video system. Figure 5 illustrates this flowchart of capsule clinic health care services for doctors. As noted, patient prescriptions are generally filled immediately after the web-based consultation. The big data system collects the characteristics of the population in the area where the capsule clinic is located, along with past medication habits. It configures the drug resources in the intelligent pharmacy in accordance with the provided information and the doctor’s advice. The available drug resources are configured to meet the needs of more than 90% of the population. Patient prescriptions are provided by an intelligent pharmacy that focuses on the community residents’ demands and their physical conditions. Furthermore, if residents only want to refill medications at the capsule clinic, they can receive the same prescription with the help of web-based doctors.

Unlike the internet hospital, the capsule clinic not only provides consultation services but also supports patients’ medication needs with the intelligent medicine cabinet under the approval of the offline physical clinic. This measure provides great convenience for patients with long-term drug demands due to chronic diseases. When patients complete a physical hospital visit and have qualified for web-based prescription refills, they can pick up their next refill at the capsule clinic. If patients are identified as needing further examination or hospitalization during the remote consultation, the doctor will schedule an appointment for them at the parent hospital.

The number of capsule clinic users reached 12,219 by October 2022 (Table 2). Since 2020, the number of users has shown a gradual upward trend. Most of these patients are aged between 18 and 60 years; older adults have problems using the capsule clinics owing to the digital divide. Regarding the choice of drug purchased at the capsule clinic, 94.09% of patients buy Rx (Receptor X) medication, whereas only 4.49% choose over-the-counter medication, and 1.42% visit the capsule clinic simply to use the health monitoring program. This shows that most patients use both diagnosis and treatment and medicine dispensing functions in capsule clinics. In terms of the total cost of related drugs, more than 80% of patients spent less than 300 Chinese Yuan (US $43.20) per visit. Perhaps most of the PHC needs of nearby residents can be met by the capsule clinics, where people can avail of basic diagnostic services and buy essential medicines.

Various data from the past 3 years show that the number of capsule clinics in use is increasing year by year. Residents who live nearby are curious about the new medical institutions and may try the health monitoring program when walking past. Users are very interested in the 24-hour service provided by the clinic. In addition, the low number of users may have resulted from inadequate publicity; therefore, many people, especially older adults, are afraid to try this novel medical model. Although older adults still experience obstacles to using the new intelligent medical model, the use of capsule clinics is on the rise owing to the rapid development of internet technology, the expansion of internet access facility coverage, and the facilitation of using facilities. The gap between patients’ “willingness to use” and “ability to use” capsule clinics may influence the digital divide.

Thus far, based on the geographical distribution mentioned above, the capsule clinics are distributed in each residential location, which satisfies the residents’ desire to avail of PHC services close to home and at any time. Compared with the distance that the residents previously had to travel to reach community hospitals, the geographic accessibility of PHC services has been largely addressed by the capsule clinics. Furthermore, the issue of economic accessibility has also been addressed as people with jobs can avoid the cost of absence for medical treatment and the cost of transportation. With the help of big data, the capsule clinic’s smart pharmacy could estimate and allocate drugs in accordance with residents’ medication characteristics, meet residents’ medication needs, and achieve good coverage of PHC services.
Figure 4. Flowchart of capsule clinic health care services for patients.

Figure 5. Flowchart of capsule clinic health care services for doctors.
Table 2. Characteristics of patients of capsule clinics from 2020 to 2022.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Patients, n (%)</th>
<th>2020</th>
<th>2021</th>
<th>2022</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total users</td>
<td></td>
<td>2392</td>
<td>4595</td>
<td>5232</td>
<td>12,219</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤18</td>
<td></td>
<td>430</td>
<td>879</td>
<td>925</td>
<td>2234</td>
</tr>
<tr>
<td>18-40</td>
<td></td>
<td>874</td>
<td>1923</td>
<td>1740</td>
<td>4537</td>
</tr>
<tr>
<td>40-60</td>
<td></td>
<td>739</td>
<td>1168</td>
<td>1650</td>
<td>3557</td>
</tr>
<tr>
<td>≥60</td>
<td></td>
<td>349</td>
<td>625</td>
<td>917</td>
<td>1891</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td>839</td>
<td>1531</td>
<td>1982</td>
<td>4352</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td>1553</td>
<td>3064</td>
<td>3250</td>
<td>7867</td>
</tr>
<tr>
<td>Medicine category</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rx&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td>2212</td>
<td>4297</td>
<td>4989</td>
<td>11,498</td>
</tr>
<tr>
<td>OTC&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td>84</td>
<td>250</td>
<td>214</td>
<td>548</td>
</tr>
<tr>
<td>No medicine&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td>96</td>
<td>48</td>
<td>29</td>
<td>173</td>
</tr>
<tr>
<td>Total drug cost range (Chinese Yuan&lt;sup&gt;d&lt;/sup&gt;)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-100</td>
<td></td>
<td>957</td>
<td>1742</td>
<td>2048</td>
<td>4747</td>
</tr>
<tr>
<td>100-200</td>
<td></td>
<td>872</td>
<td>1492</td>
<td>1664</td>
<td>4028</td>
</tr>
<tr>
<td>200-300</td>
<td></td>
<td>251</td>
<td>544</td>
<td>481</td>
<td>1276</td>
</tr>
<tr>
<td>&gt;300</td>
<td></td>
<td>312</td>
<td>817</td>
<td>1039</td>
<td>2168</td>
</tr>
</tbody>
</table>

<sup>a</sup>Rx: Receptor X.
<sup>b</sup>OTC: over the counter.
<sup>c</sup>No medicine: patients do not obtain medicine from the capsule clinic and only use the health monitoring system services, such as measuring blood pressure, height, weight, and other basic items.
<sup>d</sup>1 Chinese Yuan=US $0.14.

Routine Management and Operation

In the context of the United Nations (UN) pilot program, since the capsule clinic concept will be implemented throughout China, its routine management should be considered. The physical medical institutions (community health care stations) that have obtained the "Medical Institution Practicing License" can apply to establish a capsule clinic. The institution and application report should be submitted together to the local health administration department. Doctors from the parent institution provide remote consultations for the associated capsule clinic. Additionally, traditional hospitals (community health care centers), doctors, and the local health administration department are jointly responsible for the routine management. It is important to encourage video surveillance at capsule clinics and to incorporate procedures that ensure the traceability of drugs dispensed automatically. The physical health care center is primarily responsible for not only the unified management of capsule clinics set up by subordinate branches but also the safety of the medical services provided and the quality of the drugs dispensed at the capsule clinic. Drugs dispensed by the intelligent medicine cabinet at the capsule clinic should adhere to the wholesale drug distribution enterprises’ unified procurement guidelines.

Conclusions and Limitations

With the emergence of capsule clinics, China has a new medical model that can alleviate PHC accessibility problems. Capsule clinics are currently being promoted in every township, village, and community in Ningbo, with the goal of ensuring that every community resident has access to PHC services at their doorstep. The clinics make it convenient for people in remote areas to achieve higher-quality and equitable health resources. In the future, the UN and China will reach consensus on promoting capsule clinics nationwide and become the global template for PHC service delivery. Capsule clinics in remote areas can improve the geographic accessibility for people who lack PHC resources. Capsules can also work well in cities through the use of big data. In many cases and situations, capsule clinics can provide people with the timely PHC services they need, thus reducing financial losses.

Capsule clinics are aimed at people with both common diseases and chronic diseases, and they provide web-based follow-up services, medication disbursement, web-based medical insurance settlement, and more. With the goal of enhanced PHC accessibility, the UN and China intend to place capsule clinics in schools, nursing homes, suburbs, islands, remote mountain
areas, and other modern communities. Establishing capsules in these areas can have important effects on many populations. Figure 6 shows that the capsule clinics can be used in the future as follows:

1. Capsule clinics can be placed in remote mountainous areas and islands to increase residents’ access to PHC services. In many cases, low geographic accessibility is a barrier to residents’ access to effective health resources. The capsule clinic can help solve the problem. People in remote areas can receive basic health care services through telehealth. The advantages of capsule clinics, especially in terms of access to essential medicines, are greatly highlighted for these populations.

2. Capsule clinics can use big data to effectively allocate resources, accurately identify regional population characteristics, allocate appropriate PHC resources, and increase investments in the resources needed to meet the various needs of the people. As such, capsule clinics can be placed in homogeneous, crowded places such as schools and nursing homes. As a special place in the community, the school takes the safety of students very seriously, and an on-site capsule clinic can help ensure it.

3. Communities should take advantage of capsule clinics’ 24-hour services. Capsule clinics could be installed in crowded locations with nighttime activity to prevent accidents and, at the same time, offer people timely treatment or access to medication. For example, factories and hotels are key places where things go wrong; an on-site capsule would allow people to seek timely medical help. Doctors are always available digitally to provide reliable advice for seekers.

Figure 6. Future application scenarios planned.

Of course, the capsule clinic has some limitations. First, data security supervision and drug use safety must be strengthened. Second, the publicity for capsule clinics is not sufficient. People seem to be unaware of this new model or have a wait-and-see attitude. Influenced by traditional health care concepts, numerous patients are more likely to use physical hospitals and are unwilling to use the capsule clinic; consequently, capsule clinic usage is still low. Third, capsule clinics in China are currently limited to urban communities in Ningbo; the single location application and scenario reflect the lack of promotion efforts.

Although the capsule clinic still has some challenges and limitations, it is meaningful in terms of developing convenient and less costly intelligent PHC treatments. People who live in remote areas such as villages and poor mountainous areas far from health care centers face barriers to accessing PHC. Telehealth needs to be enhanced in terms of user experience, and the equipment should be more user-friendly and universal. People are more likely to use smart medical treatment to improve the quality of health services on the basis of the familiar medical treatment mode. Many patients are likely to be driven into poverty by the indirect economic burden of disease. Meanwhile, with higher living standards and enhanced
conceptions of health, more and more people are demanding better health products and higher-quality health services. If we are to reach our goal of improving universal health coverage, we must commit to investing in and scaling up proven solutions. Thus, promoting capsule clinics is highly relevant. The intelligent medical industry is still in its infancy, and there is great room for improvement and sufficient potential demand. In the future, the capsule clinic may help solve fundamental imbalances in the distribution of medical resources and contradictions among the growing health care needs of the population.

Countries need to make more and smarter investments in foundational health systems, with an emphasis on PHC, essential services, and marginalized populations. We must make great efforts to ensure equal access to public health services for all.

Acknowledgments

We would like to thank Yinzhou District Health Bureau, Ningbo, for their assistance and support in the field research. This study was sponsored by The National Natural Science Foundation of China (72274141); the Zhejiang Provincial Natural Science Foundation (LY22G030006); the World Health Organization, which collaborated on a pilot project to build a high-quality and efficient health service system (GJ2-2021-WHOPO-C2); and the General Research Project of Education Department (Y202147813). The sponsors were not involved in the design and conduct of the study; the collection, management, analysis, and interpretation of data; or the preparation, review, and approval of the manuscript.

Authors’ Contributions

DLL, RJZ, CC, MH, and LMS conceived and designed the study. YYH, XYW, XYZ, and QRY participated in field research. CC, RJZ, YYH, and XBJ wrote the original draft. DLL, RJZ, LMS, and XYZ substantively revised the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

References


One Digital Health Intervention for Monitoring Human and Animal Welfare in Smart Cities: Viewpoint and Use Case

Arriel Benis1,2,3*, PhD; Mostafa Haghi4,5,6*, PhD; Thomas M Deserno5,6, PhD; Oscar Tamburis2,3,7, PhD

1Department of Digital Medical Technologies, Holon Institute of Technology, Holon, Israel
2Working Group “One Digital Health”, European Federation for Medical Informatics (EFMI), Le Mont-sur-Lausanne, Switzerland
3Working Group “One Digital Health”, International Medical Informatics Association (IMIA), Chene-Bourg, Geneva, Switzerland
4Ubiquitous Computing Laboratory, Department of Computer Science, HTWG Konstanz – University of Applied Sciences, Konstanz, Germany
5Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Germany
6Working Group “Accident & Emergency Informatics”, International Medical Informatics Association (IMIA), Chene-Bourg, Geneva, Switzerland
7Institute of Biostructures and Bioimaging, National Research Council, Naples, Italy
*these authors contributed equally

Corresponding Author:
Arriel Benis, PhD
Department of Digital Medical Technologies
Holon Institute of Technology
52 Golomb Street
Holon, 5810201
Israel
Phone: 972 03 5026892
Email: arrielb@hit.ac.il

Abstract

Smart cities and digital public health are closely related. Managing digital transformation in urbanization and living spaces is challenging. It is critical to prioritize the emotional and physical health and well-being of humans and their animals in the dynamic and ever-changing environment they share. Human-animal bonds are continuous as they live together or share urban spaces and have a mutual impact on each other’s health as well as the surrounding environment. In addition, sensors embedded in the Internet of Things are everywhere in smart cities. They monitor events and provide appropriate responses. In this regard, accident and emergency informatics (A&EI) offers tools to identify and manage overtime hazards and disruptive events. Such manifold focuses fit with One Digital Health (ODH), which aims to transform health ecosystems with digital technology by proposing a comprehensive framework to manage data and support health-oriented policies. We showed and discussed how, by developing the concept of ODH intervention, the ODH framework can support the comprehensive monitoring and analysis of daily life events of humans and animals in technologically integrated environments such as smart homes and smart cities. We developed an ODH intervention use case in which A&EI mechanisms run in the background. The ODH framework structures the related data collection and analysis to enhance the understanding of human, animal, and environment interactions and associated outcomes. The use case looks at the daily journey of Tracy, a healthy woman aged 27 years, and her dog Mego. Using medical Internet of Things, their activities are continuously monitored and analyzed to prevent or manage any kind of health-related abnormality. We reported and commented on an ODH intervention as an example of a real-life ODH implementation. We gave the reader examples of a “how-to” analysis of Tracy and Mego’s daily life activities as part of a timely implementation of the ODH framework. For each activity, relationships to the ODH dimensions were scored, and relevant technical fields were evaluated in light of the Findable, Accessible, Interoperable, and Reusable principles. This “how-to” can be used as a template for further analyses. An ODH intervention is based on Findable, Accessible, Interoperable, and Reusable data and real-time processing for global health monitoring, emergency management, and research. The data should be collected and analyzed continuously in a spatial-temporal domain to detect changes in behavior, trends, and emergencies. The information periodically gathered should serve human, animal, and environmental health interventions by providing professionals and caregivers with inputs and “how-to’s” to improve health, welfare, and risk prevention at the individual and population levels. Thus, ODH complementarily combined with A&EI is meant to enhance policies and systems and modernize emergency management.

(JMIR Med Inform 2023;11:e43871) doi:10.2196/43871

https://medinform.jmir.org/2023/11/e43871
Introduction

Background

The current data-driven society pushes people to seek and ask for an always-increasing number of highly personalized services [1]. Citizens of smart cities use ubiquitous and mobile technologies and expect to obtain what they need before asking for it. Thus, smart cities are models of urbanization development wherein technologies are actively deployed to enhance the quality of life (QoL) of both humans and animals [2].

Accordingly, personal data such as location, interests (eg, queries on search engines, reads on social media, answers to surveys, and purchases), contacts (eg, calls and social media connections), and activities (eg, neighborhood meetings and personal announcements) are shared continuously [3]. Concerning health-related issues, citizens, as patients, demand proactive suggestions from a modern and evolving public health sector to improve self-care and lower the number of critical events. To this end, their digitized health records encompass the entire process, from the physician’s consultation (in person or remote) to purchasing prescribed drugs (in person or on the web). People use web-based services to check their laboratory tests or reports of certified physical training. Furthermore, they may use applications for sleep monitoring that are provided by a health care management organization [4-8].

Smart cities and (digital) public health share aspects related to a healthy lifestyle. The Internet of Things (IoT); we will use IoT as an encompassing term including the Internet of Medical Things [IoMT] and the Internet of Animal Health Things [IoAHT]) monitors several subtopics of public health: environmental conditions, electromagnetic radiation, health conditions, fitness activities, food quality, emotions, and accidents [9-11]. Therefore, we need signals, data, and information produced by sensors in wearables or mobile apps or existing on social media networks to run health interventions [10] and cope with emergencies [12]. Accordingly, this has been called to expand the capabilities of health records. In smart cities, IoT devices also trace and track animals. Such applications reflect social and cultural norms, the safety and wellness of animals (eg, pet activity and feeding trackers [13]), and the collective composed of humans and animals [14]. In such communities, large amounts of data are produced continuously and exchanged wirelessly [15]. Its purposes range from home automation to trip management and QoL. These sensors range from microelectromechanical systems to advanced medical devices.

It is worth noting that, during the first waves of the COVID-19 pandemic, an increasing rate of shared information was observed. The need arose to timely deliver health care for both humans and animals as specific clusters of customers belonging to the same ecosystem [5,9]. Valid data collection was needed [16,17] that fit the essential elements of smart (healthy) cities [18,19]. For instance, McConnell et al [20] analyzed the influence of owning animals on stress. Other work has assessed the impact of the surrounding ecosystem on humans’ and animals’ QoL [21,22].

A practical setup for the entire data management process comprises aspects of generation, collection, transmission, storage, extraction, analysis, reporting, and decision-making as codified according to the principles of accident and emergency informatics (A&EI), with the dual purpose of preventing harm and supporting decisions [23]. Furthermore, aspects such as education, citizen engagement, and the large vision of human nature are notable elements of the One Digital Health (ODH) framework.

The ODH Framework to Set Up an ODH Intervention

The ODH framework develops around 2 so-called keys [24]. On the one hand, One Health looks at monitoring and assessing environmental hazard interactions and their impacts on health and biodiversity [25]. In contrast, Digital Health stands as the mature deployment of currently available technology to improve individuals’ health and care [26]. The framework also includes 3 perspectives, 5 dimensions, and a technological ring.

As ODH proposes a new holistic, data-driven approach encompassing human, animal, and environmental health and welfare, a solid common ground with this working scenario emerges. Table 1 adjusts the common aspects of smart cities and digital public health using the ODH framework.

Health professionals traditionally define an intervention as the whole process, starting by defining a protocol; implementing it; systematically evaluating its effects; and then using the results to define, for example, a new therapeutic strategy or a new policy. According to the Digital Health perspective, we refer to ODH Intervention as the use of digital and mobile technologies for specific initiatives addressing human, animal, and environmental system needs. In other words, an ODH intervention is a real-world implementation of the theoretical ODH framework, which relies on Findable, Accessible, Interoperable, and Reusable (FAIR) data [27-29]. In this regard, we recognize an appropriate alignment with the Classification of Digital Health Interventions v1.0 proposed by the World Health Organization (WHO), which deals with the principles of implementing and delivering formal or informal care using electronic health services [10,30].

KEYWORDS

One Health; Digital Health; One Digital Health; accident and emergency informatics; eHealth; informatics; medicine; veterinary medicine; environmental monitoring; education; patient engagement; citizen science; data science; pets; human-animal bond; intervention; ambulatory monitoring; health monitoring; Internet of Things; smart environment; mobile phone
Table 1. Relationships between smart cities, digital public health, and One Digital Health (ODH).

<table>
<thead>
<tr>
<th>Topics in common between (digital) public health and smart cities</th>
<th>ODH dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Citizen engagement</td>
</tr>
<tr>
<td>Surveillance</td>
<td></td>
</tr>
<tr>
<td>Accidents and emergencies</td>
<td>✓</td>
</tr>
<tr>
<td>Environmental conditions</td>
<td></td>
</tr>
<tr>
<td>Electromagnetic radiation</td>
<td></td>
</tr>
<tr>
<td>Health condition of older adults</td>
<td>✓</td>
</tr>
<tr>
<td>Emotions</td>
<td>✓</td>
</tr>
<tr>
<td>Epidemics</td>
<td></td>
</tr>
<tr>
<td>Fitness activities</td>
<td>✓</td>
</tr>
<tr>
<td>Food quality</td>
<td></td>
</tr>
<tr>
<td>Healthy lifestyle promotion</td>
<td>✓</td>
</tr>
</tbody>
</table>

Our main objective was to show and discuss how the ODH framework—combined with A&EI—can dynamically support the comprehensive monitoring and analysis of the daily life events of humans and animals in technologically integrated environments such as a smart home and a smart city.

The paper is structured as follows. After the introduction, we present and analyze a timely use case from a smart healthy city context [31]. We then describe the planning of an ODH intervention to deal with the use case features. Therefore, we remodel and update the original use case as an actual outcome of an ODH intervention. In the end, we introduce a part of the generic evaluation process of an ODH intervention and discuss its implementation potential. We conclude by pointing out the strengths and limitations of this use case, such as future developments of the ODH framework jointly with A&EI.

Use Case

Overview
A smart and healthy city well acknowledges several core ODH challenges. Our use case describes the sequence of steps needed to set up an ODH intervention around a user-centered demand and measure several parameters continuously, comprehensively, and reliably for effective multipurpose and integrative welfare monitoring. In this use case, we deal with human and animal health and welfare monitoring, continuously measuring physiological and behavioral parameters during their different activities—alone or together.

Tracy is a healthy single woman aged 27 years who shares her apartment in a metropolitan area with her dog Mego. Tracy is a junior finance analyst (ie, a tertiary position with responsibilities that might cause stress). During working weekdays, Tracy follows some typical activities of daily living. Mego is a healthy midsized male Shiba Inu dog aged 5 years. Mego lives with his owner Tracy. Most of Mego’s time is spent at home (Figure 1).
A Regular Day in Tracy and Mego’s Lives

At 6 AM, Tracy usually wakes up and checks that Mego’s smart pet feeder delivered kibble and water. She then goes for a 30-minute run with Mego. Returning at 6:50 AM, Tracy takes a shower and has breakfast while briefly checking emails, social media, and the news. At 7:30 AM, she goes to the office by car or bicycle depending on the weather. Tracy works from 8 AM to 5:30 PM. She starts by planning and reviewing professional emails and then tackles high-priority tasks before handling analytical tasks. She takes a 1-hour lunch break with colleagues approximately at noon. The afternoon is organized around meetings and administrative tasks.

To return home, if Tracy used the bicycle on her way to work and there are issues on the way back, she may choose to take it on the bus or train. During this time, Mego, similar to other dogs, rests for a large part of the day and occasionally wakes up and barks to react to surrounding noises, plays with his toys, or drinks some water [13].

At around 6 PM, Mego warmly welcomes Tracy as he is excited to go out. If the weather, unexpected events, and mood conditions allow, they spend approximately 1 hour on a walk. During this time, Tracy does some shopping and meets friends in a park [32]. By 7:15 PM, they return home, and Mego receives his evening food ration. Tracy cooks dinner, watches television, reads a book, and prepares her plans for the next day. At 11:30 PM, Tracy goes to sleep.

Figure 1 shows a 24-slice clock in which the main activities of Tracy and Mego are reported depending on the moment of the day. In particular, an inner circle is recognizable that reports Mego’s activities (e.g., sleeping or eating kibble). An external circle features Tracy’s activities (e.g., moving to or from work or reading a book). Activities involving both of them are instead reported on the partition line between the aforementioned circles.

Some Important Spatial-Temporal Facts in Tracy and Mego’s Daily Lives

Activity

Tracy spends most of her time indoors at home, her working place, shopping centers, restaurants, and the gym. Outdoor activities include running, working, hanging out with friends in parks and other open spaces, and using public transportation [33-35].

Tracy uses different means of transportation, such as walking, bicycle, car, and public vehicles. From home to work and back, she takes approximately 1 hour of biking—a considerable physical activity. Tracy’s smartwatch informs her of caloric consumption, pulse, and oxygen saturation. The last physiological measurements are regularly stored in the cloud.
while, some people came along, and finally, the ambulance arrived. Tracy stayed a few days in the hospital and another week at home to recover. During this time, Tracy’s friends had been taking care of Mego, hosting him and trying to preserve his routine to reduce the negative impact of Tracy’s absence (ie, during hospitalization) or unavailability to take care of him on her own (ie, during recovery).

Methods

Planning an ODH Intervention

An ODH intervention can be seen as the implementation of a set of digital functionalities, or digitalities [24], designed and deployed to (1) support specific initiatives that address human, animal, and environmental system needs and challenges; (2) assess and study these systems’ expected and unexpected outcomes and effects and collect related data; and (3) select timely metrics for the outcomes of multicriteria decision analyses.

An ODH intervention is implemented to (1) address One Health–related challenges; (2) achieve One Health–related important and strategic outcomes for clinical follow-up and practice, such as for technology improvements needed; and (3) achieve FAIR uses of digital technologies [30,44].

In our case, this translates to (1) a challenge aiming to enhance aspects of healthy lifestyle promotion and surveillance; (2) an outcome consisting of performing effective monitoring of human and animal welfare within the context of a smart city, where health is a pivotal component; and (3) sensors for monitoring as digital technologies through which the intervention is implemented.

Effective monitoring of humans and animals in a smart environment must cover all relevant locations (indoors and outdoors) and define the critical parameters of health care and QoL. [45]. Table 2 reports the 9 major types of interactions between the intervention recipients considering the extant strict interconnectedness between them [46]. Each cell reports instances of how one actor (row) positively or negatively affects another (column).

Table 2. Examples of interactions between the 3 One Health components.

<table>
<thead>
<tr>
<th>Human</th>
<th>Animal</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face to face [6]</td>
<td>Food, habitat provision, emotional bond, and health and well-being control [48]</td>
<td>Compliance and sustainability for a technological environment at home, on a bicycle, in the steering wheel of a car, in the fridge handle, and more [52]</td>
</tr>
<tr>
<td>Interactive (technology-based; eg, via social media) [8,47]</td>
<td>Population follow-up and birth control [49,50]</td>
<td>Urbanization [53-55]</td>
</tr>
<tr>
<td></td>
<td>Environment destruction [51]</td>
<td>Climate change, pollution, and regulation [51]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Animal</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional and physical bond; health, well-being, and safety feelings (eg, dog) [56]</td>
<td>Soil fertilization [61,62]</td>
</tr>
<tr>
<td>Disease vectors (zoonoses) [57]</td>
<td>Natural hazard (invasive species) [63-66]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Environment</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food chain [58,59]</td>
<td>Soil fertilization [61,62]</td>
</tr>
<tr>
<td>Natural hazards and disasters [67]</td>
<td>Natural hazard (invasive species) [63-66]</td>
</tr>
<tr>
<td>Space for well-being development [68-70]</td>
<td>Wildfire impacts on slope stability triggering in mountain areas [73]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Environment</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability of survival resources (food and shelter) [71]</td>
<td>Natural hazards and disasters [67]</td>
</tr>
<tr>
<td>Chemicals’ influence on animal reproduction [72]</td>
<td>Wildfire impacts on slope stability triggering in mountain areas [73]</td>
</tr>
</tbody>
</table>
Therefore, all the collected physiological, behavioral, and environmental data must (1) comprise indoor and outdoor locations where subjects are active and in contact [74,75], (2) be continuously time stamped, and (3) be shareable in a FAIR way [27].

Adverse Health Events
The ODH framework delivers a variety of observations for continuous health monitoring. It aims at analyzing health data on environmental, behavioral, physiological, and psychological domains, which is in line with the WHO QoL definition [76]. Continuous health monitoring allows continuous data analytics and even subtle trends to become recognizable at the early stages. This then opens many options for preventive medicine, hindering adverse health events (AHEs) [77]. However, AHEs will still occur but maybe not as frequently as without continuous monitoring. In such an event, the measurements taken from the subject are helpful. As of today, smart homes, smart cars, smart wearables, and smart clothes send out emergency calls, but so far, useful information—although available at the site—is not transferred because of lacking protocols and communication standards. This is addressed by A&EI [23]—in an emergency, whether an accident or a health-related adverse event, every second in fact counts as it carries data correlated with the individual involved. In a temporal-spatial continuous monitoring supported by the dynamic point of perception, the subject’s associated data are distributed in each measuring smart environment and device. Upon the occurrence of an event, such isolated data silos have to concatenate to build the complete informative understanding of the latest health status of the subject. The International Standard Accident Number (ISAN) takes over the role, aiming at (1) standardizing an event by associating a unique number (token) composed of time and location of the event and ID number, (2) automatically collecting the corresponding isolated data slices of the individual from the alerting system (ie, smart environments and wearables), (3) automatically generating the alert and transferring it to the responding system (ie, emergency service), and (4) simultaneously transferring the vital and nonvital data to the curing system (ie, hospital) for informing the medical personnel before the subject is delivered to the hospital [78,79].

ODH Intervention Steps
Overview
We deconstruct the ODH intervention into single independent events and steps (globally called “activities”) as items to be organized in tabular form, called the ODH intervention table (see Multimedia Appendix 1). For each activity, a number of fields must be identified, which relate to, for example, ODH domains and dimensions considered, digitalities involved, and their eventual mutual linkages (FAIRness levels).

Activity Identification
Each activity is assigned to a single row. The field “Activity UID” indicates the unique numerical identifier of the activity within the ODH intervention. The “Activity name” describes the single ODH intervention step. If the activity is broken down into subtasks, each one of them will be evaluated individually (ie, as a single row).

ODH Dimension Scores
An ODH dimension score assesses the dimensions of the ODH framework that directly relate to the activity. Each dimension obtains a specific importance value (increasing integer from 1 to 5) that indicates its connection rate to the corresponding activity of the intervention. The dimensions are reported as follows: C (citizen engagement), E (education), H (human and veterinary health care), I (health care Industry 4.0 as health care services and technologies involving, eg, robots, 3D printing, cutting-edge ITs, and artificial intelligence [24,80]), and S (surrounding environment).

Main Digitality Domain
The field refers to 1 of the 3 areas of digital functionalities (humanities, animalities, and environmentalities) encompassed by the technology ring within the ODH steering wheel [24]. This field is divided into 2 subfields.

The first subfield, named “Speciality,” reports the digitality explicitly deployed in the corresponding activity. Each speciality is assigned to a single row. This means that each single activity is operationalized by one or more specialities and is characterized by the marking letter of the domain it refers to: H (human), A (animal), or S (surrounding environment). Accordingly, H(Si) denotes the ith speciality that operationalizes the general digital functionality from the human domain, which delivers the corresponding activity.

The second subfield, named “Technology,” refers instead to one or more technological solutions through which the speciality is deployed. Each technology is assigned a single row as well.

FAIR Data and Data FAIRness
Assessing an ODH intervention’s FAIRness is a basic requirement to correctly process the entire management cycle and stewardship of the data collected and shared during the intervention itself. Thus, the design and deployment steps of an ODH intervention need for its data to be FAIR. In this context, it means that the data must be (1) “Findable” by allowing for their discovery and sharing continuously between different monitoring, analyzing, and alerting systems; (2) “Accessible” by allowing the relevant and approved individuals and connected systems (ie, on behalf of accredited organizations) to deal with data and information when, where, and how needed to manage regular and disruptive events; (3) “Interoperable” by involving health-related communication, data exchanges, and processing standard protocols offered in a secured technological framework; and (4) “Reusable” to allow for a systematic, continuous, and intelligent integration of big multidimensional data for primary (eg, real-time clinical and environment monitoring) and secondary (eg, clinical and epidemiological follow-up) uses [29].

More specifically, a FAIR data assessment is conducted for every technology singled out so that an overall score is extracted for the speciality that the technologies refer to [29]. We mark satisfying, good, and total versus inadequate FAIRness of an individual technology with the symbols + or −, respectively.
**Related Digitality Domain**

This points out 1 of the other 2 areas of the technology ring (Figure 2) involved in the same activity as the main one. Similar to the main digitality domain, we divide related digitalities into the subfields *speciality* and *technology* and assess their level of FAIRness.

**Figure 2.** Example of a graphical representation of digitalities involved in a One Digital Health intervention.

---

**Data Linkage**

In total, 2 digitalities from 2 different domains form a relationship. If 3 digitalities—one from each of the 3 domains—connect with each other, we refer to it as a *triality*, and the field is duplicated. The “Data Linkage” field relates to the relationship or triality between the specialities described in the previous subsections. Equation 1 describes an example of a relationship between 2 specialities:

\[
r_{HA} = H(S_i) \oplus A(S_j) \text{ (1)}
\]

where \( r_{HA} \) is the relationship between specialities, \( H(S_i) \) is the \( i \)-th speciality related to a digital functionality in the human domain, \( A(S_j) \) is the \( j \)-th speciality related to a digital functionality in the animal domain, and \( \oplus \) denotes a special defined operation similar to addition. It implies a possible confrontation between different specific deployments (in different domains) of the same kind of digital functionality.

In the absence of a relationship between one speciality from one domain and another speciality from another domain, the symbolism of equation 1 becomes as follows:

\[
r_{Ho} = H(S_i) \oplus \emptyset \text{ (2)}
\]

where \( \emptyset \) is the null element (there is no speciality connecting with \( S_i \)) and \( r_{Ho} \) is the absence of a relationship (the second member is null). Consequently, we represent a triality as:

\[
t = r_{HAS} = H(S_i) \oplus A(S_j) \oplus S(S_k) \text{ (3)}
\]

where \( t = r_{HAS} \) is the triality (the specialities relate to all 3 domains) and \( S(S_k) \) is the \( k \)-th speciality related to a digital functionality in the environment domain.

**Technology Ring**

The technology ring serves as a catalyst among all the digital functionalities that the ODH framework relies on [24]. Accordingly, an ODH intervention takes place ideally within it through a series of steps described as follows. Each of the 3 domains is divided into as many circular sectors as digitalities are involved in the ODH intervention (Figure 2).

The width of each circular sector is directly proportional to the number of technologies deployed by the single speciality. Then, the bisector of the angle at the center is identified for each sector and represented by a dashed line. A dot is located along the bisector to represent the rate of involvement of the referring digitality within the ODH intervention.
The position of each dot is the output of the overall assessment of the FAIRness of the digitality considered, which involves all the possible technologies related at once; the lesser the output of the assessment, the farther the dot is from the center, and the higher the associated score (decreasing importance from 1 to 5; integer values) of the digitality.

We define the harmonic average of the associated mentioned scores as the weight of a relationship between 2 digitalities:

where $M = M(S)$ is the harmonic average, $N=2$ is the number of digitalities related to each other, and $S_i$ is the score associated with each of the specialities that operationalize the digital functionalities involved in the relationship.

We use a harmonic average as it is based on all observations, gives larger weights to smaller observations, and is thus more robust to strong observations and fluctuations of samples. Moreover, it can cope with variable time factors—in our case, the same speciality deployed via different technologies in different domains and in different time settings.

For a triality, we calculate the $M$ value for each of the 3 single relationships and then their arithmetic average. $M$ is a function of the scores associated with the 2 related dots. The higher the output of the FAIRness assessment of the digitalities, the closer the corresponding dot is to the center of the technology ring, the lower the ranking of the relationship (decreasing importance from 1 to 5), and the more effective that portion of the ODH intervention is. From a graphical point of view, the dots and relationships shall tend toward the center of the ring. For an optimal ODH intervention, its ODHness (ie, an overall evaluation of how well the ODH intervention is delivered) has to tend to 0 [24].

Eventually, the weight associated with the dimensions involved in the ODH intervention is a corrective factor that improves the scores obtained in the previous steps. Accordingly, the greater the number of dimensions involved in the ODH intervention, the more effective the corrective factor is supposed to be.

**Ethical Considerations**

Our research is not an observational study and does not produce or analyze any data taken from humans. Rather, we present a viewpoint and a use case as a potential scenario for an ODH intervention. Accordingly, a review by an ethics committee is not required.

However, planning an ODH intervention must cope with data privacy and security. In our opinion, confidentiality protection is a sensitive issue. In this use case, we do not focus specifically on this aspect. Nevertheless, in a real-world operationalization, it will be critical to have a dedicated data protection framework that allows us to deal when needed with data anonymization and deidentification.

In addition, anyone interested in running (primary use) or using data collected (secondary and tertiary uses) during an ODH intervention in real-world conditions must, according to the relevant local rules, obtain the approval of the relevant research ethics and data protection committees.

**Results**

**Use Case (Updated): A Day in Tracy’s Life in a Smart Healthy City**

To picture this vision, we flashback to our use case with Tracy and Mego. This time, we tell the same story but set in a smart city using smart devices.

Tracy wakes up while her smart home processes the physiological measurements acquired overnight with IoT devices: a radar sensor on top of the bed monitors the respiratory rate, and the bed is equipped with a capacitive electrodiograph in the mattress. Tracy also uses ballistocardiography and seismocardiography mounted on the bed frame for cardiorespiratory measurements and sleep assessment. When Tracy opens a door or presses a switch, her body temperature, electrodermal activity (EDA), and photoplethysmography (PPG) are taken, which deliver heart and respiratory rates upon touching the smart door handle or smart switch, respectively. In addition, passive infrared sensors gather activity data for both Tracy and Mego. In the bathroom, the smart toilet analyzes her urine and measures the pH value and density of excrements. In front of the mirror, Tracy’s body temperature is measured by an infrared thermal sensor, whereas the weight scale integrated into records PPG and EDA. When watching television, Tracy’s electrocardiogram (ECG) chair records her heart and respiratory rates.

Before Tracy leaves her apartment, she checks for Mego’s general health status, provided by his smart dog collar, on her app. Upon Tracy’s exit, her smart home transfers the health monitoring to her wearable sensors: Tracy’s smart clothes are equipped with an embedded ECG. During all her outdoor activities, Tracy’s wearable device performs the measurement, and data on the air quality are recorded [17,45].

Today, Tracy uses her car to drive to work. Opening the smart car door activates the measuring system, including body temperature, EDA, and ECG from the steering wheel as well as heart and respiratory rates from image-based PPG computed from the smart car’s indoor camera. It is worth mentioning as well that Tracy’s bicycle is equipped with a smart handlebar that also delivers health data.

**A Particular Day in Tracy’s Life (Updated)**

Some days ago, the continuous health monitoring system reported to Tracy that she should consult a physician—the sequence and length of atrial fibrillation periods had slightly increased, a well-known harbinger of stroke. However, because of time conflicts, the physician’s appointment was still outstanding when Tracy collapsed during her evening walk in the park with Mego. Instantaneously, Tracy’s smartwatch detected the AHE and asked Tracy to release the alert return. The countdown had not yet ended, Mego’s collar reported Mego’s extraordinary excitement to Tracy’s smartwatch, which immediately generated the ISAN and sent an emergency alert. Using the ISAN, the rescue team requested access to the...
smartwatch so as to check previous and ongoing measurements and, therefore, be well prepared when arriving at Tracy’s location. This made it possible to perform the right intervention at the earliest point in time. Tracy recovered soon without staying in the hospital overnight.

Analyzing the ODH Intervention

Multimedia Appendix 1 showcases examples of “how-to” analysis of several “activities” (2 in regular time and 1 during a health-related disruptive event) as part of a wider ODH intervention. For each one of them, the rates of involvement of the ODH dimensions are scored, and the relevant technical fields are evaluated in light of the FAIR principles prism. Eventually, for each activity, the data linkage formula is also reported.

The first example relates to the monitoring of Tracy’s respiratory rate via a radar sensor on top of her bed. From the A&E viewpoint, this activity allows for the detection of respiratory abnormalities to be reported, if needed, in real time to an emergency medical service or be collected for future investigation.

The second example relates to the possibility for Tracy to check on the activity of the wearable devices she and Mego use. This may allow Tracy to know that, “in case of emergency” (eg, fever and muscular pain detected in Mego [before] and Tracy [after]) may be signs of a zoonotic phenomenon, such as a bacterial infection known as leptospirosis, the system will be able to send relevant data from each of them to their own medical records [81].

In both cases, the industrial- and health care–related dimensions of the ODH framework appear to be highly involved. The surrounding environment is only slightly affected, whereas no notable involvement emerges for the human-related dimensions.

The last example involves all the ODH dimensions. In this case, using a smartwatch to send an alert message reflects some kind of citizen engagement (C) in sharing data with health care providers (M) that need to be trained (E) to use advanced technologies based on industrial standards (I) that can be used in different environments (S).

Discussion

Overview

In this viewpoint and use case, our main objective was to show and discuss how the ODH framework will support 24/7/365 technology-based health and environment monitoring to provide the right answer to the right event (related to a human, animal, or place) in the right place at the right time and with the right means. The central challenge in applying ODH in real-world conditions is the integration of digital development and technologies into 1 health concept. This comes with their use to achieve One Health goals rather than redefining and reconceptualizing One Health in the face of technological advancements [10,11,24,29]. In this regard, planning an ODH intervention is critical for taking forward the integration of the 3 main One Health domains (human, animal, and ecosystem) in light of their digital and computational components. The analyzed use case emerges from the combination of smart cities and digital public health. It suggests an around-the-clock scenario of sensor-based welfare monitoring for a human and a pet in a smart environment context—a large part of Tracy and Mego’s lives involves technological measuring and monitoring systems related to IoMT or IoAHT [8,12]. To enhance the understanding of the impact of an ODH intervention and the subsequent assessment of its ODHness, we proposed a way to analyze human and animal activities in different environments by quantifying their relationships with the ODH framework features [24] and their links to technical fields by considering the assessment of the FAIR guiding principles for the 3 areas identified and included within the ODH technology ring [29,76].

Principal Findings

In the use case supporting our viewpoint, as in real life, the large amount of data generated by IoT sensors and technologies allows for an effective analysis of behaviors, habits, pattern extraction, and medical conditions of different types of subjects in the mid- and long term. In addition to the initial aim of health condition prediction, prevention, and early alert perspective—usually reported in the existing literature for single contexts such as home [82], hospital [83], or even veterinary epidemiology [84] and that ODH already addresses in a unified way—a growing importance has arisen for disruptive events and punctual abnormalities.

For Tracy, this could be related, for instance, to a medical emergency, such as dyspnea or cardiac arrest during sleep or a physical activity, detected by physiological sensors. It can be, as another example, a bicycle accident detected by physiological sensors, accelerometers, and automatic alerts to emergency services. An additional example relates to contact tracing via an app for highly contagious diseases such as COVID-19, which can alert Tracy and support her in making timely decisions according to the public health authorities’ recommendations.

For Mego, a disruptive event can be a veterinary emergency in case of an accident, as seen for Tracy, or a change in the hydration frequency that can suggest a food-based intoxication.

Moreover, a disruptive and dangerous event can involve both Tracy and Mego, such as a fire, with Tracy’s physiological parameters dramatically changing within a couple of minutes overnight and Mego’s similarly agitated behavior.

Such aspects of emergency management demand a continuous follow-up of humans’, animals’, and surrounding environments’ health care in a P5 medicine approach [85-88] that combines (1) prediction, acquiring data and building relevant models supporting emergency and disaster preparedness [8]; (2) prevention, detecting (weak) signals of abnormalities and treating them before strong deviations (eg, preventive confinement to avoid an epidemic); (3) participation, involving the handling of human, animal, or environmental issues (eg, engaging in vaccination campaigns [89-91]); (4) personalization, proposing the use of the adapted solutions to a detected issue (a drug-based treatment for a human or an animal or the development of a repopulation program for a hurt vegetable ecosystem); and (5) precision, delivering the right intervention at the right time (as much as possible in real time) by the right
individual on the right ecosystem components to get to, for example, efficient and dedicated evacuation plans [92,93].

In addition, a specific interpretation is also delivered for what concerns (1) citizen engagement and education aimed at information development and large-scale information gathering (so-called citizen science [94]) for first aid, basic health-related technologies, and sustainability using, for example, mobile apps to increase their penetration rate in the grand public and (2) interoperability from an industrial perspective to facilitate the interconnection and communication between the different IoT systems and alert systems to enhance end-to-end accident and disaster management [78,79].

As shown, singling out the activities that the ODH intervention in a “smart city meets public health” scenario would be made up of and working the technology ring out accordingly (also) to define its ODHness aimed to commence the dialogue with topics from different perspectives dealing with health communication (eg, health promotion based on engagement and education [6,95]) and surveillance (eg, monitoring health conditions and physical activity, detecting and preparing the health care system for disruptive and long-term events, and behavior or environmental changes) [9].

A&EI discipline showed its contributions to the use case regarding the short-term and abrupt event and abnormality detection as well as long-term prediction and prevention. The use-case analysis indicated that A&EI aims to (1) turn smart private spaces into diagnostic spaces unobtrusively (eg, home and car) [96], (2) build continuous measurement and monitoring via dynamic points of perception, and (3) achieve interconnectivity and communication of the means of measurement. The field focuses on the data linkage and interconnectivity of medical and nonmedical sensors and devices in smart environments to construct and support the onboard and distributed data processing and analysis in hierarchical levels of abstracts [12]. The integration of ODH and A&EI in this use case contributes to the development of educational projects and programs, allowing health, environment, engineering, design, and business students and trainees to develop their creative and critical abilities by proposing new concepts and systems [42,95,97]. These education projects could be used in future smart cities wherein One Health and disaster (ie, accidents and emergencies) prevention and preparedness will be daily life pillars [98,99]. This aspect fits with the United Nations 2030 Agenda for Sustainable Development, which points out that young people should be educated in such a way that smart health monitoring is not only for unhealthy humans and older adults [100,101]. For example, analyzing walking steps is a default functionality of mobile devices and is used by young and healthy people to measure their physical activity in real time. Moreover, from an A&EI perspective, preventing complications of a potential medical emergency is something that smartwatches have yet to achieve by detecting and alerting their owners of the first signs of a cardiac event [102].

What emerges is that the very act of planning an ODH intervention placed particular emphasis on aspects of emergency management, which in this context refers directly back to A&EI. This is actually a way to deal with the same dimensions as ODH. Therefore, it is plausible to say that the ODH framework can be enriched with a new layer-like cross-sectional element so that an infinitely recurring loop is created between the 2 models. In summary, ODH and A&EI as a whole contribute to enhancing (1) the overall QoL of the smart city inhabitants (both humans and animals); (2) the public (digital) health policies and processes frame, such as the entire smart city ecosystem development and management, that is, architecture (eg, accessibility for people with disabilities and reduction of energy consumption) and urbanization (eg, communication systems, transportation networks, malls, education, and health care center locations); and (3) the communication between the different health care system actors involved, such as clinicians, engineers, regulators, and administrators and, more generally, the 3 domains that the ODH framework is made of. The existing Medical Informatics and Digital Health Multilingual Ontology could be expanded and adapted to this end [103,104].

Viewpoint Constraints
A correct development of the “smart city meets digital public health” field is currently still hindered by a number of factors, as briefly discussed in the following sections.

Governance
The control and governance of environmental sustainability can be best approached by assessing ecosystem services that are capable of quantifying and valuing all the goods and services that are generated within the ecosystems themselves. Recently, such networks have been increasingly endowed with digital technologies such as (1) environmental management and monitoring information systems; (2) automated and scalable approaches for collecting, digitalizing, and assembling geocoded big data; and (3) information-fusion algorithms and artificial intelligence that use multiple data streams and clinical decision support algorithms that integrate population-based, public health–focused perspectives into outbreak detection–focused management systems [45,105,106].

Data Security and Privacy
However, it is necessary to consider that using electronic and computational systems to collect, store, and analyze data and react and deliver an appropriate response induces at least 2 challenges. Dealing with health care customer data requires dealing with their security and privacy from an ethical viewpoint to ensure patient rights [107] and prevent data theft that can induce, for example, health-related device hijacking with dramatic consequences [108]. Moreover, the health-related data acquisition and monitoring systems used in smart cities demand seamless and secured bidirectional communication, supporting high-speed, large-bandwidth, and low-latency internet infrastructures materialized by the fifth and sixth generation of wireless network technology (respectively known as 5G and 6G) and future versions. Moreover, the use of cloud-based storage and computational resources is essential to allow (near) real-time data collection and analysis.

Sustainability
Accordingly, from a sustainability perspective, as smart cities are electric power–based and digital green, digital health green
infrastructures must be developed to reduce the environmental fingerprint [109,110].

Therefore, from an industrial perspective, reducing power consumption, electronic waste, and biodegradable health care–related materials is a global challenge [111,112].

**Interoperability**

Furthermore, in smart cities, the heterogeneity of IoT devices, and specifically those related to the health care industry, must benefit from improved interoperability standards and procedures and device compatibility. This last point must critically affect the health service providers’ and customers’ acceptance and use of mobile and ubiquitous technologies [113].

**Health Communication and Disparity**

Nevertheless, a limitation of having a generalized use of IoT devices or, more globally, a healthy smart city lies on health communication and disparity. Indeed, using IoT may be challenging for different groups such as older adults as well as individuals with communication disabilities or limited abilities. Accordingly, even though IoT is taking health care to the next stage worldwide, it could paradoxically exclude some people who need it [6,114]. Therefore, a similar threat may affect the veterinary sector, whose development rate is still quite lagging behind that of human medicine.

**Ethical and Cultural Limitations**

In addition, collecting continuous data on anyone and anything may entail ethical and cultural limitations both related to privacy. In a smart city, data collection and processing are performed such that the customers of health care services are not (fully) aware of the sharing of personal data [115]. Moreover, developing a smart city wherein health monitoring is a pillar requires a high technology acceptance level when we consider that technologies are everywhere. Just to mention another critical sector such as transportation safety, the development of IoT-based vehicle accident detection is important, but privacy limitations exist regarding the fact that any driving event can be recorded even if it is not related to an accident [116,117].

In the use case presented, the use of the different activity trackers points out that the technology is accepted by our persona (Tracy), yet it is an open question whether she was fully aware of “giving up” part of her privacy rights [4,118].

**Limitations of the Study**

Some limitations of this work arise from the way the use case for the ODH intervention was determined. First, the “Animal” domain was only represented by a pet. The presence or role of other kinds of animals—such as nonconventional pets (eg, rabbits and reptiles), herds (eg, cows and goats), or wildlife (eg, wolves and boars)—was not addressed. In addition, the “Environment” domain was not properly included in the use case, whereas in a smart and healthy city, the management of environmental resources, made itself smart via IoT and information and communications technologies, is a critical element toward actual e-governance policy planning [119].

Furthermore, a major focus was on the general category of IoT that, although an all-encompassing type of technology, is currently dealing with peculiar declinations for what concerns the general (human) health care sector with IoMT and, in more recent times, animals’ health and well-being status with IoAHT [120]. Another findable limitation is that the proposed use case only describes a single day in Tracy and Mego’s lives. Although Tracy’s working days are characterized by many routine activities, the analysis of a longer period would have likely involved many more things to be discussed.

**Future Perspectives**

Tracy’s daily occupations and activities presented in this fictional use case are applicable to a large part of the population living in similar contexts.

One of the potential and expected implications of ODH interventions lies on supporting digital health literacy, from children to older adults, to engage with personal health, public health, and environmental monitoring systems to increase awareness. Improvement in this aspect is likely to lead to better health outcomes and a more proactive approach to medical practice, thus reducing both the digital divide and health inequalities [6,121].

On the basis of these assumptions, the development of policies regarding health data management and sharing for secondary use [122-126] is expected to be facilitated and perhaps automated. A timely and FAIR deployment of ODH- and A&EI-based data collection modalities would lead to the establishment of more comprehensive guidelines and decision support systems for what concerns the many (interrelated) features of health care and to a better understanding of user expectations.

The combination of health literacy, human expectations, and data collection—or, rather, the logical flow through these points—yields scenarios where information is stored in electronic health records in a timely manner and available in health care management organizations’ repositories and allows for the implementation of personalized medicine processes considering environmental measurements and human [127] or animal [120,128,129] behaviors. In a broader, holistic view, it looks at a proactive medical practice that no longer addresses patients but humans outside the human-centered health care systems whose improved awareness even steps beyond the concept of “empowerment” [130,131]. This means that electronic health record data shall be processed with other data such as lifestyle habit data, for example, patient-generated health data (ie, not reported to health care professionals) and environmental data (eg, pet owning [42] and hobbies). An integrative understanding of the human way of life and human needs and expectations will help efficiently and effectively enhance health communication–targeted campaigns to improve disease prevention, detection, and follow-up at a large scale [6].
Conclusions

Our vision for the future sees smart cities as an emerging paradigm involving a large set of technologies (eg, IoT and, thus, digital health, telehealth, mobile health, and web-based social networks of patients and caregivers) and behavior-changing tools to refine education, engagement, the consumption of food, physical activity, and the use of technology. The main aim is to improve QoL and life expectancy. This has shifted health care focus from treatment after diagnosis to prediction and prevention and from a health care professional–centered (so-called paternalistic [132]) approach to a health care service user–centered follow-up management. Thus, health care is no longer limited to the walls of the health care centers (eg, clinics and hospitals) and the “homespitals” (also known as hospital at home) but is ubiquitous and based on continuous, comprehensive, and reliable measurement of several physiological and behavioral parameters. The integration of the ODH and A&EI viewpoints will allow for the reduction of disparities and loss of time in managing disruptive health-related events and for looking at health as a whole, wherein human and animal well-being in a secure and proactive environment. ODH and A&EI are triggers for developing and implementing precision public health, currently defined as imaginary, by dealing with the entire data management process from end to end [11].

Acknowledgments

The authors are grateful to Mrs Orly Seligmann (Bachelor of Design), visual communication designer at the Holon Institute of Technology, for her assistance.

Data Availability

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

Authors' Contributions

AB contributed to conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, supervision, validation, visualization, and writing (original draft, review, and editing). MH contributed to conceptualization, formal analysis, investigation, methodology, validation, and writing (original draft, review, and editing). TMD contributed to methodology, validation, and writing (review and editing). OT contributed to conceptualization, formal analysis, investigation, methodology, supervision, validation, and writing (original draft, review, and editing).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Excerpts of the One Digital Health (ODH) intervention table for this use case. [PDF File (Adobe PDF File), 252 KB - medinform_v11i1e43871_app1.pdf ]

References


https://medinform.jmir.org/2023/1/e43871 Jmir Med Inform 2023 | vol. 11 | e43871 | p.128 (page number not for citation purposes)


17. Lepenes R, Zakari IS. Citizen science for transformative air quality policy in Germany and Niger. Sustain 2021 Apr 02;13(7):3973 [FREE Full text] [doi: 10.3390/su13073973]


74. Herrera F, Oh SY, Bailleisen JN. Effect of behavioral realism on social interactions inside collaborative virtual environments. Presence (Camb) 2018 Feb 1;27(2):163-182. [FREE Full text] [doi: 10.1162/pehe_a_00324]


97. Ricci FL, Consorti F, Pecoraro F, Luzi D, Tamburis O. A petri-net-based approach for enhancing clinical reasoning in... [MEDLINE: 33525460]


Abbreviations

A&EI: accident and emergency informatics
AHE: adverse health event
ECG: electrocardiogram
EDA: electrodermal activity
FAIR: Findable, Accessible, Interoperable, and Reusable
IoAHT: Internet of Animal Health Things
IoMT: Internet of Medical Things
IoT: Internet of Things
ISAN: International Standard Accident Number
ODH: One Digital Health
PPG: photoplethysmography
QoL: quality of life
WHO: World Health Organization

©Arriel Benis, Mostafa Haghi, Thomas M Deserno, Oscar Tamburis. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.05.2023. 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.
Monitoring the Implementation of Tobacco Cessation Support Tools: Using Novel Electronic Health Record Activity Metrics

Jinying Chen1,2,3,4, PhD; Sarah L Cutrona1,3, MD; Ajay Dharod2,5,6,7, MD; Stephanie C Bunch8, MPH; Kristie L Foley1,5, PhD; Brian Ostasiewski9, BS; Erica R Hale1,2, MS; Aaron Bridges8, BS, MBA; Adam Moses2, MHA; Eric C Donny10, PhD; Erin L Sutfin11, PhD; Thomas K Houston1,2, MD; iDAPT Implementation Science Center for Cancer Control12

1iDAPT Implementation Science Center for Cancer Control, Wake Forest University School of Medicine, Winston-Salem, NC, United States
2Department of Internal Medicine, Wake Forest University School of Medicine, Winston-Salem, NC, United States
3Department of Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, MA, United States
4Department of Preventive Medicine and Epidemiology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States
5Department of Implementation Science, Division of Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, NC, United States
6Wake Forest Center for Healthcare Innovation, Winston-Salem, NC, United States
7Wake Forest Center for Biomedical Informatics, Winston-Salem, NC, United States
8Center for Health Analytics, Media, and Policy, RTI International, Research Triangle Park, NC, United States
9Clinical & Translational Science Institute, Wake Forest University School of Medicine, Winston-Salem, NC, United States
10Department of Physiology and Pharmacology, Wake Forest University School of Medicine, Winston-Salem, NC, United States
11Department of Social Sciences and Health Policy, Wake Forest University School of Medicine, Winston-Salem, NC, United States
12Wake Forest University School of Medicine, Winston-Salem, NC, United States

Corresponding Author:
Jinying Chen, PhD
iDAPT Implementation Science Center for Cancer Control
Wake Forest University School of Medicine
1 Medical Center Blvd
Winston-Salem, NC, 27101
United States
Phone: 1 617 358 5838
Email: jinchen@wakehealth.edu

Abstract

Background: Clinical decision support (CDS) tools in electronic health records (EHRs) are often used as core strategies to support quality improvement programs in the clinical setting. Monitoring the impact (intended and unintended) of these tools is crucial for program evaluation and adaptation. Existing approaches for monitoring typically rely on health care providers’ self-reports or direct observation of clinical workflows, which require substantial data collection efforts and are prone to reporting bias.

Objective: This study aims to develop a novel monitoring method leveraging EHR activity data and demonstrate its use in monitoring the CDS tools implemented by a tobacco cessation program sponsored by the National Cancer Institute’s Cancer Center Cessation Initiative (C3I).

Methods: We developed EHR-based metrics to monitor the implementation of two CDS tools: (1) a screening alert reminding clinic staff to complete the smoking assessment and (2) a support alert prompting health care providers to discuss support and treatment options, including referral to a cessation clinic. Using EHR activity data, we measured the completion (encounter-level alert completion rate) and burden (the number of times an alert was fired before completion and time spent handling the alert) of the CDS tools. We report metrics tracked for 12 months post implementation, comparing 7 cancer clinics (2 clinics implemented the screening alert and 5 implemented both alerts) within a C3I center, and identify areas to improve alert design and adoption.

Results: The screening alert fired in 5121 encounters during the 12 months post implementation. The encounter-level alert completion rate (clinic staff acknowledged completion of screening in EHR: 0.55; clinic staff completed EHR documentation of screening results: 0.32) remained stable over time but varied considerably across clinics. The support alert fired in 1074 encounters during the 12 months. Providers acted upon (ie, not postponed) the support alert in 87.3% (n=938) of encounters, identified a
patient ready to quit in 12% (n=129) of encounters, and ordered a referral to the cessation clinic in 2% (n=22) of encounters. With respect to alert burden, on average, both alerts fired over 2 times (screening alert: 2.7; support alert: 2.1) before completion; time spent postponing the screening alert was similar to completing (52 vs 53 seconds) the alert, and time spent postponing the support alert was more than completing (67 vs 50 seconds) the alert per encounter. These findings inform four areas where the alert design and use can be improved: (1) improving alert adoption and completion through local adaptation, (2) improving support alert efficacy by additional strategies including training in provider-patient communication, (3) improving the accuracy of tracking for alert completion, and (4) balancing alert efficacy with the burden.

Conclusions: EHR activity metrics were able to monitor the success and burden of tobacco cessation alerts, allowing for a more nuanced understanding of potential trade-offs associated with alert implementation. These metrics can be used to guide implementation adaptation and are scalable across diverse settings.

(JMIR Med Inform 2023;11:e43097) doi:10.2196/43097

KEYWORDS
medical informatics; electronic health records; EHR metrics; alerts; alert burden; tobacco cessation; monitoring; clinical decision support; implementation science; smoking cessation; decision tool

Introduction

Background
Provider-facing computerized clinical decision support (CDS) tools in electronic health records (EHRs) are common digital health interventions supporting health care quality improvement programs [1-6]. Monitoring (ie, continual evaluation) of the impact of these tools is important for program evaluation and may ultimately contribute to implementation success [7,8]. Approaches for evaluating CDS tools largely rely on surveys, qualitative interviews, and data collected through direct observation or audio/video recording [9-12]. These approaches require substantial human effort (from implementation staff and clinical teams) for data collection. Automated methods leveraging EHR activity data offer a promising solution to reduce the data collection burden, but research on these methods is still in the earliest stage.

This study aimed to develop automatic metrics to monitor the implementation of EHR-embedded CDS tools and demonstrate their use within the context of a smoking cessation program sponsored by a National Cancer Institute (NCI)–designated cancer center.

Tobacco Control Programs in NCI Cancer Centers
Tobacco use increases the risk of cancer and leads to poor prognosis after cancer diagnosis [13-16]. Clinical practice guidelines recommend routine screening for tobacco use and referral to evidence-based cessation interventions in patients with cancer [17,18], but this practice is underused [19]. To address this practice gap, the NCI’s Beau Biden Cancer Moonshot program launched the Cancer Center Cessation Initiative (C3I) in 2017 to provide funding to NCI-designated cancer centers to implement or enhance their tobacco treatment services [20].

Electronic alerts (e-alerts) are common CDS tools in EHRs, promoting adherence to practice guidelines [2-5,21,22], including tobacco screening and treatment at the point of care [23-25]. This strategy has been adopted by some C3I-funded cancer centers [26,27]. However, effective implementation of alerts into the clinical workflow is nontrivial [28-30]. Monitoring of provider responses to newly implemented alerts can identify barriers to adoption and the burden imposed by the alerts.

Study Objectives
We developed and applied EHR activity metrics to answer three questions. (1) Did the alert completion rate change over time or vary across clinics? (2) What was the burden introduced by the alerts? (3) What factors were associated with variation in alert completion? Our research questions were motivated by three factors. First, sustainability (eg, sustained use and completion of the alerts) is a key construct of implementation outcomes [8] and should be monitored over time. Second, monitoring variations in alert completion across clinics can support the adaptation of alert implementation to the local context. Third, alerts could add a “burden” on providers [30-33], which should be evaluated.

Methods

Study Design
We developed and applied new EHR activity metrics to monitor the tobacco cessation tools (two Best Practice Advisory [BPA] alerts) implemented in cancer clinics for 12 months (Figure 1).
Figure 1. Study overview. The support alert would fire only if the screening result was positive (ie, patient being a current smoker) and answers to both Smoking Screener questions (Q1: “When did you last smoke (even 1 or 2 puffs)?”; Q2: “Quitting smoking could help improve your health. Are you interested in quitting?”) were documented (see Multimedia Appendix 1, step B1). EHR: electronic health record.

Ethics Approval
The study was approved by the Wake Forest School of Medicine Institutional Review Board (IRB00066841). Deidentified EHR data were used, with informed consent for data access waived by the institutional review board.

Digital Health Intervention
The CDS tools were two conditional sequential alerts integrated into the Epic EHR, a commercial cloud-based EHR system (Figure 1; detailed in Multimedia Appendix 1): (1) a screening alert to remind clinic staff to complete tobacco screening, triggered if “current smoker” or “unknown smoking status” was previously documented in the EHR, and (2) a support alert to prompt the clinical provider to discuss support and referral to a tobacco cessation clinic, triggered if the screening result was positive and answers to both Smoking Screener questions were answered (see note for Figure 1). Each alert had two modalities: (1) interruptive (triggered when the patient chart was opened; if postponed, presenting again after 10 minutes or when the patient chart was reopened) and (2) noninterruptive (in the general BPA section of the EHR).

Implementation Context
The Tobacco Control Center of Excellence (TCCOE) at the Wake Forest Baptist Comprehensive Cancer Center implemented the alerts in the Epic EHR system used by 7 cancer clinics (medical oncology: n=3; radiation oncology: n=3; cancer survivorship: n=1) in the Atrium Health Wake Forest Baptist Comprehensive Cancer Center in 2019 and 2020. The alerts were integrated into the Epic EHR as BPAs, a form of CDS in the EHR that reminds providers to attend to important tasks [4]. The implementation team from the TCCOE worked with the hospital information technology team on implementing the alerts. The alerts were customized by using rule-based logic (eg, rules on who will receive the alerts and when to fire the alerts; detailed in Multimedia Appendix 1). All 7 clinics implemented the screening alert; 5 implemented the support alert. Training was provided to clinic staff and providers (1-month weekly before or in the first month of implementation and monthly check-in after alert implementation). Some clinics used extensive support from patient navigators and tobacco treatment specialists to complete screening documentation and referral to the cessation clinic.

Evaluation
Metrics Development and Automation
Metrics development took three steps: (1) identifying relevant EHR variables, (2) developing SQL queries to extract variables from the EHR database, and (3) developing computer code to calculate the metrics. We used EHR data associated with 2 clinics to develop and test the metrics. A team of experts in health informatics and implementation science, EHR specialists, and physicians participated in the metrics development.

Using the computer code we developed, EHR data extraction takes about 10 minutes, and the calculation of each metric takes tens of seconds. This speed is adequate for monitoring CDS tools used by implementation programs. Full automation of these metrics is possible after their integration into the EHR.
EHR Variables Used to Derive the Metrics

We extracted alert activity data from event log files of the Epic EHR system. The variables used to develop the metrics included alert id, alert instance id, alert name (eg, a tobacco screening alert), timestamps corresponding to alert firing and provider responding (called alert firing time point and alert response time point for convenience), alert triggering condition (eg, triggered by opening the patient chart), subsequent actions taken (eg, acknowledge/override warning), alert override reason, and alert-associated signed orders.

Each alert id is associated with a unique encounter id and a patient id. An alert id corresponds to multiple alert instance ids if the alert is fired again after being postponed. Alert triggering condition was used to distinguish interruptive alerts from noninterruptive ones.

We used subsequent actions taken, alert override reason, and alert-associated signed orders to identify providers’ actions on the alerts. When the clinic staff completed or postponed the screening alert, subsequent actions taken recorded a value “acknowledge/override warning,” and alert override reason recorded whether the staff acknowledged screening completion (ie, hit the button “Documented in Flowsheet,” step A in Figure A1-1, Multimedia Appendix 1), postponed the alert (hit “Defer”), or determined that the patient was inappropriate for screening (hit “Not appropriate”). For the support alert, subsequent actions taken recorded a value “acknowledge/override warning” when the provider hit the buttons under “acknowledge reason,” and alert override reason recorded the provider’s actions (eg, discussed or not discussed with patients) and patient’s readiness to quit (Figure A1-2, Multimedia Appendix 1). Alert-associated signed orders recorded whether the provider placed an order for a referral to the cessation clinic.

In addition, we used two encounter-level variables, flowsheet name and flowsheet value, to determine whether the clinic staff documented screening results (ie, answers to Q1-Q3 in step B1 in Figure A1-1, Multimedia Appendix 1) in the EHR.

Metrics

We defined three metrics to measure alert completion and burden (Multimedia Appendix 2).

The alert completion rate was defined as the number of encounters where a provider completed alert-prompted actions divided by the number of encounters where the alert fired. We defined screening alert completion by either staff’s acknowledging completion of screening or completion of EHR documentation of screening results. We defined support alert completion at two levels: (1) discussing with patients and assessing patient readiness to quit (“discussion”) or (2) referring patients to the on-site tobacco cessation clinic (“referral”).

We measured the burden of interruptive alerts by two metrics: alert firing rate and alert handling time. We focused on interruptive alerts because they were more likely to add a “burden” on providers [30-33]. We defined alert firing rate as the number of times the alert fired during a specific period divided by the number of times the alert was completed during that period. We calculated the average time providers spent completing an alert per encounter, using encounters in which the alert was completed; similarly, we calculated the average time spent postponing alerts per encounter using encounters in which the alert was postponed at least once.

Data Collection

For each clinic, we collected EHR data about clinic characteristics, patient characteristics, and EHR activities related to tobacco cessation alerts. We collected EHR alert activity data as described previously. Each instance of an alert was linked to a specific patient encounter and the patient’s demographic information (sex and race), using encounter and patient IDs.

Data Analysis

We summarized clinic and patient characteristics for each clinic. We then used EHR activity metrics to address the research questions related to alert completion and burden. Statistical analyses were conducted using STATA/MP 15.1 (StataCorp LLC) [34].

We measured the overall and per-clinic alert completion rates for the screening alert during every 3-month period across 12 months post alert implementation. The 12-month postimplementation period was specified for each clinic. We measured the support alert completion rate at two levels: “discussion” and “referral.”

We measured the alert firing rate and handling time of interruptive screening alerts and support alerts to assess the burden of interruptive alerts.

Factors Associated With Alert Completion

As a secondary analysis, we examined the distribution of alert completion over patients’ demographics (sex and race) and encounter types.

Three physicians reviewed all encounter types and selected “relevant encounter” types as those in which screening for smoking status was an appropriate part of routine care (Multimedia Appendix 3).

Results

Clinic Characteristics

The clinics varied in the number of encounters (from n=1464 to n=110,553) and patients (from n=328 to n=9410) during 12 months post alert implementation (Table 1). The typical structure of these clinics was for nurses to support multiple providers across multiple days.

---

https://medinform.jmir.org/2023/1/e43097

Chen et al

JMIR Med Inform 2023 | vol. 11 | e43097 | p.138
### Table 1. Clinic and patient characteristics during the 12 months after implementing tobacco cessation alerts.

<table>
<thead>
<tr>
<th>Clinic characteristics</th>
<th>Medical oncology&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Radiation oncology&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Cancer survivorship (S&lt;sup&gt;c&lt;/sup&gt;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service area</td>
<td>Urban</td>
<td>Rural</td>
<td>Urban</td>
</tr>
<tr>
<td>Staffing, n&lt;sup&gt;e&lt;/sup&gt;</td>
<td>5-10</td>
<td>10-20</td>
<td>M1-M3: medical oncology clinics 1-3.</td>
</tr>
<tr>
<td>Encounters, n</td>
<td>30,727</td>
<td>9102</td>
<td>R1-R3: radiation oncology clinics 1-3.</td>
</tr>
<tr>
<td>Patients, n</td>
<td>4688</td>
<td>1193</td>
<td>S: cancer survivorship clinic.</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>64 (14)</td>
<td>65 (13)</td>
<td>M1 and R1 implemented only the screening alert.</td>
</tr>
<tr>
<td>Sex, n (%)&lt;sup&gt;f&lt;/sup&gt;</td>
<td></td>
<td></td>
<td>The approximate number of clinic team members (physicians, advanced practice practitioners, nurses, and other clinical staff) in a clinic. The number is not precise due to staff turnover and the hiring of temporary staff.</td>
</tr>
<tr>
<td>Female</td>
<td>3122 (66.6)</td>
<td>4956 (52.7)</td>
<td>Some clinics have a small percentage of patients missing information on sex (0.03% missing for R3; complete for other clinics), race (complete for M2; less than 0.3% missing for other clinics), and ethnicity (1% missing for M1, 0.6% missing for R1, 0.2% missing for M3 and R3; complete for other clinics).</td>
</tr>
<tr>
<td>Male</td>
<td>1566 (33.4)</td>
<td>4454 (47.3)</td>
<td>Other: American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, Latin American or Hispanic, and other.</td>
</tr>
<tr>
<td>Race, n (%)&lt;sup&gt;f&lt;/sup&gt;</td>
<td></td>
<td></td>
<td>The percent of patients who were active smokers during 12 months post alert implementation. The denominator is the number of patients who had their smoking status documented in the electronic health record.</td>
</tr>
<tr>
<td>African American</td>
<td>1070 (22.8)</td>
<td>1731 (18.4)</td>
<td>M1 and R1 implemented only the screening alert.</td>
</tr>
<tr>
<td>White</td>
<td>3350 (71.5)</td>
<td>7227 (76.8)</td>
<td>The approximate number of clinic team members (physicians, advanced practice practitioners, nurses, and other clinical staff) in a clinic. The number is not precise due to staff turnover and the hiring of temporary staff.</td>
</tr>
<tr>
<td>Other&lt;sup&gt;g&lt;/sup&gt;</td>
<td>259 (5.5)</td>
<td>436 (4.6)</td>
<td>Some clinics have a small percentage of patients missing information on sex (0.03% missing for R3; complete for other clinics), race (complete for M2; less than 0.3% missing for other clinics), and ethnicity (1% missing for M1, 0.6% missing for R1, 0.2% missing for M3 and R3; complete for other clinics).</td>
</tr>
<tr>
<td>Hispanic or Latino, n (%)&lt;sup&gt;f&lt;/sup&gt;</td>
<td>114 (2.4)</td>
<td>349 (3.7)</td>
<td>Other: American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, Latin American or Hispanic, and other.</td>
</tr>
<tr>
<td>Yes</td>
<td>4545 (96.9)</td>
<td>9042 (96.1)</td>
<td>The percent of patients who were active smokers during 12 months post alert implementation. The denominator is the number of patients who had their smoking status documented in the electronic health record.</td>
</tr>
<tr>
<td>No</td>
<td>1170 (98.1)</td>
<td>1034 (97.6)</td>
<td>M1 and R1 implemented only the screening alert.</td>
</tr>
<tr>
<td>Insurance, n (%)</td>
<td></td>
<td></td>
<td>The approximate number of clinic team members (physicians, advanced practice practitioners, nurses, and other clinical staff) in a clinic. The number is not precise due to staff turnover and the hiring of temporary staff.</td>
</tr>
<tr>
<td>Medicare</td>
<td>2721 (58.0)</td>
<td>4961 (52.7)</td>
<td>Some clinics have a small percentage of patients missing information on sex (0.03% missing for R3; complete for other clinics), race (complete for M2; less than 0.3% missing for other clinics), and ethnicity (1% missing for M1, 0.6% missing for R1, 0.2% missing for M3 and R3; complete for other clinics).</td>
</tr>
<tr>
<td>Medicaid</td>
<td>243 (5.2)</td>
<td>598 (6.4)</td>
<td>Other: American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, Latin American or Hispanic, and other.</td>
</tr>
<tr>
<td>Other insurance</td>
<td>1659 (35.4)</td>
<td>3537 (37.6)</td>
<td>The percent of patients who were active smokers during 12 months post alert implementation. The denominator is the number of patients who had their smoking status documented in the electronic health record.</td>
</tr>
<tr>
<td>No insurance</td>
<td>65 (1.4)</td>
<td>19 (1.8)</td>
<td>M1 and R1 implemented only the screening alert.</td>
</tr>
<tr>
<td>Smoking rate, n/N (%)&lt;sup&gt;h&lt;/sup&gt;</td>
<td>590/4606</td>
<td>10069230</td>
<td>The approximate number of clinic team members (physicians, advanced practice practitioners, nurses, and other clinical staff) in a clinic. The number is not precise due to staff turnover and the hiring of temporary staff.</td>
</tr>
</tbody>
</table>

---

**Patient Characteristics**

The patients seen by the cancer survivorship clinic were 5-8 years older than patients seen by other clinics (mean age for each clinic 59-67; Table 1). Most patients were non-Hispanic White and were beneficiaries of Medicare. The smoking rate ranged between 8.9% (n=45 among 506 patients who had smoking status documented in the EHR; cancer survivorship clinic) and 17.8% (58/325; radiation oncology clinic 2).

**Alert Completion Rate**

The screening alert fired in 5121 (2.8% of 180,647) encounters 12 months post implementation. The alert completion rate was 0.55 (2817/5121) based on the staff’s acknowledgment of screening completion in EHRs and 0.32 (1647/5121) based on the completion of EHR documentation of screening results. Both alert completion rates remained stable over time (Figure 2A) but varied considerably across clinics (Figure 2B-D).
Among the 2817 encounters where the staff acknowledged completion of screening, 84.7% completed interruptive alerts and 15.4% completed noninterruptive ones.

The support alert was implemented for 5 clinics (medical oncology clinic 2 and 3, radiation oncology clinic 2 and 3, and cancer survivorship clinic) and fired in 1074 encounters. Providers responded without postponing (n=938, 87.3%), discussed tobacco use treatment options (n=640, 59.6%), identified patients who were ready to quit (n=129, 12%), and placed referrals to the cessation clinic (n=22, 2%).

**Figure 2.** Completion rate of tobacco screening alert for (A) all clinics and (B-D) individual clinics. Clinics in (C) and (D) were categorized into three levels based on the number of encounters in which a screening alert was fired during 12 months post alert implementation. Level 1: >1000; level 2: >100 and ≤1000; level 3: ≤100. Line thickness was used to represent these three levels. EHR: electronic health record. M1-M3: medical oncology clinics 1-3. R1-R3: radiation oncology clinics 1-3. S: cancer survivorship clinic.

### The Burden of Interruptive Alerts

On average, the number of times a screening alert was fired before completion was 2.7 (range 1.0-12.7 for individual clinics; Table 2); the average number of times a support alert was fired before completion was 2.1 (range 1.8-3.3 for individual clinics; Table 2).

**Table 2.** Alert firing rate of the screening alert and the support alert by clinics.

<table>
<thead>
<tr>
<th></th>
<th>Medical oncology</th>
<th>Radiation oncology</th>
<th>Cancer survivorship (S)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M1&lt;sup&gt;d&lt;/sup&gt;</td>
<td>M2&lt;sup&gt;d&lt;/sup&gt;</td>
<td>M3&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td>Screening alert</td>
<td>4.9&lt;sup&gt;e&lt;/sup&gt;</td>
<td>1.3</td>
<td>2.2</td>
</tr>
<tr>
<td>Support alert</td>
<td>N/A&lt;sup&gt;f&lt;/sup&gt;</td>
<td>1.8</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R1&lt;sup&gt;d&lt;/sup&gt;</td>
<td>R2&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td>Support alert</td>
<td></td>
<td>1.8</td>
<td>3.3</td>
</tr>
</tbody>
</table>

<sup>a</sup>M1-M3: medical oncology clinics 1-3.

<sup>b</sup>R1-R3: radiation oncology clinics 1-3.

<sup>c</sup>S: cancer survivorship clinic.

<sup>d</sup>M1 and R1 implemented only the screening alert.

<sup>e</sup>We defined the alert firing rate as the number of times the alert fired during 12 months post alert implementation divided by the number of times the alert was completed during the same period. We did not calculate the alert firing rate at the encounter level because it was undefined (ie, division by 0) for encounters that did not complete the alert.

<sup>f</sup>N/A: not applicable.
On average, time spent completing the screening alert per encounter was 53 seconds (50 seconds for support alert); time spent postponing screening alerts per encounter was 52 seconds (67 seconds for support alerts).

Factors Associated With Alert Completion

Completion rates of the screening alert and the support alert were balanced across patient subgroups (sex, race, and their interaction).

Among 5121 encounters for which the screening alert was fired, 4425 (86.4%) were “relevant” and 696 (13.6%) were “less relevant” to routine tobacco screening. The alert completion rate for “relevant” encounters was higher than that for “less relevant” ones (2793/4425, 63.1% vs 24/696, 3.5%; \( P < .001 \)).

Table 3. Key findings from the application of the electronic health record (EHR) activity metrics and implications for clinical decision support (CDS) tool design and use.

<table>
<thead>
<tr>
<th>Key findings from the application of EHR activity metrics</th>
<th>Implications for CDS tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variation in alert completion</td>
<td>Potential for improving alert adoption/completion through local adaptation</td>
</tr>
<tr>
<td>• The screening alert completion rates varied substantially across the clinics.</td>
<td>Strategies to support use:</td>
</tr>
<tr>
<td>• The screening alert completion rate was higher for encounters perceived as relevant to routine tobacco screening by physicians.</td>
<td></td>
</tr>
</tbody>
</table>

Limited alert efficacy

• Providers responded to most support alerts, but few patients were ready to quit, and referral to the tobacco cessation clinic was rare.

Inconsistencies between the acknowledgment of alert completion and documented screening

• EHR documentation of screening results was rare for some clinics, even though their clinic staff acknowledged completion of screening for most encounters.

Interruptive alerts received more responses but also added burden to providers

• Providers were more responsive to interruptive alerts than noninterruptive ones.
• Postponing the interruptive alert did not save providers time compared with completing the alert.

Improving Alert Adoption and Completion Through Local Adaptation

Clinics varied substantially in completing the alert, calling for clinic-specific strategies to improve alert adoption. We also identified a modifiable factor (ie, the alert encounter relevance) that affects alert completion. Our physician coauthors considered

certain encounter types (eg, initial consultation and office visit) to be relevant for routine tobacco use screening, while others (eg, lab visit and radiation oncology treatment visit) were deemed less relevant. While existing guidelines recommend repeating the smoking assessment at every encounter [17,18], we found that the completion rate of the screening alert was much lower for “less relevant” encounters, which may appear

Discussion

Principal Results

We developed and applied EHR activity metrics to monitor two tobacco cessation CDS alerts implemented in 7 cancer clinics. Our metrics were able to capture variation in alert completion across clinics, monitor alert efficacy, identify discrepancies between staff-acknowledged screening completion and screening documentation, and provide insights into the balance between alert efficacy and imposed burden. These findings inform four areas where CDS tool design or use can be improved (Table 3), which we discuss below.
to be guideline noncompliance. This finding could be informative for committees that develop tobacco screening and treatment guidelines. Implementation teams that want to enforce the “screening at every encounter” rule may need additional strategies. These could include using provider orientation and local champions to influence the culture surrounding tobacco screening [35].

Improving Support Alert Efficacy

Although providers responded to support alerts frequently, referral to the tobacco cessation clinic was rare. One reason was that few patients were ready to quit at the point of care. Future programs may incorporate additional strategies, such as patient education, provider training in patient-provider communication, and addressing patient-level barriers (eg, barriers associated with health beliefs and socioeconomic factors) [36,37]. Note that the 2% referral rate may underestimate the effect of the support alert because it was calculated based on referrals directly linked to the alert. If tobacco treatment specialists contacted the patients interested in quitting after the patient visits, these follow-up activities would be documented elsewhere without a link to the alert, or if a patient chose other treatment methods (eg, quitline or medications), the alert-driven referral would not happen.

Improving Accuracy of Tracking for Alert Completion

The completion rates of the EHR documentation of screening results were lower for some clinics, even though their clinic staff acknowledged screening completion for most encounters. Through discussion with the team coordinating the tobacco cessation program, we identified one major reason for this gap. In clinics using support from patient navigators to complete screening documentation, the clinic staff were likely to bypass the screening but still acknowledged completion. Therefore, measuring EHR documentation is important for the accurate tracking of alert success. We used encounter-level data for this measurement. Alert-level tracking may be necessary for the future development of targeted strategies (eg, provider-specific training) to improve alert adoption. The alert design can be improved to allow this, for example, by disabling the button for acknowledging the completion of screening until the EHR documentation is completed.

Balancing Alert Efficacy With Burden

Although commonly used, effective integration of e-alerts into the clinical workflow has proven difficult [29-33,38,39]. Medication alerts were frequently overridden by health care providers [29,30,33,40], and providers experienced alert-related burden and fatigue [9,29,31,41]. Our study found that postponing the interruptive alert did not save providers time compared with completing the alert. This was partly due to the refriring of postponed alerts. An overabundance of interruptive alerts in EHRs may lead to frequent “postpone” or “override” actions and user dissatisfaction [31-33]. However, our findings do not support disabling the interruptive alerts, as we found that providers were much more responsive to interruptive alerts than noninterruptive ones. One way to alleviate the alert burden is increasing the time interval between postponing and refiring or setting the maximum number of times (eg, 2 or 3) to fire a tobacco cessation alert during each encounter.

Contribution to Implementation Science Methods

New methods are needed for monitoring implementation, including automated approaches that reduce the data collection burden [7,42]. We contributed to this literature by developing automatic EHR activity metrics for monitoring the implementation of CDS tools. Our approach has three merits. First, automatic metrics are suitable for rapid periodic evaluation of implementation programs. These metrics can identify deviations and variations of CDS use at clinic and provider levels, which may inform the selection of key informants for interviews to identify causes of deviation and variation, and the development of strategies to improve CDS design and use. Second, EHR activity data work “behind the scenes” to capture EHR use behavior without interruptions [43-45]. Metrics built on this data can reduce reporting bias and may minimize Hawthorne effects (ie, participants’ engagement with an intervention changes when they are aware of attention from observers) [46]. Third, EHRs have been adopted by most US hospitals [47], and EHR-embedded CDS tools are frequently used to support health care quality improvement [1-6]. The ubiquity of EHRs contributes to the generalizability of our approach.

Our work relates to studies using EHR audit logs (one type of EHR activity data) but is different in methodology. The metrics described in these studies measure EHR use and associated burden (eg, total time on EHR, time spent using the EHR after hours, time spent on chart review per patient per day) [48-52] nonspecific to CDS tools. Using EHR audit logs to measure providers’ response to a specific EHR tool is challenging, typically involving manual mapping of low-level actions recorded in the log files to EHR use activities [39,50,53]. We used alert activity data generated by Epic’s built-in functions to eliminate manual mapping.

Prior studies on alert burden focused on medication alerts and used alert override rate and alert volume as markers for burden in the context of de-implementation [30-33,40]. To our knowledge, this study is the first to systematically measure the burden of preventive care alerts. Our findings did not support a simple de-implementation approach but call for better local adaptation to balance alert efficacy and burden.

Limitations

This study has several limitations. First, the EHR data we analyzed only contained alert-linked referrals to the tobacco cessation clinic. Our analysis may underestimate the actual effect of the support alert. Second, EHR activity data only capture provider interaction with the EHR and lack information about other clinical activities (eg, discussion with patients, pager ringing) during an encounter. In-depth investigations on clinical workflows and their impact on alert response are needed to better understand the variation of alert completion across clinics.

Conclusions

This study developed EHR activity metrics and demonstrated their use in monitoring the impact of CDS tools implemented
by a C3I-funded implementation program that promotes tobacco cessation in patients with cancer. These metrics can be used to guide implementation adaptation and are scalable and adaptable to other settings that use e-alerts to promote adherence to health practice guidelines.

Acknowledgments

The tobacco cessation alerts were implemented by the Tobacco Control Center of Excellence for clinics at the Wake Forest Baptist Comprehensive Cancer Center. The Wake Forest Baptist Comprehensive Cancer Center was renamed Atrium Health Wake Forest Baptist Comprehensive Cancer Center in 2021.

This work was supported by the National Cancer Institute of the US National Institutes of Health (grants P50CA244693 and P30 CA012197, and CRDF award 66590 through the Cancer Center Cessation Initiative Coordinating Center contract) and the National Heart, Lung, and Blood Institute of the US National Institutes of Health (grant K12HL138049 to JC through the Massachusetts Consortium for Cardiopulmonary Implementation Science Scholars K12 Training Program). The study funder had no role in the design of the study; the collection, analysis, and interpretation of the data; and the decision to submit for publication.

Data Availability

Data supporting the study reported in this paper can be made available in deidentified form subject to establishing a data use agreement with the Wake Forest University School of Medicine. The code supporting this study can be accessed on GitHub [54].

Authors' Contributions

All authors take responsibility for the manuscript content, made critical revisions or contributed important intellectual content, and took the decision to submit the final manuscript. KLF, TKH, SLC, JC, ECD, and ELS obtained funding for this study. JC, TKH, and SLC conceptualized the study. JC designed and developed the electronic health record metrics, with advice from TKH and input from all coauthors. AB, AM, BO, SCB, and ERH obtained or provided data. JC, TKH, SLC, and AD analyzed data. JC and TKH visualized and interpreted the results. SLC, AD, SCB, KLF, ECD, and ELS provided critical feedback on metrics design and result interpretation. JC, TKH, SLC, and KLF wrote the first draft of the manuscript. JC, TKH, SLC, AD, SCB, KLF, BO, ERH, AB, AM, ECD, and ELS contributed to the manuscript revision.

Conflicts of Interest

AD serves as an EHR Consultant for the AAMC CORE program. AD is a co-inventor of WHIRL, which is licensed to IllumiCare, Inc. They have an ownership interest in the WHIRL application. AD is also a co-inventor of mPATH. They have equity in Digital Health Navigation (DHN) Solutions, which has licensed mPATH. None of these potential COIs overlap in any way with the content of the current study.

Multimedia Appendix 1

Clinical workflows associated with tobacco cessation alerts.

Multimedia Appendix 2

Electronic health record activity metrics.

Multimedia Appendix 3

Encounter types relevant to tobacco use screening for patients with cancer.

References


34. Stata Statistical Software: Release 15. Stata. College Station, TX: StataCorp LLC; 2017. URL: https://www.stata.com [accessed 2023-02-03]


**Abbreviations**

BPA: Best Practice Advisory  
C3I: Cancer Center Cessation Initiative  
CDS: clinical decision support  
e-alerts: electronic alerts  
EHR: electronic health record  
NCI: National Cancer Institute  
TCCOE: Tobacco Control Center of Excellence

---

Edited by C Perrin; submitted 29.09.22; peer-reviewed by J Hefner; comments to author 21.10.22; revised version received 21.11.22; accepted 18.01.23; published 02.03.23.

Please cite as:  
Chen J, Cutriona SL, Dharod A, Bunch SC, Foley KL, Ostatiewski B, Hale ER, Bridges A, Moses A, Donny EC, Sutfin EL, Houston TK, iDAPT Implementation Science Center for Cancer Control  
Monitoring the Implementation of Tobacco Cessation Support Tools: Using Novel Electronic Health Record Activity Metrics  
JMI Med Inform 2023;11:e43097  
URL: https://medinform.jmir.org/2023/1/e43097  
doi:10.2196/43097  
PMID:36862466

©Jinying Chen, Sarah L Cutriona, Ajay Dharod, Stephanie C Bunch, Kristie L Foley, Brian Ostatiewski, Erica R Hale, Aaron Bridges, Adam Moses, Eric C Donny, Erin L Sutfin, Thomas K Houston, iDAPT Implementation Science Center for Cancer Control. Originally published in JMI Med Informatics (https://medinform.jmir.org), 02.03.2023. 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 JMI Med 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.
Scalable Causal Structure Learning: Scoping Review of Traditional and Deep Learning Algorithms and New Opportunities in Biomedicine

Pulakesh Upadhyaya¹,², PhD; Kai Zhang¹, PhD; Can Li¹, MSPH; Xiaqian Jiang¹, PhD; Yejin Kim¹, PhD

¹School of Biomedical Informatics, University of Texas Health Science Center at Houston, HOUSTON, TX, United States
²Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA, United States

Corresponding Author:
Pulakesh Upadhyaya, PhD
Department of Biomedical Informatics
Emory University School of Medicine
101 Woodruff Circle
Suite 4127
Atlanta, GA, 30322
United States
Phone: 1 9794225161
Email: pulakeshupadhya@gmail.com

Abstract

Background: Causal structure learning refers to a process of identifying causal structures from observational data, and it can have multiple applications in biomedicine and health care.

Objective: This paper provides a practical review and tutorial on scalable causal structure learning models with examples of real-world data to help health care audiences understand and apply them.

Methods: We reviewed traditional (combinatorial and score-based) methods for causal structure discovery and machine learning–based schemes. Various traditional approaches have been studied to tackle this problem, the most important among these being the Peter Spirtes and Clark Glymour algorithms. This was followed by analyzing the literature on score-based methods, which are computationally faster. Owing to the continuous constraint on acyclicity, there are new deep learning approaches to the problem in addition to traditional and score-based methods. Such methods can also offer scalability, particularly when there is a large amount of data involving multiple variables. Using our own evaluation metrics and experiments on linear, nonlinear, and benchmark Sachs data, we aimed to highlight the various advantages and disadvantages associated with these methods for the health care community. We also highlighted recent developments in biomedicine where causal structure learning can be applied to discover structures such as gene networks, brain connectivity networks, and those in cancer epidemiology.

Results: We also compared the performance of traditional and machine learning–based algorithms for causal discovery over some benchmark data sets. Directed Acyclic Graph-Graph Neural Network has the lowest structural hamming distance (19) and false positive rate (0.13) based on the Sachs data set, whereas Greedy Equivalence Search and Max-Min Hill Climbing have the best false discovery rate (0.68) and true positive rate (0.56), respectively.

Conclusions: Machine learning–based approaches, including deep learning, have many advantages over traditional approaches, such as scalability, including a greater number of variables, and potentially being applied in a wide range of biomedical applications, such as genetics, if sufficient data are available. Furthermore, these models are more flexible than traditional models and are poised to positively affect many applications in the future.

(JMIR Med Inform 2023;11:e38266) doi:10.2196/38266

KEYWORDS
causal inference; causal structure discovery; deep learning; biomedicine; networks
Introduction

Background

Many applications in biomedicine require the knowledge of the underlying causal relationship between various factors beyond association or correlation. Randomized controlled trials are widely used to uncover causality, but these experiments can be prohibitively expensive or unethical in many cases. Therefore, it has sparked an enormous amount of interest in identifying causal effects from observational data [1-3].

In this paper, we discuss causal structure learning; that is, learning causal relationships that are represented as directed graph structures between different factors and its application to biomedicine. The causal structure is represented by a causal graph (also called a causal Bayesian network), which is a directed acyclic graph (DAG), in which the nodes represent variables and edges represent causation (Figure 1). An edge is drawn from a variable that represents the cause to a variable that represents the effect of that cause. Based on a variety of methodologies, causal structure learning identifies which causal models represented by DAGs accurately represent the observed data.

For example, consider the example of a gene regulatory network [4-7], which is an abstract representation of the gene regulation processes as shown in Figure 1. By observing the data of multiple variables such as gene expression profiles, causal structure learning attempts to discover causal relationships among the genes. For example, if a gene A regulates another gene B, it is represented by an arrow between gene A and gene B.

Many researchers in the biomedical field are interested in causality and not just correlation (eg, whether a particular treatment affects a particular outcome). Unlike association- or correlation-based studies that simply indicate that any 2 variables are correlated, this approach seeks to determine the directional relationship between any 2 variables (eg, between a treatment variable and an outcome variable). In biomedicine, causal structure learning can be applied in a variety of applications.
Examples

Gene Regulatory Networks

A gene regulatory network is a network in which molecular regulators and genes are the nodes, and the directed edges denote the interactions among them [5]. This is in contrast to association-based methods such as finding correlation or mutual information among the genes (finding Pearson, Kendall, Spearman correlation coefficients, etc) that do not have any directional information [8]. Such methods can only be accurate to a certain extent when it comes to deducing extensive gene regulatory structures from data sets with a large set of observations. Correlational studies can only indicate gene-gene association and not the direction of regulation. A gene regulatory network is an example of a causal structure that can be used to develop interventions to control gene expression.

Causal structure learning algorithms have been used to jointly deduce the phenotype network structure and directional genetic architecture [9]. It uses a difference causal inference method and compares it with another causal structure learning algorithm (difference-based Greedy Equivalence Search [GES]) as a baseline. Another study proposed a hybrid algorithm that combines Simulated Annealing with Greedy Algorithm to predict intergene transcriptional regulatory relationships [10], which are also directional in nature. In cancer, somatic genome alterations and differentially expressed genes have causal relationships. A correlational study cannot provide directional information in any of these applications.

The tumor-specific causal inference algorithm proposed by Xue et al [11] uses a Bayesian causal learning framework to find those relationships. Unlike association-based studies, this study is based on a causal structure learning framework across the
whole genome where Ha et al [12] found gene signatures that were the causes of clinical outcomes and were not merely correlated to them. Apart from these examples, there are also networks such as those represented in the Sachs data set [13] that simultaneously incorporates measurements of 11 phosphorylated proteins and phospholipids to find causal pathways linking them. This is different from association-based correlation studies because protein signaling pathways are directional.

In our comparative analysis of the performance of this data set, we found that machine learning models can also be effective at finding causal structures (details are available in the Results section). In the case of more complicated protein signaling networks with many nodes, machine learning–based methods might be particularly effective.

**Brain Connectivity Networks**

Different regions of the brain have distinct functions. Previous studies have used correlation-based methods [14] to find nondirectional functional connectivity among cortical regions. Spatial localization of brain functions has been studied using methods such as functional magnetic resonance imaging [15]. Regions within the brain are the nodes, and a directed edge between regions represents some functional connection (see Figure 2 in the paper by Brovelli et al [16] for the difference between coherence and causality graphs). Such connections are directional, can have different strengths (weights), and can be both inhibitory or excitatory [17]. Scalable causal structure learning models can also model such connection strengths in addition to directionality, which makes them more expressive than an association. In addition, brains have large-scale structural cortical networks that are directional with respect to information flow and can only be captured by causal structure instead of correlation.

**Epidemiology**

Causal structure learning has also been used in epidemiology with patients’ medical records. Many complex diseases are multifactorial in which a combination of these factors contributes to disease predisposition. Causal structure learning considers multiple confounders to determine causal effects solely from one factor of interest to another. For example, causal structure has been used to disentangle psychological factors that predispose adolescents to smartphone addiction [18]. Incorporating a large set of medical claim records, a recent study used a scalable causal structure learning to elucidate the clinical pathways from comorbid illnesses to Alzheimer disease [19].

**Challenges**

However, there are a few challenges. The general approach to solving this problem of learning a DAG from data, which has been studied for a long time [20], has a time complexity that scales exponentially with the number of observed variables. This is because the problem is generally nondeterministic polynomial-time complete [21]. In practice, if the number of variables is greater than a few hundred, the problem becomes intractable to solve optimally.

Several approaches have been used to solve this problem of intractable time complexity. Traditionally, constraint-based and score-based methods, which search for the optimal graph from a discrete space of candidate graphs, have been used to learn the DAG from data. Constraint-based methods such as the Peter Spirtes and Clark Glymour (PC) and Fast Causal Inference
(FCI) algorithms (which will be discussed in detail in the *Discrete Space Algorithms* section) rely on statistical tests to estimate the correct causal structure. However, biological data usually involve hundreds to thousands of variables, and the complexity of algorithms increases exponentially as the number of variables increases. For example, typical human RNA sequence data contain at least 20,000 genes. Therefore, the complexity of the PC algorithm is proportional to $2^{20000}$, which is infeasible within a reasonable amount of time.

Hence, researchers have investigated various score-based methods that assign scores based on the data to each DAG and select the one with the best score. Although score-based methods scale better than constraint-based methods, they do not scale well for several thousand variables. On the other hand, patient medical records in electronic health records or claim data raise severe scalability concerns, because they include up to 144,000 International Classification of Diseases-Ninth Revision or 69,823 International Classification of Diseases-Tenth Revision diagnosis codes, >2000 US Food and Drug Administration–approved drugs, and >10,000 Current Procedural Terminology procedures or laboratory test codes.

To overcome the limited scalability of traditional methods, recent advances in machine learning algorithms have relaxed the problem of finding an optimal DAG into a continuous optimization problem with smooth acyclicity constraints. This enables the use of nonheuristic machine learning (including deep learning) algorithms to determine the optimal causal structure. This is a promising development in the field of biomedicine. In this study, we focus on scalable algorithms. Table 1 summarizes the algorithms discussed in this study. The tools available for some of these algorithms are listed in *Multimedia Appendix 1*. A list of ground truth causal structures can be found in the *bnlearn* repository [22].

There are 2 distinct approaches in the context of treatment effect evaluation: the structural approach and potential outcome framework approach [23]. In this study, we consider the first approach, in which there are 2 distinct types of algorithms for finding the causal DAG structure. In all of these examples, the goal is to learn a DAG that shows the directional relationship among variables from observational data.
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>( \text{DS}^a )</th>
<th>( \text{CS}^b )</th>
<th>Summary</th>
<th>Remarks</th>
<th>Scalability</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC(^c)</td>
<td>✓</td>
<td>✗</td>
<td>A partially directed acyclic graph (CPDAG(^d)) is produced by iteratively checking the conditional independence conditions of adjacent nodes, conditioned on an all-size subset of neighbors.</td>
<td>Outputs usually converge to the same equivalence class; high FPR(^e) on experimental data</td>
<td>+(^f)</td>
</tr>
<tr>
<td>1C(^f)</td>
<td>✓</td>
<td>✗</td>
<td>Returns the equivalent class of the DAG(^b) based on the estimated probability distribution of random variables and an underlying DAG structure.</td>
<td>Outputs usually converge to the same equivalence class.</td>
<td>+</td>
</tr>
<tr>
<td>FCI(^i)</td>
<td>✓</td>
<td>✗</td>
<td>Modified PC algorithm to detect unknown confounding variables and produces asymptotically correct results.</td>
<td>Faster than PC with similar TPR(^j); converges to the same asymptotic result; high experimental FPR</td>
<td>++</td>
</tr>
<tr>
<td>GES(^k)</td>
<td>✓</td>
<td>✗</td>
<td>Starts with an empty graph and iteratively adds and deletes edges in the graph by optimizing a score function.</td>
<td>Faster than PC with higher TPR; stable result for the same score function</td>
<td>++</td>
</tr>
<tr>
<td>Fast GES</td>
<td>✓</td>
<td>✗</td>
<td>Improved and parallelized version of GES</td>
<td>Faster than GES; same TPR; stable result for the same score function</td>
<td>++</td>
</tr>
<tr>
<td>K2</td>
<td>✓</td>
<td>✗</td>
<td>Performs a greedy heuristic search for each nodes’ parents.</td>
<td>Greedy searches might return very suboptimal solutions.</td>
<td>++</td>
</tr>
<tr>
<td>MMHC(^l)</td>
<td>✓</td>
<td>✗</td>
<td>MMHC to find the skeleton of the network and constrained greedy search for edge orientation.</td>
<td>Greedy searches might return suboptimal solutions.</td>
<td>+</td>
</tr>
<tr>
<td>LiNGAM(^m)</td>
<td>✓</td>
<td>✗</td>
<td>Transfer the linear structure model ( \text{LiNGAM} ) to the form ( \text{LiNGAM} ) and optimize for matrix ( \text{LiNGAM} ).</td>
<td>Works very well on linear data but not on nonlinear data.</td>
<td>++</td>
</tr>
<tr>
<td>NOTEARS</td>
<td>✗</td>
<td>✓</td>
<td>Uses smooth function ( h(A) ), whose value characterizes the “DAG-ness” of the graph with adjacency matrix ( A )—that is, ( h(A)=0 ) for DAG—and optimizes using continuous optimization.</td>
<td>Might converge to many different DAGs; GPUs(^n) can speed up the process.</td>
<td>+++</td>
</tr>
<tr>
<td>NOBEARS</td>
<td>✗</td>
<td>✓</td>
<td>Proposed a new acyclicity constraint that allows for faster optimization and scalability, and a polynomial regression loss to infer gene regulatory networks from nonlinear gene expressions.</td>
<td>Might converge to many different DAGs; GPUs can speed up the process.</td>
<td>+++</td>
</tr>
<tr>
<td>DAG-GNN(^o)</td>
<td>✗</td>
<td>✓</td>
<td>Uses an autoencoder framework and deep learning to train it and infer the causal structure from the weights of the trained network and is more scalable than NOTEARS.</td>
<td>Might converge to many different DAGs; GPUs can speed up the process.</td>
<td>++++</td>
</tr>
<tr>
<td>NOFEARS</td>
<td>✗</td>
<td>✓</td>
<td>Modify NOTEARS so the scoring function remains convex to ensure local minima.</td>
<td>Might converge to many different DAGs; GPUs can speed up the process.</td>
<td>++++</td>
</tr>
<tr>
<td>GAE(^p)</td>
<td>✗</td>
<td>✓</td>
<td>Scalable graph autoencoder framework (GAE) whose training time increases linearly with the number of variable nodes.</td>
<td>Good accuracy; might converge to many different DAGs; GPUs can speed up the process.</td>
<td>++++</td>
</tr>
<tr>
<td>GRAN-DAG(^q)</td>
<td>✗</td>
<td>✓</td>
<td>Extends the NOTEARS algorithm for nonlinear relationships.</td>
<td>Works on nonlinear data; better accuracy than NOTEARS; might converge to many different DAGs; GPUs can speed up the process.</td>
<td>++++</td>
</tr>
<tr>
<td>CGNN(^t)</td>
<td>✗</td>
<td>✓</td>
<td>Generative model of the joint distribution of variables reducing MMD(^r) between the graph and data.</td>
<td>Does not always converge to a single class of equivalent DAGs; GPUs can speed up the process.</td>
<td>++++</td>
</tr>
<tr>
<td>SAM(^t)</td>
<td>✗</td>
<td>✓</td>
<td>Structurally agnostic model for causal discovery and penalized adversarial learning.</td>
<td>Does not always converge to a single class of equivalent DAGs; GPUs can speed up the process.</td>
<td>++++</td>
</tr>
<tr>
<td>RL-BIC(^u)</td>
<td>✗</td>
<td>✓</td>
<td>Reinforcement learning-based algorithm that uses both the acyclicity constraint and the BIC(^v) score.</td>
<td>Very good accuracy; does not always converge to a single class of equivalent DAGs; GPUs can speed up the process.</td>
<td>++++</td>
</tr>
</tbody>
</table>
Paper Structure

This study attempts to provide a comparative study of various scalable algorithms that are used to discover causal structures from observational data to the biomedicine community. Some of these traditional and score-based methods have been extensively studied [24], but many of the algorithms discussed here focus on scalable causal structure learning. Although we do not list all possible approaches as Vowels et al [25], we sample a few important algorithms and evaluate their performance on synthetic data sets and the Sachs data set [13].

This tutorial paper presents algorithms for causal structure identification in biomedical informatics. In the Methods section, we discuss the methodology and examine the traditional algorithms that determine the optimal causal graph in a discrete space. We also discuss algorithms that use continuous space optimization to discover causal relationships. We compare the performance of these algorithms in the Results section. Finally, we present the discussion and conclusions. A brief overview of the methods and results is presented in Figure 2.

Methods

Overview

In this section, we discuss 2 paradigms of algorithms for causal structure learning. First, we consider algorithms that search for the optimal DAG in the discrete space of all possible DAGs (space of all possible discrete DAGs for a given number of variable nodes) or discrete space algorithms. Second, we consider scalable algorithms that use continuous optimization methods to find the optimal DAG (ie, algorithms that search the continuous space of all possible weighted DAGs to find the optimal one), known as continuous space algorithms.

Discrete Space Algorithms

Overview

The first type that we discuss in this section is discrete space algorithms for causal discovery; that is, algorithms that search for the optimal DAG in the discrete space of candidate causal graphs. This is in contrast to continuous space algorithms (discussed in the Continuous Space Algorithms section) that search for the optimal DAG from the continuous space of weighted candidate graphs.

The discrete space algorithms can be divided into the following 4 types: combinatorial constraint-based models, score-based models, hybrid models, and functional models. In combinatorial constrained-based methods, we consider methods that check the conditional independence relations of 2 adjacent nodes conditioned on all-size subsets of variables. Score-based methods perform optimization by considering a score representing the goodness of fit and can handle more variables than constraint-based methods. Hybrid methods combine constraint- and score-based algorithms. Functional models find structural equations to describe the causal relationship and are useful mostly when the variables can be assumed to be expressed by some linear or nonlinear equations.

Combinatorial Methods

We now focus on combinatorial optimization methods, where conditional independence relationships in the data are used for finding the optimal DAG.

PC and Its Variants

The PC algorithm was proposed by PC and is named after them [26]. This algorithm produces a completed partially DAG (CPDAG) by iteratively checking the conditional independence relations of 2 adjacent nodes conditioned on all-size subsets of variables.
their neighbors. Three assumptions underlie the algorithm: no confounder variable, the causal Markov condition, and faithfulness. Under these conditions, this algorithm generates a partially directed causal graph that is proven to be asymptotically correct.

The PC algorithm is order-dependent; that is, the output of the algorithm can depend on the order in which the variables are provided to the algorithm. To address this problem, Colombo and Maathuis [27] developed a PC-stable algorithm in which the deletion of an edge takes place at the end of each stage (considering any 2 nodes’ relations within a predetermined neighborhood). Thus, any ordering of vertices will result in the same edge deletions, resulting in the same stable output. The PCMI and PCMI+ [27-29] are 2 extensions of the PC algorithm proposed to handle large-scale time-series data sets.

**Inductive Causation Algorithm and Its Variants**

The inductive causation (IC) algorithm uses the estimated probability distribution of random variables with an underlying DAG structure and outputs the equivalent class of the DAG. In contrast, PC provides a schematic search method and is thus considered a refinement of the IC.

The IC* algorithm [30,31] is an extension of the IC algorithm, which searches for causal relations using observations of a set of variables, even when they appear as latent variables. The output of the IC algorithm is a CPDAG that only has directed edges (identified causation) and undirected edges (undetermined causation). The output of the IC* algorithm is an embedded pattern; that is, a hybrid graph containing ≥2 types of edges.

**FCI and Its Variants**

The FCI is a modification of the PC algorithm [30,32] that detects unknown confounding variables and produces asymptotically correct results. FCI improves the PC algorithm by adopting 2 rounds of phases of the PC algorithm. The algorithm first uses PC-phase I to find an initial skeleton, then uses the separation set to orient all v-structure triples (a→c←b) and outputs a CPDAG; then performs another round of skeleton searching based on the CPDAG and repeats the orientation for unshielded triples. The really fast causal inference algorithm [33,34] skips the second step, which is the most time-consuming part of the task, and therefore significantly accelerates the FCI procedure. A set of 10 rules was added to the algorithm to orient the edges of the skeleton.

**Score-Based Methods**

**Overview**

In addition to traditional combinatorial methods such as PC and FCI, score-based methods have also been used to uncover causal structures. In these methods, algorithms determine the optimal DAG by optimizing a particular score.

A typical score function is the Bayesian information criterion (BIC) score. The GES algorithm uses different score functions for different data types as follows: the BIC score (for continuous data), likelihood-equivalence Bayesian Dirichlet uniform joint distribution score (for discrete data), and Conditional Gaussian score (for continuous or discrete mixture data).

where $L_{max}$ is the maximized likelihood function of the model, $n$ is the number of observational data points, and $k$ is the degree of freedom. The definition of the Bayesian Dirichlet uniform joint distribution scoring function was found in a study by Buntine [35]. The conditional Gaussian score is defined on the ratios of joint distributions, and Andrews et al [36] have proved that the Conditional Gaussian score is score equivalent; that is, a scoring function that scores all DAGs in the same Markov Equivalence Class equally.

Score-based methods include the GES algorithm, the fast GES algorithm, and the K2 algorithm.

**GES Algorithm**

The GES algorithm was proposed by Chickering [37], and its underlying principles were obtained from Meek [37,38]. The algorithm starts with an empty graph and iteratively adds and deletes the edges in the graph by optimizing a score function. During the forward phase, the algorithm searches iteratively from the space of the DAGs created by one edge addition on the current DAG and selects the edge with the best score. The forward phase ends when the score is no longer increasing. In the second phase, the algorithm repeats the above step but deletes one edge at a time and selects the edge that improves the score the most. The algorithm stops as soon as there are no more edges to be deleted.

**Fast GES Algorithm**

Fast GES is an improved and parallelized version of the GES. Significant speedup was achieved by storing the score information during the GES algorithm [39]. In addition, several insights regarding parallelization were offered in the paper. First, the precalculation of covariances can be parallelized by variables. Second, it is possible to parallelize the process of calculating the edge scores when an edge addition is being performed on the graph. A greater speedup can be achieved for sparse graphs.

**K2 Algorithm**

The main idea of the K2 algorithm [40] is to perform a greedy heuristic search of the parents of each node. For each node, the algorithm iteratively determines the parents. When visiting node $X_i$, the algorithm searches for all possible parents of $X_i$ ($X_j$ such that $j$ has a lower ordering of $i$). The algorithm greedily adds $X_j$ to the parent set of $X_i$ if it could increase a predefined score function. The iteration for node $X_i$ stops when the number of parent nodes reaches the (preset) maximum or when adding an $X_j$ does not increase the score anymore. The entire algorithm finishes after completing the iteration for all $X_i$.

**Hybrid Algorithms**

Hybrid algorithms use a combination of score-based and combinatorial constraint-based optimization methods to determine the optimal DAG. An example is the Max-Min Hill Climbing (MMHC) algorithm. The MMHC algorithm is a combination of constraint- and score-based algorithms [41]. It uses the Max-Min Parents and Children algorithm. A detailed
description is provided in [41,42] to find the skeleton of the Bayesian network and then perform the constrained greedy search to orient the edges.

**Algorithms for Functional Causal Models**

**Overview**
Functional causal models or structural equation models (SEMs) assume structural equations that define the causal relationships. Such structural equations may describe the linear and nonlinear relationships among variables. In addition to the discrete methods discussed here, SEMs are also an important assumption in many machine learning–based methods that use the continuous optimization techniques in the Continuous Space Algorithms section.

**Linear Non-Gaussian Acyclic Model**
The linear non-Gaussian acyclic model (LiNGAM) was originally proposed by Shimizu [43] to learn linear non-Gaussian acyclic causal graphs from continuous-valued data. The LiNGAM transfers to the form of , and the causal structure problem becomes an optimization problem for matrix B. There are several extensions of the LiNGAM model using different estimation methods, including independent component analysis–based LiNGAM [43,44], DirectLiNGAM [45], and Pairwise LiNGAM [46].

**Additive Noise Models**
A nonlinear additive noise model is proposed in [47]. The model assumes that the observed data are generated according to the following equation:

\[ y_i = f_i(x_i) + e_i \]

where \( f_i \) is an arbitrary function, \( x_{pa(i)} \) denotes the ancestor nodes of node \( x_i \) in the true causal graph, and \( e_i \) is the noise variable of an arbitrary probability density function. This study proves the basic identifiability principle for the 2 variables case and generalizes the results to multiple variables.

**Continuous Space Algorithms**

**Overview**
Traditional causal discovery algorithms attempt to discover a causal graph, which is usually a DAG, while searching for an optimal graph in the space of candidate graphs. The score-based optimization problem of DAG learning (discussed in the Score-Based Methods section) is mathematically given by the following equation:

\[ \sum_{G \in \mathcal{G}} h(G) \]

Here, \( \mathcal{G} \) is the set of all DAGs with \( n \) nodes, and \( h(G) \) is the cost or score function. The problem of searching for all DAGs is usually intractable and superexponential in the number of nodes in the graph.

An alternative approach would be to model the problem as a continuous space optimization problem, which would then allow the application of various learning techniques. Recently, several publications have explored continuous optimization methods that learn DAGs by adding an acyclicity constraint. In these approaches, the discrete acyclicity constraint \[ \square \] is replaced by \[ \square \], where \( h(A) \) is a smooth function that ensures acyclicity of \( G(A) \).

The hard constraints on acyclicity can be relaxed and incorporated into the loss function to be optimized. This smooth continuous constraint allows the use of machine learning–based tools, which in turn can make the algorithms scalable in the presence of substantial amounts of data. These algorithms are based on SEMs.

**NOBEARS Algorithm**
Several other improvements such as the NOBEARS algorithm [48] have improved the scalability of the NOTEARS algorithm. A fast approximation of a new constraint is proposed, and a polynomial regression loss model is proposed to account for nonlinearity in gene expression to infer gene regulatory networks.

**NOTEARS Algorithm**
This algorithm considers the acyclicity constraint and comes up with the constraint.

Here, \[ \square \] is the element-wise product. \( h(A) \) equals 0 if and only if \( G(A) \) is acyclic, and more severe deviations from acyclicity would increase the value of the function. This study assumes a linear SEM:

\[ X_i = f_i(Z_i) + e_i \]

Here, \( X_i \) is a \( d \)-dimensional sample vector of the joint distribution of \( d \) variables and \( Z_i \) is a \( d \)-dimensional noise vector. We denote \( n \) such samples by matrix \( X \), and the loss function (with \( l_1 \)-regularization) is given as follows:

\[ \sum_{i=1}^{n} h(A) = 0 \]

The constraint is given by \( h(A)=0 \) and is used in the final Lagrangian formulation of the loss function. The paper on learning sparse nonparametric DAGs is an extension of NOTEARS, which tries to define a “surrogate” of the matrix above for general nonparametric models to optimize [49].

**Directed Acyclic Graph-Graph Neural Network Algorithm**
A Directed Acyclic Graph-Graph Neural Network (DAG-GNN) [50] generalizes the NOTEARS algorithm by considering the nonlinearity in the SEMs. It can be modeled with a variational autoencoder neural network with a special structure, with an encoder \[ \square \], and a decoder \[ \square \] and where \( g_1,g_2 \) are parameterized functions that can be assumed to serve as the inverse of \( f_1,f_2 \), respectively.

This variational framework considers \( Z \) to be a latent vector (instead of viewing it as noise in linear SEMs), which can have dimensions other than \( d \). The decoder then attempts to
reconstruct the data from this latent variable. The encoder and decoder can be trained together from $n$ samples of $\mathbf{Z}$ such that the loss function:

$$\text{is minimized, where } KLD \text{ is the Kullback-Liebler Divergence.}$$

The constraint in this optimization process to ensure the acyclicity of matrix $A$ is slightly modified to:

$$\alpha$$

where $\alpha$ is an arbitrary parameter. This constraint can be implemented more easily in graphical processing unit-based deep learning libraries owing to the algorithm’s parallelizability and scalability of the algorithm.

**NOFEARS Algorithm**

Wei et al [51] demonstrated that the NOTEARS algorithm fails to satisfy the Karush-Kuhn-Tucker regularity conditions. Therefore, they reformulated the problem to ensure that the convexity of the scoring function can still ensure local minima even when the constraints are nonconvex. This new algorithm called the NOFEARS algorithm has the following acyclicity constraint.

**Graph Autoencoder**

Ng et al [52] propose another graph autoencoder (GAE) framework for causal structure learning, which improves the training speed and performance over DAG-GNN for both linear and nonlinear synthetic data sets.

Some other similar machine learning–based continuous learning algorithms include gradient-based neural DAG [53], Causal Generative Neural Network [54], and structurally agnostic model [55].

**Reinforcement Learning-Based Methods**

Reinforcement learning-based methods have been proposed recently that consider both the acyclicity constraint and BIC score in the reward function and attempt to learn the DAG [56]. They used an actor-critic model, where the actor is an encoder-decoder framework that takes data as input and outputs the graph. The critic uses the reward function for this graph and updates the proposed graph.

**Results**

This section provides the results to compare the effectiveness of some causal structure learning algorithms on synthetic and real data.

**Benchmark Methods**

The synthetic data were generated in the same manner as in the DAG-GNN paper [50]. An Erdos-Renyi model with an expected node degree of 3 was used to generate the random graph, and the adjacency matrix was formed by assigning weights to the edges from a uniform distribution. The samples were generated using the following structural equation:

$$\text{Here, } Z \text{ is random Gaussian noise. We consider 2 functions for } g(X). \text{ The first is a (linear) identity function:}$$

$$\text{and the second is a nonlinear function}$$

We considered 5 data sets for both linear and nonlinear functions. For each data set, we generated $n=5000$ independent samples according to the above equations. We used 6 algorithms, 4 of which are discrete space algorithms. PC and Greedy Fast Causal Interface (GFCI) are constraint-based methods, GES is a score-based method, and MMHC is a hybrid method. We also considered 2 continuous space methods, DAG-GNN and GAE.

We also evaluated these algorithms on the publicly available Sachs data set [13] using the above 4 metrics and showed the results are shown in Table 2. For other data sets that have the ground truth but are not covered in our experiments, please refer to the bnlearn repository [22]. The algorithms were implemented in Python for machine learning–based continuous space methods and R for discrete space algorithms.

**Table 2. Benchmark experiments on the Sachs data set.** We evaluated 6 algorithms—Peter Spirtes and Clark Glymour (PC), Greedy Equivalence Search (GES), Greedy Fast Causal Interface (GFCI), Max-Min Hill Climbing (MMHC), Directed Acrylic Graph-Graph Neural Network (DAG-GNN), and graph auto encoder (GAE), on 4 metrics—structural hamming distance (SHD), true positive rate (TPR), false positive rate (FPR), and false discovery rate (FDR)—and show their results in Figure 3. In all these evaluations, we consider any edge whose direction is reversed as half discovered.

<table>
<thead>
<tr>
<th>Metric</th>
<th>PC</th>
<th>GES</th>
<th>GFCI</th>
<th>MMHC</th>
<th>DAG-GNN</th>
<th>GAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHD ($\downarrow$)</td>
<td>24.50</td>
<td>26.50</td>
<td>29.50</td>
<td>22.00</td>
<td>19.00 $^a$</td>
<td>22.00</td>
</tr>
<tr>
<td>FDR ($\downarrow$)</td>
<td>0.77</td>
<td>0.72</td>
<td>0.79</td>
<td>0.68</td>
<td>0.71</td>
<td>0.89</td>
</tr>
<tr>
<td>TPR ($\uparrow$)</td>
<td>0.32</td>
<td>0.56</td>
<td>0.44</td>
<td>0.47</td>
<td>0.11</td>
<td>0.05</td>
</tr>
<tr>
<td>FPR ($\downarrow$)</td>
<td>0.49</td>
<td>0.64</td>
<td>0.72</td>
<td>0.45</td>
<td>0.13</td>
<td>0.21</td>
</tr>
</tbody>
</table>

$^a$Italicized values represent the best results for each metric.
Figure 3. Accuracy comparison. We evaluated 6 algorithms: Peter Spirtes and Clark Glymour (PC), Greedy Equivalence Search (GES), Greedy Fast Causal Interface (GFCI), Max-Min Hill Climbing (MMHC), Directed Acrylic Graph-Graph Neural Network (DAG-GNN), and graph auto encoder (GAE) on 4 metrics, structural hamming distance (SHD), true positive rate (TPR), false positive rate (FPR), and false discovery rate (FDR). In all these evaluations, we considered any bidirectional edges as half discovered. In experiments (A-D), first column, the data are drawn from a distribution according to the underlying causal graph where relationships between nodes are linear, and experiments (E-H), second column, is for the nonlinear case. In all experiments, the number of nodes of the graph ranges from 10, 20, 50, to 100. For each graph size, we drew 5 different data sets from the graph structure with a sample size of 1000 and calculated 4 evaluation metrics and obtained the average.

Observations
All algorithms were tested on both linear and nonlinear data. The accuracies of some of these algorithms are shown in Figure 3. We used the following 4 evaluation measures: structural hamming distance (SHD), true positive rate (TPR), false positive rate (FPR), and false discovery rate (FDR). SHD refers to the number of edge insertions, deletions, and reversals. In our case, we used a modified SHD, where a reversal contributes half of the SHD score instead of 1. The TPR is the ratio of the algorithm’s correctly discovered edges to the number of edges in the ground truth graph. FPR is the ratio of the algorithm’s falsely discovered edges to the number of nonedges in the ground truth graph. FDR is the ratio of the algorithm’s falsely discovered edges to the total number of discovered edges. We also evaluated these algorithms on the Sachs data set using the above 4 metrics, and the results are shown in Table 2.

Time Complexity
The relative scalability of different algorithms is presented in Table 1. The worst-case time complexity of the PC, GES, and GFCI algorithms was $O(m^3)$, where m is the number of variables (nodes in the DAG). For the GES, the best-case time complexity was $O(n^2 \log n)$. For the GAE and DAG-GNN, the time complexity of the algorithm is $O(kV)$, where k is the number of iterations. The time complexity for MMHC is $O(|V|^2)$, where V is the set of
variables, \( S \) is the largest set of parents and children, and \( I \) is a parameter of the algorithm that denotes the size of the largest conditioned subset [41].

In our experiments, we observed that in the worst-case scenario, the running time for a maximum of 100 variables was of the order of hours for MMHC and of the order of minutes for the other algorithms. However, as the number of variables increases to a few thousand, machine learning–based methods such as DAG-GNN and GAE can provide solutions in a reasonable time. The trade-off between complexity (number of iterations) and accuracy can provide a choice between a method that is less accurate but faster or vice versa.

**Discussion**

**Interpretation of Results**

It is clear from the results that the algorithms have different advantages and disadvantages. Although the PC algorithm performs well across both linear and nonlinear data, it has a low TPR and is computationally intensive. The GES, GFCI, and MMHC algorithms show a very high FPR, but their TPR is higher than that of the PC algorithm. The SHD of the 2 machine learning–based methods—DAG-GNN and GAE—was also considerably lower for both data sets.

Continuous constraint-based algorithms generally exhibit a very low FDR, except for the benchmark Sachs data set. This is generally because both linear and nonlinear models are based on SEMs with the same causal relationship function at every node, which is an algorithm assumption when they learn the causal structure, but one cannot guarantee the same for Sachs data [13], because such constraints cannot be defined a priori.

This is corroborated by recent results from Zhu et al [56] where such gradient-based methods performed poorly on data generated by a nonlinear model, in which every causal relationship (node function) was sampled from a Gaussian distribution. However, this is a growing research area. In general, in areas such as gene regulatory networks and brain connectivity networks where the number of variables is large, machine learning–based methods can provide comparable results to traditional methods with a much more efficient time complexity and scalability.

**Challenges**

Machine learning for causal structure learning is not without its limitations, which may present several challenges. First, in many applications, there is no ground truth about causal structure, which makes it difficult to evaluate the performance of these algorithms. Furthermore, many scalable methods use stochastic gradient descent; thus, the final output graph is not always deterministic. When the number of data samples or variables is low, traditional or score-based methods are a better choice, especially when the application requires fewer false positives. For the PC, GES, and GFCI algorithms, we observed that these algorithms require considerable running time, as the number of variables is more than 100 [57].

However, when it comes to large samples of data (eg, more than 100,000 samples) or hundreds of variables (eg, in many gene networks), machine learning methods can provide a reasonable solution, because other methods fail owing to scalability issues. As machine learning algorithms are highly parallelizable, the solutions can be computed much faster, particularly through the use of a graphical processing unit. These algorithms are potentially useful for many applications related to genetics and biomedicine, especially those with an abundance of observational data.

The continuous space machine learning models are more scalable and might be useful in the era of big data. Traditional methods might have complexities that grow exponentially with the number of attributes. Despite the nonconvexity of the optimization proposed by Zheng et al [58], optimization and learning strategies can be used to help find the optimal solution. Several methods have been used to solve this problem using augmented Lagrangian approaches [50,52].

The NOBEARS algorithm reduces the computing complexity of NOTEARS from cubic to quadratic in terms of the number of attributes [48], allowing for smooth implementation in data sets that have more than 4000 attributes. The algorithms are also highly parallelizable, and most of the algorithms use deep learning libraries such as Tensorflow [59] and PyTorch [60].

Machine learning techniques for causal discovery, which use continuous space optimization, are an emerging area of research, which can lead to more efficient causal discovery, particularly in applications where directed graphs are used to specify causal relations more clearly. With sufficient data, machine learning models can be robust to certain discrepancies such as sample bias, missing data, and erroneous measurements. Many of these applications have also focused on weaker concepts of causality such as pairwise directionality during the analysis of gene networks and brain connectivity networks [61,62].

It is noteworthy that machine learning methods are usually black box methods, which might provide lesser insight into the process of derivability of the causal structures. For higher interpretability, an option that has been explored is to develop parallel versions of these algorithms, such as PC [63]. In the future, options such as ensemble learning can be explored for the same.

Some other challenges can be found in finding causal structure from data. In the case of learning causal structure from electronic health record data, they might have several problems, such as missing values or noise in the data, which are very common [64]. If the number of missing values or the amount of noise is significant, the application of causal discovery methods might yield unreliable results.

Furthermore, most causal discovery methods assume that the distribution of data is stationary, which may not be true in certain medical applications [65]. Hence, it is very important to consider the aforementioned problems as well as issues related to selection bias before causal structure learning methods are applied. Glymour et al [24] discuss some general guidelines to avoid such problems in causal structure learning. These generalized learning algorithms are ineffective in many biomedical applications, such as in learning biological or gene networks, because they do not consider specific network
constraints. These constraints can be incorporated into causal structure learning methods for greater efficiency.

Conclusions

In this paper, we have discussed the motivation for causal structure discovery in biomedicine as well as some interesting applications. Two paradigms of causal discovery algorithms have been reviewed. Combinatorial or score-based algorithms are used in the first paradigm for optimizing discrete spaces of candidate causal graphs, whereas machine learning algorithms are used in the second paradigm to solve continuous optimization problems with acyclicity constraints. In addition to listing these methods, we have also included resources that readers can use to find appropriate applications. Furthermore, we tested several algorithms against synthetic benchmark data sets and against the Sachs real-world data set and evaluated their relative performances. We have also discussed their theoretical time complexity. Our discussion of the limitations and challenges of various algorithms is intended to offer readers a guide for choosing an algorithm from among the many available options. Finally, we highlight several challenges associated with finding causal structure from real-world data (eg, missing values, nonstationarity, noise, and sampling bias).

Acknowledgments

XJ is a Cancer Prevention and Research Institute of Texas scholar in cancer research (RR180012) and was supported in part by the Christopher Sarofim Family Professorship, University of Texas Stars award, University of Texas Health Science Center startup, and the National Institutes of Health under award numbers R01AG066749 and U01TR002062.

Authors' Contributions

The survey and experiments on deep learning–based methods and the survey on potential applications were conducted by PU. The survey and experiments on traditional methods were conducted by KZ and CL. XJ and YK conceived the study and provided useful inputs for the potential applications of scalable structure learning.

Conflicts of Interest

None declared.

Multimedia Appendix 1

List of tools for causal discovery.

[DOCX File, 21 KB - medinform_v11i1e38266_app1.docx]

References

4. TransmiR v2.0 database. The Cui Lab. URL: http://www.cuilab.cn/transmir [accessed 2021-10-14]


Abbreviations

- **CPDAG**: completed partially directed acyclic graph
- **DAG**: directed acyclic graph
- **DAG-GNN**: Directed Acyclic Graph-Graph Neural Network
- **FCI**: Fast Causal Inference
- **FDR**: false discovery rate
- **FPR**: false positive rate
- **GAE**: graph autoencoder
- **GES**: Greedy Equivalence Search
- **GFCI**: Greedy Fast Causal Interface
- **IC**: inductive causation
- **LiNGAM**: linear non-Gaussian acyclic model
- **MMHC**: Max-Min Hill Climbing
- **PC**: Peter Spirtes and Clark Glymour
- **SEM**: structural equation model
- **SHD**: structural hamming distance
- **TPR**: true positive rate

©Pulakesh Upadhyaya, Kai Zhang, Can Li, Xiaoqian Jiang, Yejin Kim. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 17.01.2023. 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.

Xiaojie Chen¹*, MS; Han Chen²*, PhD; Shan Nan¹, PhD; Xiangtian Kong³, PhD; Huilong Duan¹,4, PhD; Haiyan Zhu²,5, PhD

¹Key Laboratory of Biomedical Engineering of Hainan Province, School of Biomedical Engineering, Hainan University, Haikou, China
²Hainan Hospital of Chinese People’s Liberation Army General Hospital, Sanya, China
³IMWare, Wuhan, China
⁴College of Biomedical Engineering and Instrumental Science, Zhejiang University, Hangzhou, China
⁵First Medical Center of Chinese People’s Liberation Army General Hospital, Beijing, China

*these authors contributed equally

Corresponding Author:
Haiyan Zhu, PhD
First Medical Center of Chinese People's Liberation Army General Hospital
28 Fuxing Road, Haidian District
Beijing, 100037
China
Phone: 86 13521361644
Email: xiaoyanzhibj301@163.com

Abstract

Background: In emergency departments (EDs), early diagnosis and timely rescue, which are supported by prediction modes using ED data, can increase patients’ chances of survival. Unfortunately, ED data usually contain missing, imbalanced, and sparse features, which makes it challenging to build early identification models for diseases.

Objective: This study aims to propose a systematic approach to deal with the problems of missing, imbalanced, and sparse features for developing sudden-death prediction models using emergency medicine (or ED) data.

Methods: We proposed a 3-step approach to deal with data quality issues: a random forest (RF) for missing values, k-means for imbalanced data, and principal component analysis (PCA) for sparse features. For continuous and discrete variables, the decision coefficient $R^2$ and the $\kappa$ coefficient were used to evaluate performance, respectively. The area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) were used to estimate the model’s performance. To further evaluate the proposed approach, we carried out a case study using an ED data set obtained from the Hainan Hospital of Chinese PLA General Hospital. A logistic regression (LR) prediction model for patient condition worsening was built.

Results: A total of 1085 patients with rescue records and 17,959 patients without rescue records were selected and significantly imbalanced. We extracted 275, 402, and 891 variables from laboratory tests, medications, and diagnosis, respectively. After data preprocessing, the median $R^2$ of the RF continuous variable interpolation was 0.623 (IQR 0.647), and the median of the $\kappa$ coefficient for discrete variable interpolation was 0.444 (IQR 0.285). The LR model constructed using the initial diagnostic data showed poor performance and variable separation, which was reflected in the abnormally high odds ratio (OR) values of the 2 variables of cardiac arrest and respiratory arrest (201568034532 and 1211118945, respectively) and an abnormal 95% CI. Using processed data, the recall of the model reached 0.746, the $F_1$-score was 0.73, and the AUROC was 0.708.

Conclusions: The proposed systematic approach is valid for building a prediction model for emergency patients.

(JMIR Med Inform 2023;11:e38590) doi:10.2196/38590

KEYWORDS
emergency medicine; prediction model; data preprocessing; imbalanced data; missing value interpolation; sparse features; clinical informatics; machine learning; medical informatics
**Introduction**

In the emergency department (ED), early identification of high-risk patients can improve clinical decisions, avoid waste of resources, and lead to better patient prognosis [1,2]. A prospective study showed that the incidence of adverse events due to improper emergency care is about 5%-10%, of which half can be prevented through early detection [3]. However, early identification is difficult as these patients often show little obvious signs before rapid deterioration [4].

Prediction models for high-risk patients in EDs can greatly support caregivers [5]. Electronic medical record (EMR) data, which fully capture patients’ status, are an important source for developing disease risk prediction models [6]. As a typical high-risk disease in EDs, sudden death is a major public health problem worldwide, accounting for 15%-20% of all deaths [7,8]. A previous study showed that cardiogenic diseases, potassium, mean platelet volume, creatinine, chloride, and sodium are important variables to predict the risk of death in patients [5]. A survey showed that age, male, hypertension, diabetes, hypercholesterolemia, and a family history of coronary heart disease are all associated with increased risk of sudden death [9]. A study evaluating the relationship between the variables of laboratory tests and the occurrence of acute death in patients found that serum sodium, glucose, and the leukocyte count show a U-shaped relationship with mortality [10]. In addition, total bilirubin, creatine kinase, the international normalized ratio, aspartate aminotransferase, and lactate dehydrogenase are all risk factors associated with acute death in patients [11-13]. However, the data quality of EMRs limits their effective use for developing prediction models [6,14]. Prediction of sudden death needs a variety of clinical data, which are frequently missing, imbalanced, and having sparse features.

Missing values, imbalanced data, and sparse features are 3 common problems of EMR data. Missing values indicate not enough data collected due to improper use of the hospital information system or other reasons [14]. Imbalanced data refer to the imbalanced distribution of negative and positive samples. This leads to more features of negative samples in the learning model, which is not suitable for the prediction of arbitrary patients [15,16]. Sparse features are zero features that are much larger than nonzero features and increase computing memory and reduce generalization ability [17,18]. Especially in small samples, a large amount of noise in sparse features makes model training impossible to converge. Therefore, tackling these quality issues of EMR data is an essential step to improve the predictive performance of machine learning (ML) models.

To solve the aforementioned 3 problems, we propose a series of ML approaches to increase fitting ability and generalization ability. Using the approach, we developed a sudden-death prediction model. The risk factors related to sudden death obtained through logistic regression (LR) model were consistent with the results reported in the earlier literature on the analysis of risk factors of in-hospital death. These results show that our data-preprocessing approach can effectively maintain the rich information contained in emergency data and provide a reliable data source for the development of a sudden-death prediction model.

**Methods**

**Study Design**

Our methods of data preprocessing consisted of 5 steps, as shown in **Figure 1**. The last 3 steps tackle 3 low-quality issues: missing values, imbalanced data, and sparse features. Finally, postprocessing data quality is evaluated by a sudden-death prediction model case study.

**Figure 1.** Workflow of ED data preprocessing and evaluation. ED: emergency department; EMR: electronic medical record.
Data Collection and Cleaning

Data for ED patient prediction model development are summarized in Table 1.

Close investigation of each data table is required so as to know the location of our content of interest. For instance, data regarding a patient’s basic information are stored in the emg_visit table. Lab test items and results are stored in the lab_result and lab_master tables. The clinical record field in the emg_order table can be used to determine whether a sudden-death event occurred. One lab test (eg, blood test) can be performed multiple times to observe the patient status closely. Based on clinical experts’ opinions, only the last one is meaningful.

| Table 1. Description of the data table involved in the query process. |
|------------------|------------------------------------------------|
| Table name       | Data description                                           |
| emg_drug_detail  | The patient's medication record, including the prescription number, drug name, dosage, drug specification, administration time, and administration route during the treatment period |
| emg_drug_master  | Master record form of patient medication recording patient ID and prescription number |
| emg_order        | Doctor’s order record form used to record the medication, inspection, diagnosis, treatment, and other doctor’s orders of the patient during treatment |
| emg_visit        | Patient visit information table, including the patient's basic personal information, diagnosis of the current visit, triage, and other information |
| lab_test_master  | Patient’s laboratory test master record form recording the patient’s age and gender information, laboratory test items made during the visit, and the corresponding doctor’s order ID |
| lab_result       | Laboratory test results of patients, including test results of patients |

Variable Screening

The number of variables obtained from the data collection was large, so screening of important variables facilitated final analysis. Two approaches can be adopted. One is based on statistical significance. The other is based on the specific research objective, opinions of medical experts, or authoritative literature [5,12,13]. In our study, the first approach was taken. Variables with many missing values were filtered out using the threshold. For example, Alvarez et al [19] set the threshold to 2%, while Seki et al [20] set it to 25%. In this study, we set the threshold to 80%. This means that when 80% of the values of 1 variable are missing, that variable should be filtered out.

Data Interpolation for Missing Values

Missing values affect the effectiveness of ML models. Data missing show 3 different patterns: missing completely at random (MCAR), missing at random deletion (MAR), and not missing at random (MNAR). MCAR means that the missing of data is completely random and does not depend on observed or unobserved values [21]. In this case, any interpolation method will not cause deviation. However, the assumption of MCAR in actual data is difficult to satisfy [22,23]. MNAR and MAR mean that the missing of data depends on the unobserved value and does not depend on the unobserved value, respectively [24]. However, it is impossible to infer whether the missing pattern belongs to MNAR or MAR through the existing data containing the missing pattern, and the assumption based on MAR is more consistent with the actual data situation [22,25]. MAR allows us to estimate missing values using existing observation data in the data set [24].

The goal of all kinds of interpolation methods is to reasonably estimate missing values and improve the quality of data. Interpolation methods are mainly divided into single interpolation and multiple interpolation. Multiple interpolation is a commonly used and better performance interpolation method. It generates multiple possible estimates for missing data and uses statistical inference to interpolate the final value. This method can reflect the randomness of missing data, and the interpolation error is smaller [21]. In a single interpolation, interpolation methods, such as constants (ie, specific identifications), mean, median, and data distribution, can be used. However, such methods usually cause greater deviation [26,27]. The single interpolation method based on ML has attracted increasingly more attention [23], such as interpolation based on a clustering algorithm [28], an ensemble model [29], and Bayesian theory [30]. Although multiple imputation can bring smaller deviation, when the frequency of missing data is high and the sample size is small, multiple imputation should be considered [31]. However, its implementation is relatively complex, and it needs to involve the selection of an interpolation model and the number of interpolation data created [32]. When the data are sufficient and the variability of the estimated value does not need to be considered, it is feasible to choose multiple imputation or single imputation [31]. Considering that our sample size was relatively sufficient, to build a simpler interpolation method, we used a random forest (RF) [33,34] as the interpolation algorithm to realize the interpolation of missing data in the form of a single interpolation.

Altogether, the followed steps are proposed.

- For variable “i,” 1 set of patient samples without missing values work as training samples and the other set of patient samples with missing values work as test samples.
- If other variables in the 2 samples are missing, the mean (continuous variable) or mode (discrete variable) is temporarily interpolated to form a complete sample.
- Use training samples to train RF models, the model is applied to test samples to predict missing values.
- For the next variable, steps 1, 2, and 3 are repeated until all variables of the whole sample are interpolated.

Processing Imbalanced Data

Imbalanced data refer to the imbalanced distribution of negative and positive samples. For example, in the classification of rare diseases and credit predictions, there could be more negative samples than positive ones. Because most ML algorithms assume that categories (eg, positive or negative) of samples are evenly distributed, classifying models trained with imbalanced data are more likely to classify a new sample into the majority category [15].

Basic solutions for imbalanced data are to use under- or oversampling to make the data balanced, such as random oversampling [35], random undersampling, the synthetic minority oversampling technique (SMOTE) [36], and the adaptive synthetic sampling method (ADASYN) [15]. Although both undersampling and oversampling approaches can achieve data balance, the oversampling approach adds many sample copies to overfit the model. Wang and Japkowicz [16] and Chawla et al [36] also argued that undersampling is more favorable than oversampling in extreme imbalance situations. However, randomly discarding undersampling may also lose some representative samples. Segura-Bedmar et al [37] and Lin et al [38] proposed a clustering method to tackle this problem. The k-means considers the similarity between samples and uses the sample closest to the centroid of the cluster to approximate all the sample characteristics within the cluster, and the obtained samples are representative. The advantage of the clustering method over random undersampling is that all samples are used in the clustering process. This ensures that the information about all samples can be used to determine the sampling results and some important samples are not randomly discarded. In addition, we can adjust the number of clusters in k-means according to the actual data imbalance so as to achieve different undersampling ratios without other complex adjustments.

To avoid the loss of important samples, we adopted k-means based on the Euclidean distance to cluster samples. New samples were generated though clustering, which had similar characteristics in the same cluster and were distinguished in the different clusters. The centroid of a cluster represents the overall characteristics of the whole cluster. In this way, important features are not discarded. Since the centroid of the cluster is calculated based on the average of the samples in the cluster, the centroid is not necessarily a real sample. So, we took the real samples with the smallest distance from the centroid.

Processing Sparse Features

Sparse features means that the feature index is much larger than the actual number of nonzero features. In total, there were 891 different types of diagnosis in our data set. However, for a single patient, the number of diagnoses was quite few. This formed sparse-feature phenomena.

When sparse features occur, the sample is prone to having the problem of variable separation and multicollinearity. That is, a single variable or a linear combination of multiple variables can perfectly predict outcome events. However, this only works for small-size samples. It also leads to the situation in which the model gives an abnormally large weight to the variables and the results are unreliable [17,18,39]. Although there are many methods to optimize weights, such as gradient descent, a large number of zeros in features make the gradient tend to 0, and the parameters cannot be fully trained.

The processing of sparse features can be considered from both the model and the data themselves. From the point of view of the model, the parameter estimation bias of high-dimensional sparse data can be reduced through the optimization of the algorithm. For example, Firth regression [40] is used. The basic idea is to add a penalty term to the score function so as to reduce the deviation of the maximum-likelihood estimate of the parameter. This can solve the problem of variable separation and multicollinearity caused by sparse features to a large extent.

From the point of view of the data themselves, it is necessary to transform the data to be processed into nonsparse data, and this transformation should retain the amount of information contained in the original data as much as possible. Considering the theme of our paper, our goal is to improve the quality of data rather than optimize the model algorithm. Therefore, we solved the problem of sparse features from the perspective of data. At present, there are many dimensionality reduction methods for high-dimensional sparse features, such as principal component analysis (PCA) [39], singular value decomposition (SVD) [41], and linear discriminant analysis (LDA) [42]. The essence of these methods is to map the original data to a low-dimensional space through a specific transformation form to solve the problem of data sparsity. Among these methods, LDA needs to reduce dimensionality based on sample labels. Considering that the actual data may not be able to carry labels, and the difference in label definitions will greatly affect the dimensionality reduction results, this supervised dimensionality reduction method is not conducive to being extended to other data scenarios [43]. Therefore, we considered using unsupervised dimensionality reduction methods, such as PCA, to transform our data.

PCA has been widely used in analysis with high-dimensional sparse features [44-46]. PCA essentially transforms the feature space of the original sample so that the new feature is a linear combination of the original features. The basic principle of principal component (PC) selection is to keep the maximum variance, and all PCs are orthogonal to one another. Thus, the phenomenon of multicollinearity is avoided. Therefore, new samples no longer have sparse features, which makes the ML model better fit the parameters.

In detail, new data can replace the original data as the input source for regression or classification models. Suppose where each column represents a feature and each row is a sample. Assuming that the sample has been decentralized, represents the covariance of matrix X. Let the transformed matrix be D, which is derived as:

As C is a real symmetric matrix, according to the properties of the real symmetric matrix, its order m must have m unit orthogonal eigenvectors. That is, is a matrix that can make the original covariance matrix similar to diagonalization.
Therefore, by solving \( m \) eigenvalues and eigenvectors of \( \mathbf{A} \). By sorting the eigenvalues from large to small, we got \( \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m) \). There are the following relationships:

Take the first \( k \) columns of \( \mathbf{V} \) as the basis for transforming \( m \)-dimensional features into \( k \)-dimensional features and record it as \( \mathbf{Y} = \mathbf{X}\mathbf{P} \).

First, we manually merged similar diagnostic nouns according to prior knowledge, from 891 to 405. However, the data were obviously separated and sparse. For instance, none of the negative samples had a sudden cardiac arrest or sudden respiratory arrest diagnosis. Next, we only kept the diagnosis that appeared in more than 5% population. Finally, PCA was proposed for the remaining variables. The first 17 PCs that could explain 98.2% variance of the original sample were selected. Regression analysis was carried out on the samples after dimensionality reduction. The explanation of variables was achieved by counting the weight of the original variables on each PC.

**Ethical Considerations**

After preliminary review, the project was found to be in line with relevant medical ethics requirements. If it is funded by the Hainan Major Science and Technology Program in 2020, the Hainan Medical Ethics Committee will perform its duties and strictly abide by relevant regulations and requirements for medical ethics and informed consent of patients to ensure ethical supervision and review during the implementation of the project (reference number: 00824482406).

**Results**

**Data Preprocessing and Model Building**

A comprehensive evaluation was carried out on the ED data set of the Hainan Hospital of Chinese PLA General Hospital. We developed a set of Python programs to implement our methods. Specifically, the program was developed in Microsoft Windows 10 (Intel (R) core (TM) i5-9500 CPU, 3GHz). All data preprocessing and model building were completed in Python (Python 3.8 Anaconda) using multiple Python data science libraries, mainly including Numpy, Pandas, Matplotlib, and Scikit-learn. In addition, codes on data interpolation, imbalance correction, and PC regression are currently available on GitHub [47].

**Data Collection and Cleaning**

We collected the data of patients who went to the ED of the Hainan Hospital of Chinese PLA General Hospital from July 27, 2017, to May 6, 2021. In the sudden-death group, the data of 1085 patients were collected. In the non-sudden-death group, the data of 17,959 patients were collected. For the analysis of laboratory test data, we excluded patients who did not have any laboratory test records before sudden death. A total of 108 (10%) patients were excluded, and 977 (90%) patients with sudden death were used for the analysis of laboratory test data. For diagnostic data, we excluded patients who were missing diagnostic data from the visit. Finally, there were 1083 patients with sudden death and 615 patients with nonsudden death. We developed statistics on the baseline data of all patients, as shown in Supplementary Table S1 in Multimedia Appendix 1. Distributions of age and gender are visualized in Figures 2-5.

In the first group, there were 741 males (68.4%) and 342 females (31.6%), and 2 (0.2%) patients lacked gender information (Figure 2). The age varied between 45 and 80 years. The mean age was 56.4 years (SD 11.2). The quartile, median, and mode were 44, 59, and 68, respectively. In the second group, there were 9403 (52.4%) males and 8556 (47.6%) females. The age distribution is shown in Figures 4 and 5. The mean age was 41.6 years (SD 13.6). The quartile, median, and mode were 29, 42, and 48, respectively. For both groups, their distributions of age were akin to the normal distribution, which is consistent with a real-life situation.

**Figure 2.** Distribution of the gender of patients with sudden death.
Figure 3. Distribution of the gender of patients without sudden death.

Figure 4. Distribution of age of patients with sudden death.

Figure 5. Distribution of patients of age with nonsudden death.
Variable Screening

To perform variable screening, that is, filtering out insignificant variables, we counted the number of appearance and missing times. The second row of Table 2 shows the number of patients who had no corresponding data in the individual category. Moreover, we investigated the reasons missing data exist in all the 3 categories. For instance, there were 108 (10%) patients having no laboratory test. Among them, we could not find lab test data for 33 (30.6%) patients. For the remaining 75 (69.4%) patients, their lab tests appeared after the sudden-death event. There were 287 (26.4%) patients having no medication data. Sudden death had occurred before the medication was given, and the medication was in the doctor’s order record, such as an epinephrine injection, but was not recorded in the patient’s medication table.

There were 275 variables in the lab test category. For a given variable, not every patient (sample) had the value, namely a missing value. The missing ratio of a variable could be obtained by the number of cases having a missing value of that variable being divided by the total number of patients. The average ratio was 79.8%, as shown in the third row of Table 2. So, we set an 80% threshold to screen nonstatistically significant variables. Finally, 72 variables were kept in this category. These were patient age, gender, glucose, creatine kinase, inorganic phosphorus, total cholesterol, triglycerides, potassium, sodium, and calcium.

For diagnosis, 891 different types of diagnosis were obtained after the initial data collection. Because the diagnosis is recorded in the form of free text, 1 diagnosis item could have several different synonyms. By merging these texts into a unified name via manual review, we obtained 405 variables. The number of confirmed patients of each diagnostic variable was counted.

Table 2. Missing value ratios of variables of patients with sudden death.

<table>
<thead>
<tr>
<th></th>
<th>Laboratory tests (275 variables)</th>
<th>Medications (402 variables)</th>
<th>Diagnosis (891 variables)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients without data, n (%)</td>
<td>108 (10%)</td>
<td>287 (26.4%)</td>
<td>2 (0.18%)</td>
</tr>
<tr>
<td>Average ratio of missing values</td>
<td>79.8% (866/1085)</td>
<td>72.4% (786/1085)</td>
<td>99% (1080/1085)</td>
</tr>
<tr>
<td>Maximum ratio of missing values</td>
<td>90% (977/1085)</td>
<td>73.5% (797/1085)</td>
<td>100% (1085/1085)</td>
</tr>
<tr>
<td>Minimum ratio of missing values</td>
<td>25.8% (280/1085)</td>
<td>48.5% (526/1085)</td>
<td>58.4% (634/1085)</td>
</tr>
</tbody>
</table>

Data Interpolation, Processing Imbalanced Data, and Sparse Features

In addition to age and gender, we used an RF to interpolate the missing values for each of the remaining 3 nonmissing patient data were used as a training set to train the model to interpolate missing values. The training set was further split into training data (80%) and validation data (20%). The coefficient of determination $R^2$ and the $\kappa$ coefficient were used to test the consistency of the imputation results of continuous variables and categorical variables. In the interpolation process, the median of $R^2$ was 0.623 (IQR 0.647) and the median of the $\kappa$ coefficient was 0.444 (IQR 0.285).

Due to the extreme imbalance of our original data, the number of patients with sudden death only accounted for 5% (977/18,936) of the total sample size. We generated 4 different data ratios (1:10, 1:5, 1:2, and 1:1) through k-means to achieve undersampling. These data were used with the original ratio to test the consistency of the imputation results of continuous variables, nonmissing values for each of the remaining variables. Nonmissing patient data were used as a training set to train the model to test variables using a data set with a data ratio of 1:1 as the data source to filter variables. To reflect the degree of correlation between variables, continuous variables were treated as ordinal categorical variables. Taking the normal index range of the variables as a reference point, the test results of the patients were mapped into 3 categories: L (index is lower than the normal value), N (index is normal), and H (index is higher than the normal value). To determine the significant factors affecting the sudden death of patients and avoid a negative effect on the final analysis results, we first performed the chi-square test to filter out the variables and then excluded variables when $P > .10$. Next, LR univariate analysis was performed to filter out variables with $P > .05$. Tables 3 and 4, respectively, show the variables excluded by the chi-square test and the LR univariate analysis, and their $P$ values. We reintroduced some of the excluded variables into the final candidate variable set according to the literature review and the advice of consulting medical experts, including urine specific gravity, chloride, hematocrit, sodium, magnesium, lactate dehydrogenase, urine ketone body test, red blood cell count, and serum albumin. These variables have no significant statistical significance but are clinically related to sudden death. Finally, we selected 4 subgroups from the set of variables with significant statistical significance. In addition, variables not statistically significant but related to outcome events were also grouped separately. The final 5 groups were subjected to LR multivariate analysis, and the groups were as follows:

- Group 1: qualitative test of creatinine, serum uric acid, urine protein
- Group 2: $\gamma$-glutamyl transferase, alanine aminotransferase, total bilirubin

Validation by a Sudden-Death Case Study

Analyzing Risk Factors of Sudden Death

We constructed an LR model to analyze the patients’ laboratory test variables using a data set with a data ratio of 1:1 as the data source to filter variables. To reflect the degree of correlation between variables, continuous variables were treated as ordinal
For each group, 500-fold bootstrapping was used for model training and evaluation [48]. Each bootstrap randomly split 70% of the data into the training set and 30% of the data into the test set. Finally, the mean values of AUROC, recall, and $F_1$-score for 500 training sessions in each group were reported, and the AUROC also reported the 95% CI. Table 5 illustrates the model evaluation results of the 5 groups of variables. The performance parameters of group 2 were the best among the 5 groups of variables. In the recognition of patients with sudden death, a recall rate of 0.801 was obtained, the $F_1$-score was 0.835, and the model’s AUROC was 0.843 (95% CI 0.842-0.844). The results showed that this set of variables can better identify patients with sudden death. Therefore, other group variables based on the group 2 variables were added successively, and AUROC was taken as the evaluation index. The added variables would be included in the final model if AUROC could be improved. In the end, 13 laboratory test risk variables related to sudden death events were determined, and the patient’s gender variable was retained as a demographic feature. In general, the final variables used included $\gamma$-glutamyl transferase, alanine aminotransferase, total bilirubin, creatinine, serum uric acid, the international standardized ratio, creatine kinase, the platelet count, potassium, sex, sodium, magnesium, chloride, and serum albumin. These variables were used to build the final LR model. Table 6 shows the results of LR multivariate analysis.

After determining the patient features for analysis, we split the original scale data into a training set (70%) and a test set (30%). For the training set, 4 different categories of data sets (1:1, 1:2, 1:5, 1:10) were formed by undersampling to train the model. Finally, the performance of the model was evaluated on the test set. The mean and 95% CI (500-fold bootstrapping) of the final AUROC, AUPRC, $F_1$-score, and recall are shown in Supplementary Table S2 in Multimedia Appendix 1. In addition, we further used Brier scores to evaluate the calibration ability of models trained with different data ratios.

In general, as the data ratio tended to balance, the performance of the model gradually improved. Figures 6 and 7 show the model receiver operating characteristic (ROC) curve (Figure 6) and the precision-recall (PR) curve (Figure 7) of the 4 data ratios. In recognizing patients with sudden death, the best model obtained a recall rate of 0.863 (95% CI 0.862-0.865), the $F_1$-score was 0.84 (95% CI 0.839-0.842), the AUROC of the model was 0.895 (95% CI 0.894-0.896), and the AUPRC was 0.897 (95% CI 0.896-0.899). The original scale data model performed the worst, with an AUROC of 0.812 (95% CI 0.811-0.813) and an AUPRC of 0.407 (95% CI 0.404-0.409). We plotted the reliability curves of 5 training sets with different data ratios on the same test set and calculated Brier scores (Supplementary Figure S1 in Multimedia Appendix 1). Consistent with the viewpoint mentioned by Geeven et al [49], imbalance correction actually weakened the clinical application value of the model, which was mainly manifested in the poor calibration ability of the model. With the increase in sampling, the calibration of the model was worse and the Brier score was 0.16 and 0.108 in the data ratio of 1:1 and the original data ratio, respectively. Imbalance correction can balance the sensitivity and specificity of the model to a greater extent and avoid biased errors in the model. Undersampling optimizes the AUROC, $F_1$-score, and AUPRC of the model trained by the proportion of the original data. Although the Brier score in calibration improved, the gap was not large. To observe the risk factors of sudden death in patients more intuitively, we visualized the regression coefficients of the best model after performing LR(Figure 8) to observe the relationship between variables and sudden-death events.

### Table 3. Statistics of variables filtered by the chi-square test.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\chi^2$ (df)</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monocytes</td>
<td>5.433 (6)</td>
<td>.49</td>
</tr>
<tr>
<td>Basophil</td>
<td>0.705 (4)</td>
<td>.95</td>
</tr>
<tr>
<td>Eosinophils</td>
<td>0.977 (4)</td>
<td>.91</td>
</tr>
<tr>
<td>Urine specific gravity determination</td>
<td>0 (2)</td>
<td>.99</td>
</tr>
<tr>
<td>Urine tube type</td>
<td>1.25 (4)</td>
<td>.87</td>
</tr>
<tr>
<td>Urine tube type (microscopic examination)</td>
<td>6.863 (8)</td>
<td>.98</td>
</tr>
<tr>
<td>Qualitative test of urinary bilirubin</td>
<td>13.185 (4)</td>
<td>.21</td>
</tr>
<tr>
<td>Mean erythrocyte hemoglobin concentration</td>
<td>7.828 (6)</td>
<td>.25</td>
</tr>
<tr>
<td>Chloride</td>
<td>4.649 (6)</td>
<td>.59</td>
</tr>
<tr>
<td>Erythrocyte volume distribution width</td>
<td>1.148 (4)</td>
<td>.89</td>
</tr>
<tr>
<td>Hematocrit assay</td>
<td>4.982 (6)</td>
<td>.55</td>
</tr>
<tr>
<td>Sodium</td>
<td>7.915 (6)</td>
<td>.24</td>
</tr>
<tr>
<td>Magnesium</td>
<td>10.22 (6)</td>
<td>.12</td>
</tr>
</tbody>
</table>
Table 4. Statistics of variables screened by LR\textsuperscript{a} univariate analysis.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Reference range</th>
<th>OR\textsuperscript{b} (95% CI)</th>
<th>( P ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lactate dehydrogenase</td>
<td>50.0-150.0 U/L</td>
<td>1.029 (0.94-1.127)</td>
<td>.53</td>
</tr>
<tr>
<td>Urine ketone body test</td>
<td>N/A\textsuperscript{c}</td>
<td>0.912 (0.769-1.081)</td>
<td>.29</td>
</tr>
<tr>
<td>Red blood cell count</td>
<td>3.5-5.9 1012/L</td>
<td>0.827 (0.642-1.065)</td>
<td>.14</td>
</tr>
<tr>
<td>Serum albumin</td>
<td>35.0-50.0 g/L</td>
<td>0.893 (0.689-1.157)</td>
<td>.39</td>
</tr>
<tr>
<td>High-density lipoprotein cholesterol</td>
<td>1.0-1.6 mmol/L</td>
<td>0.961 (0.749-1.232)</td>
<td>.75</td>
</tr>
</tbody>
</table>

\textsuperscript{a}LR: logistic regression.  
\textsuperscript{b}OR: odds ratio.  
\textsuperscript{c}N/A: not applicable.

Table 5. Comparing the performance of 5 groups of variables.

<table>
<thead>
<tr>
<th>Group</th>
<th>Recall</th>
<th>( F_1 )-score</th>
<th>AUROC\textsuperscript{a} (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.478</td>
<td>0.6</td>
<td>0.683 (0.681-0.684)</td>
</tr>
<tr>
<td>2</td>
<td>0.801</td>
<td>0.835</td>
<td>0.843 (0.842-0.844)</td>
</tr>
<tr>
<td>3</td>
<td>0.606</td>
<td>0.687</td>
<td>0.725 (0.724-0.727)</td>
</tr>
<tr>
<td>4</td>
<td>0.484</td>
<td>0.605</td>
<td>0.686 (0.685-0.687)</td>
</tr>
<tr>
<td>5</td>
<td>0.852</td>
<td>0.651</td>
<td>0.562 (0.561-0.564)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}AUROC: area under the receiver operating characteristic curve.

Table 6. LR\textsuperscript{a} multivariate analysis.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Reference range</th>
<th>OR\textsuperscript{b} (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \gamma )-Glutamyl transferase</td>
<td>0.0-50.0 U/L</td>
<td>0.225 (0.222-0.228)</td>
</tr>
<tr>
<td>Alanine aminotransferase</td>
<td>5.0-40.0 U/L</td>
<td>1.828 (1.804-1.852)</td>
</tr>
<tr>
<td>Total bilirubin</td>
<td>0.0-21.0 ( \mu )mol/L</td>
<td>19.954 (19.7-20.2)</td>
</tr>
<tr>
<td>Creatinine</td>
<td>30.0-110.0 ( \mu )mol/L</td>
<td>1.352 (1.331-1.372)</td>
</tr>
<tr>
<td>Serum uric acid</td>
<td>104.0-444.0 ( \mu )mol/L</td>
<td>1.346 (1.334-1.359)</td>
</tr>
<tr>
<td>International normalized ratio</td>
<td>0.8-1.2</td>
<td>2.23 (2.188-2.272)</td>
</tr>
<tr>
<td>Creatine kinase</td>
<td>24.0-320.0 U/L</td>
<td>2.457 (2.431-2.483)</td>
</tr>
<tr>
<td>Platelet count</td>
<td>100.0-300.0 ( \times 10^9 )/L</td>
<td>0.623 (0.617-0.629)</td>
</tr>
<tr>
<td>Potassium</td>
<td>3.5-5.1 ( \mu )mol/L</td>
<td>1.057 (1.043-1.07)</td>
</tr>
<tr>
<td>Gender</td>
<td>Female</td>
<td>0.183 (0.182-0.184)</td>
</tr>
<tr>
<td>Sodium</td>
<td>135-145 mmol/L</td>
<td>2.182 (2.102-2.262)</td>
</tr>
<tr>
<td>Magnesium</td>
<td>0.8-1.0 mmol/L</td>
<td>4.807 (4.587-5.027)</td>
</tr>
<tr>
<td>Chloride</td>
<td>96.00-106.00 mmol/L</td>
<td>0.615 (0.603-0.627)</td>
</tr>
<tr>
<td>Serum albumin</td>
<td>35-51 g/L</td>
<td>1.284 (1.268-1.3)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}LR: logistic regression.  
\textsuperscript{b}OR: odds ratio.
Figure 6. ROC curves of different data ratio. AUC: area under the curve; ROC: receiver operating characteristic.

Figure 7. PR curves of different data ratio. AUPRC: area under the precision-recall curve; PR: precision-recall.
Development of Other ML Models

We use interpolated and undersampled data (data ratio 1:1) to train several other ML models and evaluate their performance. The training models included an RF [50], a gradient boosting machine (GBM) [51], a support vector machine (SVM) [52], and least absolute shrinkage and selection operator (LASSO) [53], which are also often used to develop medical prediction models [49,54]. We use 500-fold bootstrapping for internal validation. Each bootstrap used 70% data for training and the remaining 30% data for performance evaluation. The area under the curve (AUC), AUPRC, recall, and \( F_1 \)-score and their 95% CI values were reported. Before model training, a grid search was conducted to tune the best hyperparameter of each model through 5-fold cross-validation. The hyperparameter settings of each model are shown in Supplementary Table S7 in Multimedia Appendix 1. The ROC curve and PR curve of the models are shown in Supplementary Figures S2 and S3 in Multimedia Appendix 1, respectively, and the performance evaluation results are shown in Supplementary Table S8 in Multimedia Appendix 1. In general, the performance of the RF and GBM with an integrated scheme was the best, with an AUC of 0.936 (95% CI 0.934-0.937) and 0.931 (95% CI 0.93-0.932), respectively, and an \( F_1 \)-score of 0.857 (95% CI 0.856-0.858) and 0.80 (95% CI 0.78-0.80), respectively. This can benefit from the generalization and the ability to deal with complex feature relationships of the integrated model. The comprehensive decision results of multiple base learners are more stable than the single-model prediction results, and the performance is better. The SVM also performed better than the LR and LASSO, which are linear models, with an AUC of 0.913 (95% CI 0.912-0.914). This shows that there are some nonlinear features we used that made the linear model insufficient to recognize the relationship between these features.

Diagnostic Data Analysis Results

The final sample included 1083 patients with sudden death and 615 patients with nonsudden death. Table 7 shows the number of confirmed patients with 18 variables. The final diagnostic variables used included hypertension, myocardial infarction, cerebral hemorrhage, cardiac arrest, absolute pain, atmospheric fabric, fever, trauma, respiratory arrest, diabetes, corporate heart disease, and cerebral infarction.

We used 500-fold bootstrapping for internal validation of the model. For each bootstrap, 70% of the samples were randomly selected as the training set and 30% as the test set to evaluate the model. The final reported model performance was the mean and 95% CI of 500 results [48].

The first 17 PCs that could explain 98.2% of the variance of the original sample were selected as new variables for analysis. To observe the role of PCA, we compared the 2 schemes: the LR model using the original data and the LR model after dimensionality reduction using PCA. The LR model trained with the original data obtained a recall rate of 0.445 (95% CI 0.443-0.448), an \( F_1 \)-score of 0.562 (95% CI 0.56-0.564), and an AUROC of 0.602 (95% CI 0.6-0.603). After PCA dimensionality reduction of the original data, the PC variable was used as the data source to train the LR model, and a recall rate of 0.746 (95% CI 0.731-0.76) was obtained, the \( F_1 \)-score was 0.73 (95% CI 0.721-0.738), and the AUROC of the model was 0.708 (95% CI 0.707-0.71). Figure 9 shows the ROC curves of the 2 models. The LR model using the original data had the phenomenon of variable separation, which is reflected in the abnormally high OR values of cardiac arrest and respiratory arrest (201568034532 and 1211118945) and an abnormal 95% CI, which makes the results unreliable. In addition, the performance of the model was poor, and only a recall rate of 0.445 was obtained in the identification of patients with sudden death, which means that the identification ability of the model for patients with sudden death is not strong. After PCA dimensionality reduction, the data were no longer sparse, the model parameters were better fitted, and the model performance improved to a certain extent. In addition, data conversion also eliminated the problems of variable separation and multicollinearity.
To determine the impact of various diagnostic variables on the sudden death of emergency patients, we statistically analyzed the results of multivariate analysis on 17 PCs input into the LR model. The OR of PC4, PC5, and PC6 was 3.044, 2.859, and 3.931, respectively, showing a significant correlation with sudden-death events (Table 8). In each PC, the magnitude of the loading, the elements in the PC, reflected the importance of the original variable in the PC (Supplementary Table S3 in Multimedia Appendix 1). The loadings of all components showed that cerebral infarction, hypertension, and pulmonary infection were the top 3 variables in PC4. In PC5 and PC6, the top 3 variables were consciousness disorder, diabetes, and fever. Based on the results of the 3 PCs, we believe that the 6 diagnoses of cerebral infarction, hypertension, pulmonary infection, consciousness disorder, diabetes, and fever are significantly associated with sudden death in emergency patients.

### Table 7. Statistics of people diagnosed.

<table>
<thead>
<tr>
<th>Variable</th>
<th>People with sudden death diagnosed, n (%)</th>
<th>People with nonsudden death diagnosed, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Myocardial infarction</td>
<td>57 (5.26)/23 (3.74)</td>
<td></td>
</tr>
<tr>
<td>Chest tightness</td>
<td>8 (0.74)/35 (5.69)</td>
<td></td>
</tr>
<tr>
<td>Cardiac arrest</td>
<td>120 (11.08)/0</td>
<td></td>
</tr>
<tr>
<td>Fever</td>
<td>50 (4.62)/43 (6.99)</td>
<td></td>
</tr>
<tr>
<td>Rib fracture</td>
<td>58 (5.36)/3 (0.49)</td>
<td></td>
</tr>
<tr>
<td>Abnormal renal function</td>
<td>42 (3.88)/35 (5.69)</td>
<td></td>
</tr>
<tr>
<td>Chest pain</td>
<td>18 (1.66)/38 (6.18)</td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>65 (6.00)/66 (10.73)</td>
<td></td>
</tr>
<tr>
<td>Abdominal pain</td>
<td>30 (2.77)/45 (7.32)</td>
<td></td>
</tr>
<tr>
<td>Pulmonary infection</td>
<td>85 (7.85)/64 (10.41)</td>
<td></td>
</tr>
<tr>
<td>Respiratory arrest</td>
<td>106 (9.79)/0</td>
<td></td>
</tr>
<tr>
<td>Trauma</td>
<td>58 (5.36)/16 (2.60)</td>
<td></td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>39 (3.60)/33 (5.37)</td>
<td></td>
</tr>
<tr>
<td>Disturbance of consciousness</td>
<td>82 (7.57)/17 (2.76)</td>
<td></td>
</tr>
<tr>
<td>Cerebral hemorrhage</td>
<td>77 (7.11)/26 (4.23)</td>
<td></td>
</tr>
<tr>
<td>Cerebral infarction</td>
<td>75 (6.93)/71 (11.54)</td>
<td></td>
</tr>
<tr>
<td>Coronary heart disease</td>
<td>29 (2.68)/39 (6.34)</td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>65 (6.00)/106 (17.24)</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 9.** ROC curves of 2 models. AUC: area under the curve; LR: logistic regression; PCA: principal component analysis; ROC: receiver operating characteristic.
Table 8. PC regression results

<table>
<thead>
<tr>
<th>PC</th>
<th>OR^b (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.239 (0.235-0.242)</td>
</tr>
<tr>
<td>2</td>
<td>2.429 (2.383-2.476)</td>
</tr>
<tr>
<td>3</td>
<td>1.19 (1.126-1.253)</td>
</tr>
<tr>
<td>4</td>
<td>3.044 (2.948-3.141)</td>
</tr>
<tr>
<td>5</td>
<td>2.859 (2.687-3.031)</td>
</tr>
<tr>
<td>6</td>
<td>3.931 (3.714-4.148)</td>
</tr>
<tr>
<td>7</td>
<td>1.49 (1.405-1.575)</td>
</tr>
<tr>
<td>8</td>
<td>1.699 (1.562-1.836)</td>
</tr>
<tr>
<td>9</td>
<td>2.104 (1.949-2.259)</td>
</tr>
<tr>
<td>10</td>
<td>2.153 (2.016-2.289)</td>
</tr>
<tr>
<td>11</td>
<td>2.451 (2.191-2.711)</td>
</tr>
<tr>
<td>12</td>
<td>2.031 (1.855-2.206)</td>
</tr>
<tr>
<td>13</td>
<td>1.457 (1.339-1.575)</td>
</tr>
<tr>
<td>14</td>
<td>0.949 (0.863-1.034)</td>
</tr>
<tr>
<td>15</td>
<td>1.423 (1.231-1.614)</td>
</tr>
<tr>
<td>16</td>
<td>2.546 (2.221-2.871)</td>
</tr>
<tr>
<td>17</td>
<td>0.182 (0.164-0.201)</td>
</tr>
</tbody>
</table>

^aPC: principal component.
^bOR: odds ratio.

Discussion

Principal Findings

In this paper, 3 ML schemes were proposed to deal with missing, imbalanced, and sparse features in the process of developing sudden-death prediction models using emergency medicine data, which improves the performance of the developed model. To solve the problem of missing data, we propose an RF method to use real data to interpolate missing data. In the interpolation process, the consistency of the interpolation results is checked by determining the coefficient $R^2$ and the $\kappa$ coefficient. From the interpolation results, the method shows the ability to correctly interpolate missing data. Imbalanced data are not conducive to obtaining accurate analysis results, and the model will be more inclined to predict new samples as patients with nonsudden death [15]. In view of this phenomenon, we used the k-means algorithm to generate multiple data sets with different proportions of different categories by undersampling to evaluate the model. The method based on k-means can better preserve the patient’s characteristic information. This method will not lose some representative patient samples due to random discarding, thus reducing the bias caused by sampling. The results show that the comprehensive performance of the model gradually improves as the data tend to balance (Figures 3-5). However, imbalance correction will weaken the calibration ability of the model and increase the calibration error. Data sparsity is also not conducive to modeling and analysis. When the samples are too sparse, the results of the classifier based on maximum-likelihood estimation will become unreliable, because there may be variable separation and multicollinearity [18,55].

PC regression analysis is a method that uses PCA to extract the PC information about the original samples and uses PCs to replace the original variables for regression modeling [39]. In our diagnostic data, the LR model using the original data showed the phenomenon of variable separation, which led to unreliable results and poor performance. The performance of the PC regression model has been improved. In addition, we can analyze the diagnosis significantly related to the sudden death of emergency patients from the results of PC regression. These diagnoses are consistent with previous findings [9].

At present, there are many studies on the prediction of sudden death. Yu et al [54] constructed an ML model to predict sudden cardiac death (SCD) in 15,661 patients with atherosclerosis. The results showed that the ML model performs better than the standard Poisson regression model and the AUROC of the ML model was 0.89. Karen et al [56] trained an ML-based early warning model for identifying sudden infant death syndrome using the public data set “Lipidomic in sudden infant death syndrome.” The RF algorithm achieved an AUROC of 0.9 and a recall of 0.8. Ye et al [5] selected a variety of ML algorithms to build an early real-time early warning system (EWS) to predict the death risk of emergency patients and carried out prospective validation. The results showed that the EWS could give an early warning within 40 hours before sudden death, and the AUROC reached 0.884. Bhattacharya et al [57] used the electronic health records of 711 patients with hypertrophic myocardial cake and established an LR and naive Bayesian model with 22 variables, including statins, a family history of...
SCD, and left ventricular ejection fraction, to predict the risk of sudden death (ventricular fibrillation) in these patients. The sensitivity and specificity of the optimal model were 0.73 and 0.76, respectively, and the AUROC was 0.83. For our model, in the LR model constructed by using laboratory test data, the AUROC reached 0.895. After imbalance correction, the recall rate and AUPRC improved, reaching 0.863 and 0.897, respectively. Compared to the existing sudden-death prediction model based on ML, the performance of our model can achieve a similar effect, further indicating that our data-preprocessing methods can preserve the patient's characteristic information and improve the availability of emergency care.

Limitations
This work also has some limitations. On the one hand, we only considered a single ML algorithm for data interpolation and did not discuss and compare the application of other possible ML algorithms in interpolation. It is possible that we overlooked the better performance of other methods. For example, for our data, due to the large proportion of missing and seriously imbalanced categorical variables, although we tried to adjust the relatively balanced data set to train the model, the \( \kappa \) coefficient improved to a certain extent but the effect was still poor. Therefore, a further discussion of ML methods that can handle a large number of missing and unbalanced categories or more reasonable feature processing may achieve better imputation results. Although imbalance correction can improve the sensitivity and specificity of the model, it can avoid biased errors of the model. However, this correction will also weaken the clinical application value of the model, lowering the calibration ability of the model and making it unable to accurately estimate the risk probability of patients. For the prediction model, the calibration ability of the model was not high, even on the original scale data set. Model calibration is another important characteristic of evaluating the clinical significance of prediction models. A well-calibrated model can provide more useful information for clinical decisions [58,59].

We can further consider using isotonic regression [60] to calibrate the model to improve its clinical application value. In addition, although the solution to deal with missing, imbalanced, and sparse features proposed by us is not the latest method, it is sufficient to solve the main data quality problems encountered in the development of prediction models for sudden death, which is reflected in the improvement of model performance and the consistency of the risk factors of sudden death obtained with the earlier literature results. In the future, we need to further explore the latest methods to solve these 3 data quality problems so as to extend the data-processing process to other data sets and provide a more reliable data source for prediction models. With regard to the construction of risk factor prediction models for patients with sudden death, we have a broad definition of sudden death, including patients who have undergone rescue or death events. These patients may include some nonemergency death cases, which may have a confusing effect on the final model. In addition, our feature selection was completely based on data, and only the remaining variables were trained in groups during the model training stage. This form can reduce the complexity of manually selecting features and also explore some potential risk variables. However, some clinically significant variables will also be discarded. Therefore, whether the model has clinical guiding significance remains to be further investigated. As a case study, we used LR as the main prediction model, which facilitated us to develop and analyze the risk factors of sudden death. However, the processing capacity of the LR model for nonlinear predictors is insufficient, resulting in insufficient performance of the developed model [17]. This can be seen from the results of other ML models we additionally developed (the RF and GBM had the best performance, with an AUC of 0.936 and 0.931, respectively, which are better than LR models). Therefore, in the future, we will further optimize the data-preprocessing process and try to develop ML models with better performance to improve the clinical usability.

Conclusion
Our work proposes to use ML methods to deal with data quality issues, such as missing data, data imbalance, and sparse features in emergency data, so as to improve data availability. In addition, the risk factors of sudden death in emergency patients are obtained from our model analysis. As a preliminary analysis result, this result is also the basis for the later use of ML algorithms to build the feature selection and data analysis of the prediction model of sudden death in emergency patients.

Acknowledgments
The authors would like to show their appreciation to the engineers working in the Information Centre (ICT) department of the Hainan Hospital of Chinese PLA General Hospital for their help with data preparation. The publication of this paper was funded by grants from the National Natural Science Foundation of China (no. 82102187) and the Hainan Natural Science Foundation Youth Fund (no. 620QN380).

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

Authors’ Contributions
XC carried out the methodological study and drafted the manuscript. HC collected and processed the data and drafted the manuscript. SN made the conceptual design and made critical revisions to the manuscript. XK reviewed the methodology and reviewed the manuscript. HD also reviewed the manuscript. HD conceptualized the study and performed a critical review.
References


Abbreviations

AUC: area under the curve
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
ED: emergency department
EMR: electronic medical record
EWS: early real-time early warning system
GBM: gradient boosting machine
LASSO: least absolute shrinkage and selection operator
LDA: linear discriminant analysis
MAR: missing at random deletion
MCAR: missing completely at random
ML: machine learning
MNAR: not missing at random
LR: logistic regression
OR: odds ratio
PC: principal component
PCA: principal component analysis
PR: precision-recall
RF: random forest
ROC: receiver operating characteristic
SCD: sudden cardiac death
SVM: support vector machine
Barriers and Opportunities for the Use of Digital Tools in Medicines Optimization Across the Interfaces of Care: Stakeholder Interviews in the United Kingdom

Clare Tolley¹,², PGCert, MPharm, PhD; Helen Seymour³, BPharm, PGDip; Neil Watson¹,², MSc, MBA; Hamde Nazar¹, MPharm, PhD; Jude Heed¹, BPharm, MSc; Dave Belshaw³, RMN, MPA

¹School of Pharmacy, Newcastle University, Newcastle upon Tyne, United Kingdom
²Pharmacy Department, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, United Kingdom
³North East and North Cumbria Academic Health Science Network, Newcastle upon Tyne, United Kingdom

Corresponding Author:
Clare Tolley, PGCert, MPharm, PhD
Pharmacy Department
The Newcastle upon Tyne Hospitals NHS Foundation Trust
Queen Victoria Road,
Newcastle upon Tyne, NE1 4LP
United Kingdom
Phone: 44 0191 282 4488
Email: clare.brown@newcastle.ac.uk

Abstract

Background: People with long-term conditions frequently transition between care settings that require information about a patient’s medicines to be transferred or translated between systems. This process is currently error prone and associated with unintentional changes to medications and miscommunication, which can lead to serious patient consequences. One study estimated that approximately 250,000 serious medication errors occur in England when a patient transitions from hospital to home. Digital tools can equip health care professionals with the right information at the right time and place to support practice.

Objective: This study aimed to answer the following questions: what systems are being used to transfer information across interfaces of care within a region of England? and what are the challenges and potential opportunities for more effective cross-sector working to support medicines optimization?

Methods: A team of researchers at Newcastle University conducted a qualitative study by performing in-depth semistructured interviews with 23 key stakeholders in medicines optimization and IT between January and March 2022. The interviews lasted for approximately 1 hour. The interviews and field notes were transcribed and analyzed using the framework approach. The themes were discussed, refined, and applied systematically to the data set. Member checking was also performed.

Results: This study revealed themes and subthemes pertaining to 3 key areas: transfer of care issues, challenges of digital tools, and future hopes and opportunities. We identified a major complexity in terms of the number of different medicine management systems used throughout the region. There were also important challenges owing to incomplete patient records. We also highlighted the barriers related to using multiple systems and their subsequent impact on user workflow, a lack of interoperability between systems, gaps in the availability of digital data, and poor IT and change management. Finally, participants described their hopes and opportunities for the future provision of medicines optimization services, and there was a clear need for a patient-centered consolidated integrated health record for use by all health and care professionals across different sectors, bridging those working in primary, secondary, and social care.

Conclusions: The effectiveness and utility of shared records depend on the data within; therefore, health care and digital leaders must support and strongly encourage the adoption of established and approved digital information standards. Specific priorities regarding understanding of the vision for pharmacy services and supporting this with appropriate funding arrangements and strategic planning of the workforce were also described. In addition, the following were identified as key enablers to harness the benefits of digital tools to support future medicines optimization: development of minimal system requirements; enhanced IT system management to reduce unnecessary repetition; and importantly, meaningful and continued collaboration with clinical and IT stakeholders to optimize systems and share good practices across care sectors.
Introduction

Background

Medicines are the most common therapeutic intervention in the United Kingdom’s National Health Service (NHS). However, it is estimated that between 30% and 50% of medicines prescribed for long-term conditions are not taken as intended [1]. More than a quarter of the adult population in England live with ≥2 conditions [2], and approximately 15% of people in England take ≥5 medicines a day, with 7% taking ≥8 medicines per day [3]. Furthermore, the prevalence of multimorbidity is estimated to increase substantially, with the proportion of those with >4 diseases almost doubling to 17% by 2035 [4]. Therefore, it is important that individuals receive the maximum benefit from their medicines while minimizing harm. Effective medicines optimization contributes to improved health outcomes, patient care, safety and satisfaction, improved efficiency and use of resources, better use of professional skills, and effective clinical governance [5,6].

Transfer of Care

People with long-term health conditions frequently transition between care settings; consequently, information about a patient’s medicines is regularly transferred or translated between systems. However, the point at which patients transfer across different interfaces of care is high risk and is associated with unintentional changes to medications, errors, and miscommunication [7]. This can have consequences for patients, health care professionals, and the health system as a whole [8]. For example, each interaction with a health care professional may result in medication or treatment changes. Problematic polypharmacy may then occur, whereby multiple medicines are prescribed inappropriately or in which the intended benefits of the prescribed medications are not realized [3].

The Department of Health and Social Care’s report, Good for you, good for us, good for everybody, highlighted that to reduce overprescribing, there is a need for better shared decision-making with patients; better guidance and support for clinicians; more alternatives to medicines, such as physical and social activities and talking therapies; and more structured medication reviews for those with long-term health conditions [3]. However, these goals and initiatives must be supported by effective digital systems that are interoperable and must equip health care professionals with the right information to optimize a patient’s medications [9]. A systematic review of 13 publications found that the use of IT applications such as electronic health records (EHRs), electronic decision support tools, and electronic communication applications had a positive impact on financial and health outcomes [10]. However, contemporaneous and accurate information is often not available, with substantial local variation in practice, which can result in increased workload, duplication of tasks, and errors [11,12]. Currently, there is a complex network of different systems that contain patient health record data in distinct silos throughout a patient’s journey. There is also a range of services available to support medicines optimization activities in the United Kingdom, including those that target transitions in care. For example, the Discharge Medicines Service, New Medicines Service, and NHS Community Pharmacist Consultation Service [6,13,14]. Plans also exist to roll out the electronic prescription service to secondary care and other care settings and to develop a patient-centered consolidated medication record that can be used by health care professionals working in different settings [15]. Such records may be associated with a range of benefits including improved safety, greater flexibility, enhanced ability to respond to patient queries, reduced duplication, and lower costs [16-18]. However, their use is still at an early stage, and challenges have been identified in some studies, including problems with system reliability, technical issues, and patient concerns regarding inaccuracies and the governance around sharing data [17,19]. The government recognizes the need for information to be collected once and then shared among providers to meet an individual’s needs. Interoperability is defined as “the ability of two or more systems or components to exchange information and to use the information that has been exchanged” [20]. Nationally, a range of work is underway to enhance interoperability within health and social care settings and is a clear priority for the United Kingdom [21]. A recent policy paper, Data saves lives: reshaping health and social care with data, highlights commitments to introducing clear and open standards to make it easier to share data safely and efficiently across care settings [22].

Objectives

When considering how services might be delivered in the future and the necessary digital transformation, it is important to understand the current landscape of systems including their benefits, challenges, and opportunities. In this study, we aimed to address the following questions: what systems are being used to transfer information across interfaces of care? and what are the challenges and potential opportunities for more effective cross-sector working to support medicines optimization?

Methods

Overview

The aim of this study was to engage with key stakeholders in medicines optimization and IT across the North East and North Cumbria (NENC) integrated care system (ICS) to scope out current systems related to the transfer of information across interfaces of care. ICSs were established across regions of England on July 1, 2022, and have been described as partnerships of organizations that come together to plan and deliver joined up health and care services. Digital solutions will be central to supporting the function and role of ICSs. In addition, we sought to identify the challenges and potential
opportunities for a more effective cross-sector working to support medicines optimization and inform future priorities.

A qualitative methodology was selected to ensure gathering of a detailed understanding of participants’ experiences and perspectives. A constructivist and interpretivist approach was taken [23] together with the framework approach, which is a method developed for use in applied policy research in which there is a need to address a clear set of aims and objectives, while following an inductive approach that allows theories to develop “bottom-up” [24].

Key stakeholders in medicines optimization and digital health were invited to participate in a semistructured interview to gather their perspectives on the current medicines optimization services provided and the digital tools used to support these activities. In addition, stakeholders were asked to provide their opinions on the barriers and potential opportunities for more effective medicines optimization across sectors.

Eligibility Criteria

We included all clinicians, managers, commissioners, and stakeholders who were or had been involved in commissioning, developing, and delivering medicines optimization–related activities and who had expertise in IT, clinical informatics, and digital health solutions used within the NHS.

Recruitment

We used a snowball sampling approach to identify suitable key participants [25]. In the first instance, the researcher was introduced to an initial set of contacts by the senior medicines optimization pharmacist based at the NENC Academic Health Science Network and the digital transformation director for the NENC Academic Health Science Network. The researcher emailed potential participants and invited them to participate in a semistructured interview. This email also included attachments to a participant information leaflet and a consent form. The participants were required to provide consent via a web-based consent form before participating in the study. We proactively aimed to engage with individuals from a range of professional backgrounds and levels of experience to ensure that the data gathered were rich and representative. Data were collected until thematic saturation was reached, and we used an inductive approach to look for the nonemergence of new themes as interviews and analysis were conducted [25,26].

Data Collection

Semistructured interviews were conducted by 1 researcher (CT) between January and March 2022 to explore stakeholders’ perspectives on the current medicines optimization services provided in the region and the digital tools used to support these activities, including details regarding the flow of health information exchange and interoperability reached. In addition, stakeholders were asked to provide their opinions on the barriers and potential opportunities for more effective medicines optimization across sectors. A flexible topic guide was developed that incorporated open-ended questions and prompts. The guide was shared with a team of researchers and clinicians to review and refine the data before use, and the guide was developed iteratively throughout data collection [27]. Interviews lasted for approximately 1 hour and were conducted via video call by a researcher with clinical and postgraduate level of qualitative data collection experience, at a mutually convenient time for each participant. All interviews were recorded, transcribed verbatim together with accompanying field notes, and anonymized.

Analysis (Interviews)

Qualitative data collection and analysis were iterative, allowing themes to be generated, interpreted, explored, and disconfirming evidence identified [28]. Different data sources, for example, interviews with a range of participants, facilitated triangulation to identify where and how different data converged and diverged. The main themes and subthemes were identified using a constant comparative analysis [28]. For this purpose, data were constantly compared among interviewees to explore similarities and differences between groups and to uncover explanations for why these differences existed. Field notes contributed to the analysis by providing valuable context, for example, in which the participants used their voice to stress points or in which humor was used. Field notes were also used by the interviewer to note their own reflections and consider questions for future exploration [29]. The framework approach was used, which is a 5-staged approach to thematic analysis, enabling previous theories and insights identified through literature review or experience, to inform the development of the thematic framework, while allowing theme generation based on the data and was, therefore, open to discovering unexpected concepts based on the participants’ experiences. This was used as a complementary method along with the constant comparative analysis. Themes were discussed among team members and continually refined and applied systematically to the whole data set using the computerized software N-Vivo (QSR International). All data were analyzed by qualified members of the research staff. “Member checking” was also performed, whereby a draft of the key findings was shared with all participants, who were given a minimum of 2 weeks to provide feedback on the interpretations made and contribute any additional insight [26].

Ethics Approval

This study was approved by the Research, Policy, Intelligence, and Ethics team at Newcastle University (reference: 17851-2021).

Results

Overview

A total of 22 interviews were conducted with 23 participants lasting between 38 and 75 minutes (Table 1). Two participants (a community pharmacist and general practitioner [GP]) were unable to take part in an interview, owing to clinical commitments and availability. “Member checking” resulted in 1 correction to the results (clarification regarding work underway to create a patient medication records) and provided further information regarding national initiatives currently underway to support development of shared care records, which was incorporated into the discussion and recommendations in the Addressing Digital Gaps section.
This study revealed a range of different systems that are used across a region to support medicines optimization activities. From the interviews, a range of themes and subthemes pertaining to three key areas were identified: (1) transfer of care issues, (2) challenges of digital tools, and (3) future hopes and opportunities.

Table 1. Table of participants (N=23).

<table>
<thead>
<tr>
<th>Profession and Sector</th>
<th>Participants, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pharmacists (n=21)</strong></td>
<td></td>
</tr>
<tr>
<td>Community</td>
<td>4 (17)</td>
</tr>
<tr>
<td>Hospital</td>
<td>5 (22)</td>
</tr>
<tr>
<td>Primary care (GPa, practice, primary care network, and CCGb)</td>
<td>5 (22)</td>
</tr>
<tr>
<td>North East Ambulance Service</td>
<td>2c (9)</td>
</tr>
<tr>
<td>NHSd England and NHS Improvement</td>
<td>5 (22)</td>
</tr>
<tr>
<td>GPs</td>
<td>2 (9)</td>
</tr>
</tbody>
</table>

aGP: general practitioner.
bCCG: clinical commissioning group.
cOne interview conducted with both participants.
dNHS: National Health Service.

Medicines Optimization Systems

Informed by participant interviews and relevant literature, a simplified overview of the key systems related to medicines across primary, secondary, tertiary, and social care across the NENC was developed (Multimedia Appendix 1 [30-32]). This was not intended to provide a comprehensive overview of all systems used throughout the region but instead to illustrate the complexity of how data are stored and moved between settings. Several patient health record systems have been used in general practice and primary care. A range of different community pharmacy patient medication record systems was also identified, which lacked the ability to directly transfer information with GP systems. At the time of data collection, some hospital trusts in the region used paper-based health records and prescriptions. The transfer of information between different systems and care settings was largely facilitated through bespoke solutions delivered by third-party companies in response to a particular problem, for example, a digital referral as part of the Discharge Medicines Service from 1 hospital to a community pharmacy was typically either sent using NHS mail or a website or integrated web platform such as PharmOutcomes or Cegedim. The information flow between the systems was found to be typically unidirectional.

Transfer of Care Issues

All participants highlighted problems owing to incomplete patient records. Data were described as being held in separate silos by the GP, secondary care providers, or a community pharmacy, with ineffective data flow among them, resulting in reduced efficiency and safety. For example, participants described discrepancies between the allergy status recorded in different clinical systems (eg, missing documented allergies) and omissions owing to poor communication and clinical handover. Incomplete records also made it difficult to proactively provide care by identifying patterns of behavior that would warrant further investigation or management, for example, “if someone’s getting emergency contraception on a regular basis [from a community pharmacy], actually that should be flagging up a risk” (GP 006). A GP recalled a “significant incident where the GP hadn’t put down [methotrexate] or on the GP record was not methotrexate, and they [the patient] ended up in an ITU in Wales and [the staff] didn’t know that the patient was on methotrexate and had actually accidentally overdosed” (GP 006), but the staff were unaware that the patient was even prescribed this medication.

Challenges of Digital Tools

Participants raised concerns on specific challenges associated with the digital tools used to optimize medications. These included the use of multiple systems and workflow, interoperability, digital gaps, IT systems management, and change management.

Multiple Systems and Workflow

Participants working across all settings described the need to interact with multiple IT systems as a part of their day-to-day role. Those working in general practice, for example, would access a core EHR system alongside other systems for viewing additional information, such as hospital notes, appointment letters, or blood tests. Community pharmacists described navigating between a growing number of different systems to fulfill different tasks and purposes (refer to the quote below) and felt that exposure to “more interfaces, [presented] […] more opportunities […] for information to have to be re-transcribed [and] […] actually alert tasks to get dropped, because the right people can’t see [what needs to be done]” (Pharmacist 010). They emphasized how “the more you’re having to step out of your day-to-day workflows and go, ‘Oh my goodness, I really must remember to send a DMS [Discharge Medicines Service] referral for this thing’. You’re just not going to do it” (Pharmacist 010):

*On a Sunday, I do a COVID clinic, so I’ve got Q-Flow open, which is the appointments booking system [...] I’ve got PharmOutcomes open for CPCS referrals*
Interoperability and Safety

A GP described how community nurse practitioners do not have access to the full EHR of certain patients in their care. This occurred because the community nurses were employed by the local hospital rather than by the GP practice, and the 2 organizations used different IT systems, which did not share data in real time. Consequently, automated checks such as drug-laboratory checks or drug-interaction checks are not reliable. There was the risk that if community nurses “don’t have the blood results in their system [so] they will not get that [computerised] warning, so they may merrily go ahead and prescribe that [medication]. Then conversely, if they’ve prescribed a drug, I [the GP] don’t know that the interactions are there now” (GP 012).

Digital Gaps

The participants revealed important gaps in the availability of health data that could not be easily shared or used by health care professionals, for example, paper-based hospital notes. A locally shared health record for people living in NENC known as the “Great North Care Record” (GNCR) is currently in development, with the aim of providing health and care workers access to current medical information. However, 1 GP highlighted how their local acute hospital trust only imported some “very, very, primitive data” into the GNCR and so was not “hugely valuable” (GP 012). In addition, data from social care, for example, care homes, were not imported into the GNCR, which meant that “any changes that are made there are not displayed” (GP 006). In contrast, however, more digitally advanced hospitals transferred a greater quantity of information from their EHR into the GNCR, which was “more useful” (GP 012). This raised the possibility of inequalities among health care providers because organizations “that are [digitally] further behind are the ones that...The Great North Care Record doesn’t help” (GP 012).

Some professional sectors, such as community pharmacy, were also unable to access the GNCR, which made it “really difficult” to maximize use of that workforce and “shift patients away from some of those higher acuity services” (Pharmacist 003). Furthermore, information governance arrangements for accessing multiple systems were considered problematic. For example, 1 pharmacist suggested that the current consent model whereby community pharmacists must obtain consent before accessing information within the summary care record was inappropriate because “if you’ve got a patient’s prescription it should almost imply by informed consent, they’ve given you their script” (Pharmacist 005).

Participants also revealed a low adoption of standardized digital codes for documenting clinical data within health systems. There are “a lot of trusts [that] don’t use dm+D” (Pharmacist 002), and many hospitals are “not using SNOMED at the moment” (GP 006); without adoption of the standards, large initiatives such as shared patient medication records are “not going to work” (Pharmacist 002). Dictionary of medicines and devices is a dictionary of descriptions and codes, which represent medicines and devices used across the NHS. Systematized Nomenclature of Medicine Clinical Terms is a structured clinical vocabulary for use in EHRs and covers diagnoses, procedures, etc.

Finally, the participants noted how certain services were being rolled out without a supportive digital infrastructure. For example, the national hypertension case finding service in community pharmacies lacked “a national system for reporting it [blood pressure] [and reporting] between the two [pharmacy to GP]” (Pharmacist 011). Consequently, a range of different communication techniques, including email, letter, or pilot digital platforms, were being used, depending on local arrangements, and as data were not collected in a standardized digital format, it could not be transferred between care settings in an interoperable way.

IT Systems Management

Participants suggested that a lack of minimum standards or mandating how services are technically delivered is problematic and contributes to low adoption and delays in rolling out clinical services and digital solutions:

> So pharmacy DMS transmissions, we launched the DMS service but pharmacies [are] not seeing many of them because yes you can do it by snail mail and NHS Mail, but that’s not good enough. We shouldn’t launch a service without a platform to deliver it on. So every hospital should be told, “You can have your own, but we’re launching this new service and you must be able to provide a digital solution that’s integrated.” [Pharmacist 005]

A need was identified to rationalize the number of digital systems used, create clear expectations for suppliers, and develop standards and frameworks that outline how services should be digitally enabled. This would reduce waste from “reinventing the wheel every single time” (Pharmacist 020). A new service or digital tool was developed to limit duplication and unnecessary costs. Another pharmacist added the following:

> I think there needs to be a suite of expectations that everybody needs to have and it’s the same with hospitals, you can’t have a hospital system that doesn’t do these 10 things [...] every system needs to do that by a certain deadline. [Pharmacist 005]

Change Management

Change management was seen as important for managing the future development of digital medicines optimization services. For instance, although the clinical terminology standards have now been defined in the Systematized Nomenclature of Medicine (SNOMED) and dictionary of medicines and devices (dm+D), “the hard bit is to come, which is the adoption and more importantly the transformation around that. There’s also
a Hearts and Minds piece” (Pharmacist 002). Similarly, participants described concerns regarding data sharing, for instance:

If you suggest [data sharing] to some general practices, they’ll say: “No chance. Nobody should have access to that data.” But once you get over that barrier, and there’s a bit of trust built into it, then you can start adding to it. [Pharmacist 003]

Instead, all health care professionals need to work together to “best serve the needs of the patients” (Pharmacist 019).

**Future Hopes and Opportunities**

**Consolidated Integrated Care Record**

Participants revealed several ambitions for the future of digital medicines optimization services, although the need for a single shared consolidated medication record, giving all health and care professionals access to data, was a clear priority and reflected those working across care sectors:

One record across all organizations, that’s the blue-sky thinking. The data would sit in a data repository that would be coded and accessible, via APIs, via front-end systems that could be customised to be targeted to how GPs work, targeted too how secondary care clinicians work, targeted for acute and outpatient mental health. Fundamentally, all the data would be held in one central repository for that patient and all of the systems can pull in all that data. A patient, for example from a prescribing basis, would have one prescribing record. That prescribing record would continue out of hospital, into hospital. [GP 012]

This would “open up a number of opportunities [...] for optimisation and proper management of patients and overprescribing” (Pharmacist 002). Furthermore, by identifying and using common digital architectures and standards, such records may be linked in the future.

Some participants were worried that increased access to information could be problematic in some situations, for example, community pharmacists could have “too much information, to make a decision on” because “you could spend hours and hours and hours trailing back through communications and stuff that gets put onto records, [which] it’s probably completely inappropriate to the query” (Pharmacist 018). Consequently, several participants suggested involving end users in the development of shared records and posed that for “a sector and a workforce like community pharmacy, I think it’s big enough to warrant having a bespoke solution developed for it” (Pharmacist 001). A change in funding arrangements was also discussed as vital to support and incentivize community pharmacists to deliver clinical services at scale and to justify the need for access to shared care records in the first instance.

**Uses of Data**

There were hopes that the enhanced availability of medication data could improve workflows across care settings and population health management. Improved efficiency may be realized, for example, by sending a coded list of a patient’s medications from one system to another, so that a clinician only has to “go click, click, click, and it populates the prescribing system” (Pharmacist 010). In addition, to support better interprofessional working and better continuity of care, participants thought it would be good if they could send a “request, where you’ve got specific things that you want following up” (Pharmacist 010) alongside any information and context about a patient’s medication, directly between systems.

Participants also discussed how the development of a comprehensive shared patient record could serve as a “population health platform” (Pharmacist 002). It would then be possible to “start interrogating the information at a patient level, but [also] at a population level” (Pharmacist 002) and target public health challenges such as “overprescribing, opiate prescribing, valproate [prescribing in pregnancy]” (Pharmacist 002). Digital tools could also support better clinical prioritization, for example:

If you’ve got 100 Primary Care Network pharmacists delivering a structured medication review every year or so, that’s 1,000 reviews a week. How do you know which 1,000 patients put in for those slots? How do you caseload? That’s really important. [Pharmacist 016]

Comprehensive patient data repositories could enable clinicians to develop robust strategies to identify patients more efficiently, to avoid “each practice pharmacist going out and trying to design their own searches” (Pharmacist 013). It was clear that there are major opportunities arising from more effective data sharing, and as 1 pharmacist remarked, there are likely innovations that health care professionals have not even started to dream about.

**Discussion**

**Principal Findings**

This research has identified several challenges and potential opportunities related to the use of digital tools for delivering medicines optimization, which are categorized under 3 key concepts: transfer of care issues, challenges of digital tools, and future hopes and opportunities. There is substantial complexity in the number of various medicines management systems used throughout the region and the challenges associated with incomplete patient records. The use of multiple systems also affects the user workflow. There was a lack of intrasystem and intersystem interoperability and important gaps in digital data in some settings (eg, social care and in some areas where hospital prescribing was paper based). Several problems related to IT systems management and change management have also been described. Participants described a clear need for a patient-centered consolidated integrated health record for use by all health and care professionals across different sectors, bridging those working in primary, secondary, and social care.

We also identified a series of recommendations relevant to health service managers, policy makers, and clinical staff, which are discussed in Table 2.
Table 2. Summary table: stakeholder recommendations and objectives.

<table>
<thead>
<tr>
<th>Recommendations and specific objectives</th>
<th>Key stakeholder group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Identify the future vision for pharmacy services and support with appropriate funding plan</strong></td>
<td></td>
</tr>
<tr>
<td>Support realization of the vision through appropriate strategic planning and funding arrangements, for example, community pharmacy contractual framework</td>
<td>Policy makers and service managers</td>
</tr>
<tr>
<td><strong>Development of a patient-centered consolidated integrated health record for use by health and care professionals</strong></td>
<td></td>
</tr>
<tr>
<td>Support digitization of social care, for example, implementation of electronic prescribing and medication administration in care homes</td>
<td>Researchers and service managers</td>
</tr>
<tr>
<td>Creation of local multistakeholder working groups to ensure cocreation of digital solutions and facilitate effective teamwork and “buy-in”</td>
<td>Service managers</td>
</tr>
<tr>
<td>In-depth exploration of concerns around data sharing and governance considerations with key clinical and IT stakeholders and patients and members of the public</td>
<td>Policy makers, service managers, researchers, clinical staff, and public</td>
</tr>
<tr>
<td>Explore need for tailored views for health professional groups to shared care records</td>
<td>Service managers and researchers</td>
</tr>
<tr>
<td>Support adoption of digital solutions across NENC(^a) region’s hospitals with a pledge to end paper records and prescribing</td>
<td>Policy makers, service managers, and clinical staff</td>
</tr>
<tr>
<td><strong>Adoption of medication standards across NHS(^b) (ie, dm+d(^c) and SNOMED CT(^d))</strong></td>
<td></td>
</tr>
<tr>
<td>Identify what is the level of dm+d adoption across the region</td>
<td>Policy makers and service managers (this is in progress via NHS Digital)</td>
</tr>
<tr>
<td>Explore the local facilitators and barriers to adoption of standards</td>
<td>Policy makers, service managers, and researchers</td>
</tr>
<tr>
<td>Prioritize communication around the need to adopt standards locally</td>
<td>Policy makers, service managers, and clinical staff</td>
</tr>
<tr>
<td>Support sites through the transformation process, using tools from NHS England or FCI(^e)</td>
<td>Service managers</td>
</tr>
<tr>
<td><strong>Rationalizing number of systems and services to reduce unnecessary repetition</strong></td>
<td></td>
</tr>
<tr>
<td>Mapping process to identify services provided across sectors and highlight duplication and gaps</td>
<td>Service managers and clinical staff</td>
</tr>
<tr>
<td>Identify clinical services that would benefit from digitization in the community pharmacy sector and to support service provision and service management or audit</td>
<td>Service managers and researchers</td>
</tr>
<tr>
<td><strong>Development of frameworks and minimum standards outlining how services should be delivered using digital means at a local and national level</strong></td>
<td></td>
</tr>
<tr>
<td>Monitor guidance and toolkits from NHS organizations (eg, NHS Transformation Directorate) and publish updates, publish bullelins, and organize workshops to increase awareness of an organization’s responsibilities and share lessons across the region</td>
<td>Policy makers, service managers, researchers, and clinical staff</td>
</tr>
<tr>
<td>Use and promote the use of forums for communication across all sectors and levels from manager to frontline staff</td>
<td>Service managers and clinical staff</td>
</tr>
<tr>
<td><strong>Cross-sector communication around approaches to digital medicines optimization</strong></td>
<td></td>
</tr>
<tr>
<td>Development or use of established forums and groups to proactively collate and communicate examples of good practices between different clinical and IT stakeholders</td>
<td>Policy makers, service managers, and researchers</td>
</tr>
<tr>
<td><strong>Working with suppliers to develop integrated solutions to avoid unnecessary development of bespoke solutions</strong></td>
<td></td>
</tr>
<tr>
<td>Share lessons across organizations about successful collaborations with suppliers on innovation projects</td>
<td>Service managers, researchers, and clinical staff</td>
</tr>
<tr>
<td>Harness insight from existing projects to support scale-up of innovations</td>
<td>Service managers and researchers</td>
</tr>
</tbody>
</table>

\(^a\)NENC: North East and North Cumbria.  
\(^b\)NHS: National Health Service.  
\(^c\)dm+d: dictionary of medicines and devices.  
\(^d\)SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.  
\(^e\)FCI: Faculty Clinical Informatics.

**Addressing Digital Gaps**

The provision of health services in the United Kingdom needs to substantially change to meet the needs of an aging population, with a key focus on more integrated care across health and care settings to local needs [33]. This requires better collaboration between different professionals working across care sectors who have the right information to inform decision-making, with...
digital developments recognized as key to the transformation [33,34]. To enable this, participants emphasized the need for a single consolidated electronic patient record where all health and care professionals can read and write into and share information across traditional boundaries to deliver acute care, manage long-term conditions, and ensure patients receive the right care at the right time and place. For example, most community pharmacists in the United Kingdom do not have access to an up-to-date and comprehensive list of a patient’s medications or medical records, which hinders their ability to support individuals. However, pilot projects such as the Somerset shared care record case study (SIDeR) are underway with some promising feedback, although there is a need for further evaluation to fully understand the benefits and any unintended consequences or challenges [35]. There is a strong relationship between the maturity of digital health and comprehensive evaluation methods; therefore, this should be included as part of future locally shared care strategies to support learning and facilitate the development and adoption of systems [36]. In addition, as Cresswell et al [37] noted, there are several socio-organizational dimensions of change that must be considered to support a digitally enabled shared care agenda. They summarize key areas, such as structural and organizational complexity; variations in data management and expectations; poorly defined current shared care pathways; and issues associated with reluctance to data sharing, managing “data overload;” and configuring systems appropriately. In terms of technological dimensions, existing infrastructures and legacy systems may hinder data sharing across new technological junctions, while it may also be difficult to connect incompatible data structures and overcome supplier resistance to make distinct systems interoperable [37]. In their report, they advocated the need to map potential architectural components and designs for shared care solutions with careful consideration of their potential benefits and limitations [37].

In our study, participants highlighted that the need to undertake the process of developing shared records with end users to ensure the design and functionality is “fit for purpose.” Specifically, participants raised the question of developing tailored solutions, that is, a bespoke community pharmacy view of a patient’s health record, to enhance the usability and utility of such systems. Research has shown how the design of EHRs can influence behavior and prescribing safety [38,39]; therefore, further work is needed to explore how to create usable health records for a range of different end users working across sectors.

Participants also expressed specific ways in which data could be used to better support the transfer of care, for instance, sending digital referrals or requests for follow-up directly between existing clinical systems, further demonstrating the need for continued engagement and collaboration with end users involved in the delivery of frontline services to optimize and enhance systems over time [40].

To fully harness the benefits of a comprehensive and consolidated shared patient record, there is a need to address digital gaps within organizations across the health and social care sector. The effectiveness and utility of the tool depends on the data within. First, organizations must prioritize the use of established and approved digital information standards related to medicines and clinical information within NHS digital systems. Notably, all NHS care providers who are involved in prescribing, dispensing, or administering medicines must transfer medication information using the newest UK version of fast health care interoperability resources, use approved dose syntax to transfer the amount of medication per dose as a simple coded quantity, and use SNOMED and dm+d codes for allergy or intolerance information by March 31, 2023 [41]. However, barriers to the adoption of such standards have been identified, including a lack of cohesive national-scale digital health system; funding and support for standards; knowledge and infrastructure related, such as the impact on preexisting workflows; and lack of use of a consistent patient ID [42,43]. Therefore, local ICSs must explore how they can address such challenges and support the implementation of the standards across the region. This underscores the importance of effective clinical leadership and understanding the personal factors that influence health IT uptake [44,45]. The adoption of digitized health records, electronic prescriptions, and medication administration across primary, secondary, and social care is vital. As is embedding digital technologies across social care, particularly as estimates suggest that less than half of social care providers have any form of digital care records in England [46]. Furthermore, research has shown that a large proportion of medication errors occur in care homes [47]; therefore, tools to support the digitization of the sector to support medicines optimization and enhance the safety, quality, and efficiency are urgently needed. Our findings echo those of a recently published report by the Royal Pharmaceutical Society Scotland, *Pharmacy 2030: a professional vision document*, which outlines the changes to and enhanced roles of pharmacy professionals and key enablers, including data to inform decision-making, harnessing digital technology, developing the workforce, and increasing emphasis on multidisciplinary work [34]. A further report and policy review from the Royal Pharmaceutical Society and The King’s Fund that will inform the development of vision for pharmacy was published at the end of 2022.

To support the delivery of enhanced medicines optimization activities across care settings, there will be an increased emphasis on pharmacists and pharmacy staff undertaking clinical roles in all sectors. Community pharmacies, for example, will have a far greater role in providing enhanced clinical services and supporting the holistic prevention of ill health in a community [48,49]. The supply of medicines will be facilitated by accuracy checking technology, such as dispensing robots, or possibly through hub-and-spoke model dispensing [50]. To support this, the participants in this study highlighted the need for changes in how services are funded through the community pharmacy contractual framework. Although the demand for community pharmacies has risen since the outbreak of the Covid-19 pandemic, staff shortages are a growing concern; 1 survey reported that 91% (estimated from responses from 418 representatives of 5000 pharmacy premises) of pharmacies have experienced staff shortages [51]. There are also huge concerns around staffing shortages more widely across the NHS, which has an impact on patient care, while awaiting the results and recommendations of the NHS long-term workforce strategy [52]. This is important because changes resulting from the implementation of new digital tools, particularly systems with
poor usability, can contribute to clinician burnout and consequently reduce job satisfaction, quality and safety of care, and costs [53]. Any digital transformation relating to how medicines optimization is delivered should, therefore, be mindful of the working environment and additional stressors present, and improving the working life of health and social care providers should be a goal and actively monitored [54]. In addition, IT systems management is needed to closely monitor the number of digital systems used across an ICS footprint and rationalize how services are delivered to reduce unnecessary duplication and streamline services. This may be supported by forums and established groups with a key objective of proactively collating and communicating examples of good practices between different clinical and IT stakeholders.

Limitations
We acknowledge some limitations of this study; semistructured interviews were conducted via web-based videocall platforms, which enabled participants across a large geographical area to participate; however, we did encounter some technical issues, which may have impacted the flow and nature of discussions. We only included 2 GPs in this study; however, 5 primary care and clinical commissioning group pharmacists who provided in-depth detail about the experiences of medicines optimization of health care professionals working within GP practices were recruited, and data were collected until thematic saturation was reached. Further work may specifically explore the challenges experienced within different sectors and enhanced by using observational data collection approaches. Furthermore, the interviewer was professionally known to a small number of participants before the study, and her experience as a practicing pharmacist and academic researcher enabled her to build rapport and relationships with the participants. Throughout data collection and analysis, notes were recorded on any personal reactions or reflections to support consideration of the intersubjective reflexivity between herself and the participants [55]. This was used during the analysis stage to provide context and allowed the researcher to honestly and critically reflect on their role in interpreting the data. It also prompted further exploration of ideas and themes during the data collection stage [55]. We collected data from a range of participants across different sectors; however, this was limited to the NENC regions and so may not be representative of other parts of the United Kingdom. Finally, although the analysis and results were discussed with the research team and member checking was performed with the participants, only 1 researcher coded the data collection transcripts, which could have further decreased the rigor and replicability of our work.

Conclusions
The findings from this qualitative study of 23 clinical and IT stakeholders identified major complexity in terms of the number of different systems used throughout the NENC region and identified several important challenges in the transfer of care issues, focusing on having access to incomplete patient records. We also highlighted important barriers related to the use of digital tools, such as multiple systems and workflow, interoperability, digital gaps, IT systems management, and change management. Finally, participants discussed their future hopes and opportunities for the provision of medicines optimization services in the future, and there was a clear need for a patient-centered consolidated integrated health record for use by health and care professionals across different sectors, which would be fundamental to delivering effective and safe patient care. Further specific priorities were around understanding the vision for pharmacy services and supporting it with appropriate funding arrangements and strategic planning of the workforce, adoption of digital information standards, and development of minimal system requirements and frameworks. In addition, IT system management to reduce unnecessary repetition and, importantly, meaningful and continued collaboration with stakeholders and system suppliers to optimize systems and share good practices across care sectors were important key enablers.

Acknowledgments
This study was supported by the Academic Health Science Network and the National Health Service England Digital Primary Care First program. The funder had no input in the study design and data collection, analysis, or interpretation. The authors would like to thank the participants for their time and valuable contributions to this work.

Authors' Contributions
All the authors were responsible for the conception and study design. CT performed data collection and liaised with all authors on the analysis. CT led the writing of this manuscript, with all authors commenting on drafts. All authors have read and approved the final manuscript for submission.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Simplified overview of digital tools used across North East and North Cumbria to support medicines optimization activities.

References

https://medinform.jmir.org/2023/1/e42458


32. Chemomisc. CIS Oncology. URL: https://www.cis-healthcare.com/chemocare/ [accessed 2023-02-16]


41. DAPB4013: medicine and allergy/intolerance data transfer. NHS Digital. URL: https://tinyurl.com/yfhap3kw [accessed 2023-02-16]


47. Tolley et al. JMIR MEDICAL INFORMATICS 2023 | vol. 11 | e42458 | p.191 https://medinform.jmir.org/2023/1/e42458


Abbreviations

- **dm+d**: dictionary of medicines and devices
- **EHR**: electronic health record
- **GNCR**: Great North Care Record
- **GP**: general practitioner
- **ICS**: integrated care system
- **NENC**: North East and North Cumbria
- **NHS**: National Health Service
- **SNOMED**: Systematized Nomenclature of Medicine

©Clare Tolley, Helen Seymour, Neil Watson, Hamde Nazar, Jude Heed, Dave Belshaw. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 10.03.2023. 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.
Using the H2O Automatic Machine Learning Algorithms to Identify Predictors of Web-Based Medical Record Nonuse Among Patients in a Data-Rich Environment: Mixed Methods Study

Yang Chen¹, BSc; Xuejiao Liu¹, BEng; Lei Gao², LLM; Miao Zhu³, PhD; Ben-Chang Shia⁴,⁵, PhD; Mingchih Chen⁴,⁵, PhD; Linglong Ye⁶, PhD; Lei Qin¹,⁷, PhD

¹School of Statistics, University of International Business and Economics, Beijing, China
²School of Law, University of International Business and Economics, Beijing, China
³School of Statistics, Huaqiao University, Xiamen, China
⁴Graduate Institute of Business Administration, College of Management, Fu Jen Catholic University, New Taipei City, Taiwan
⁵Artificial Intelligence Development Center, Fu Jen Catholic University, New Taipei City, Taiwan
⁶School of Public Affairs, Xiamen University, Xiamen, China
⁷Dong Fureng Institute of Economic and Social Development, Wuhan University, Wuhan, China

Corresponding Author:
Lei Qin, PhD
School of Statistics, University of International Business and Economics
No.10, Huixin Dongjie, Chaoyang District
Beijing, 100029
China
Phone: 86 01064491146
Email: qinlei@uibe.edu.cn

Abstract

Background: With the advent of electronic storage of medical records and the internet, patients can access web-based medical records. This has facilitated doctor-patient communication and built trust between them. However, many patients avoid using web-based medical records despite their greater availability and readability.

Objective: On the basis of demographic and individual behavioral characteristics, this study explores the predictors of web-based medical record nonuse among patients.

Methods: Data were collected from the National Cancer Institute 2019 to 2020 Health Information National Trends Survey. First, based on the data-rich environment, the chi-square test (categorical variables) and 2-tailed t tests (continuous variables) were performed on the response variables and the variables in the questionnaire. According to the test results, the variables were initially screened, and those that passed the test were selected for subsequent analysis. Second, participants were excluded from the study if any of the initially screened variables were missing. Third, the data obtained were modeled using 5 machine learning algorithms, namely, logistic regression, automatic generalized linear model, automatic random forest, automatic deep neural network, and automatic gradient boosting machine, to identify and investigate factors affecting web-based medical record nonuse. The aforementioned automatic machine learning algorithms were based on the R interface (R Foundation for Statistical Computing) of the H2O (H2O.ai) scalable machine learning platform. Finally, 5-fold cross-validation was adopted for 80% of the data set, which was used as the training data to determine hyperparameters of 5 algorithms, and 20% of the data set was used as the test data for model comparison.

Results: Among the 9072 respondents, 5409 (59.62%) had no experience using web-based medical records. Using the 5 algorithms, 29 variables were identified as crucial predictors of nonuse of web-based medical records. These 29 variables comprised 6 (21%) sociodemographic variables (age, BMI, race, marital status, education, and income) and 23 (79%) variables related to individual lifestyles and behavioral habits (such as electronic and internet use, individuals’ health status and their level of health concern, etc). H2O’s automatic machine learning methods have a high model accuracy. On the basis of the performance of the validation data set, the optimal model was the automatic random forest with the highest area under the curve in the validation set (88.52%) and the test set (82.87%).
Conclusions: When monitoring web-based medical record use trends, research should focus on social factors such as age, education, BMI, and marital status, as well as personal lifestyle and behavioral habits, including smoking, use of electronic devices and the internet, patients’ personal health status, and their level of health concern. The use of electronic medical records can be targeted to specific patient groups, allowing more people to benefit from their usefulness.

(JMIR Med Inform 2023;11:e41576) doi:10.2196/41576

KEYWORDS
web-based medical record; predictors; H2O’s automatic machine learning; Health Information National Trends Survey; HINTS; mobile phone

Introduction

Background

Regular review of self–medical records by patients can enhance patient-doctor communication and facilitate disease treatment. Effective communication can increase patient satisfaction, acceptance, adherence, and co-operation with the medical team. It can also improve a patient’s physiological and functional status [1]. Conversely, poor communication between doctors and patients can lead to poor quality and continuity of care [2]. Therefore, ensuring good communication by recording, processing, and sharing health information with patients is a necessary and integral part of the health care process. Encouraging patients to use medical records can reduce unnecessary duplication of testing and treatment [3].

Before the advent of electronic medical records, traditional paper-based medical records written in technical language and comprising raw data were provided to health care professionals. However, such medical records can be worrying and confusing for patients. Consequently, clinical trials that provided written records to patients at the time reported that the use of medical records by patients had little success in enhancing communication and facilitating disease treatment [4-7]. However, with the advent of electronic storage of medical records and the internet, patients can be provided with web-based access to their medical records. Internet-accessible medical records may be particularly helpful to patients compared with centrally stored paper-based medical records. Patients can review web-based medical records repeatedly at their convenience. The readability optimization of web-based cases and the increasing popularity of internet medical information have made understanding web-based medical records easier for patients. Moreover, with the current COVID-19 pandemic, the use of web-based medical records may become more prevalent.

Studies have shown that providing patients with internet-accessible medical records may lead to modest benefits. For example, overall adherence to medical advice improved among patients using web-based medical records. A trend of improvement in satisfaction with doctor-patient communication and facilitation of disease treatment. Effective communication can increase patient satisfaction, acceptance, adherence, and cooperation with the medical team. It can also improve a patient’s physiological and functional status [1]. Conversely, poor communication between doctors and patients can lead to poor quality and continuity of care [2]. Therefore, ensuring good communication by recording, processing, and sharing health information with patients is a necessary and integral part of the health care process. Encouraging patients to use medical records can reduce unnecessary duplication of testing and treatment [3].

Some studies have applied traditional statistical methods to explore the relationship between certain factors and web-based medical record nonuse. For example, using univariable and multivariable regression models, Gerber et al [10] analyzed the use of MyChart (a personal health record portal for electronic medical record systems) among patients attending a National Cancer Institute–designated cancer center and predictors of MyChart use. Using data from the Health Information National Trends Survey (HINTS) cycle 3, Elkefi et al [9] applied descriptive statistics and chi-square tests to explore why patients tended to avoid using web-based medical records and compared patients’ perceptions of web-based medical records based on demographics and cancer diagnoses. On the basis of the 2017 to 2018 HINTS data, Patel and Johnson [11] used descriptive statistics and hypothesis testing to assess individuals’ access, viewing, and use of their web-based medical records and the use of smartphone health apps and other electronic devices in 2017 and 2018. Trivedi et al [12] used multivariable logistic regression (LR) analyses to examine the association between sociodemographic and health care–related factors on being offered access to web-based medical records and accessing web-based medical records and cited reasons for not accessing web-based medical records. These studies used traditional and relatively simple statistical methods, and the selection of predictors has certain limitations. As in the research by Elkefi et al [9], predictors relate only to demographic variables and cancer diagnoses. Screening of predictors based on a data-rich environment can be optimized.

With the rapid development of artificial intelligence, machine learning methods have received increasing attention. Machine learning algorithms are used in a wide variety of applications, such as in medicine and health care, where it is difficult or unfeasible to develop conventional algorithms for necessary tasks [13]. Compared with traditional regression-based statistical methods, machine learning is data-driven and has the advantage of not assuming the distribution and relationship of predictors, and machine learning algorithms are good at handling data that are multidimensional and multivariate. Deep learning is a step forward, which makes feature engineering part of the learning task, reducing the algorithm’s dependence on feature engineering. However, the parameters of the machine learning method greatly influence model accuracy. Incorrect parameter selection and a small sample size can both lead to reduced model performance. Some parameters (such as the number of trees, learning rate, and number of leaf nodes in the random forest)
method) determine the structure and training method of the model, which affects prediction performance. To take full advantage of the relevant machine learning algorithms, an appropriate strategy must be developed to determine the parameters.

Objectives

In this study, explanatory variables were chosen based on a data-rich environment. Data for this study were collected from the National Cancer Institute 2019 to 2020 HINTS. The HINTS regularly collects nationally representative data about the American public’s knowledge of, attitudes toward, and use of cancer- and health-related information; therefore, this study is based on the relevant background in the United States. We used almost all the questions in the questionnaire as possible predictors, thus avoiding the subjectivity of manual screening. To resolve the parameter selection problem of machine learning algorithms, this study adopted the current popular H2O (H2O.ai) automatic machine learning algorithms to realize the automation of the entire process, from construction to application of the machine learning model. At the same time, we also used the traditional statistical method of LR. To the best of our knowledge, this is the first study that applied a range of H2O’s automatic machine learning algorithms to such a large representative sample based on a data-rich environment. We implemented a combination of H2O’s automatic machine learning methods and a data-rich environment. Predictors of web-based medical record nonuse were identified based on the results of the H2O automatic machine learning methods.

Methods

Data Source

Data for this study were collected from the National Cancer Institute 2019 to 2020 HINTS. The HINTS regularly collects nationally representative data about the American public’s knowledge of, attitudes toward, and use of cancer- and health-related information. Survey researchers use the data to understand how adults (aged ≥18 years) use different communication channels, including the internet, to obtain vital health information for themselves and their loved ones. This study analyzed merged data from cycles 3 to 4. Data from cycle 3 were collected between January 2019 and May 2019, and those from cycle 4 were collected from February 2019 to June 2019. We screened the respondents based on the target-dependent variable (ie, web-based medical record nonuse), leaving respondents with no missing values in the target-dependent variable. Finally, 9072 respondents were screened.

Ethical Considerations

The HINTS administration was approved by the institutional review board at Westat Inc and deemed exempt by the National Institutes of Health Office of Human Subjects Research. This exemption also extends to this study. HINTS data are available for public use. Additional information on the survey design is available on the HINTS website.

Statistical Analysis

Explanatory variables were selected in this study based on a data-rich environment, and all questions in the questionnaire that could be answered by all participants were selected (P<141; variables that could only be answered by a specific group were not considered, such as questions only for females, eg, whether they had been screened for cervical cancer). The sociodemographic characteristics and other relevant variables of individuals who had or had not used web-based medical records were compared using chi-square tests for categorical variables and 2-tailed t tests for continuous variables. According to the results of the aforementioned statistical tests, significant variables were selected. Preliminary screening of variables was completed (P<49; some variables were merged and answers were regrouped; refer to Multimedia Appendix 1 for details). Samples with missing values for the preliminary screened variables were excluded, and accordingly, a total of 4827 samples were obtained. On the basis of these samples, 5 algorithms were used for modeling: LR, automatic generalized linear model (auto-GLM), automatic random forest, automatic deep neural network (auto–deep learning), and automatic gradient boosting machine (auto-GBM). Of them, LR is a traditional statistical method, and the last 4 are automated machine learning algorithms based on the R interface (R Foundation for Statistical Computing) of the H2O extensible machine learning platform.

We divided the data set as follows: 80% of the data were used as the training set, and 20% of the data were used as the test set. We used the method of 5-fold cross-validation on the training data to determine hyperparameters and used the selected optimal hyperparameters to fit the model using all training data and make predictions on the test set. To evaluate the predictive accuracy of the models, we reported the accuracy, precision, recall, F1-score, and area under the curve (AUC) of the validation set (validation set results for 5-fold cross-validation in the training set) and test set. The LR model selected predictors through backward selection and stepwise regression. The relative effects of the predictors in the LR model were measured based on crude odds ratios (ORs), whereas the variability and significance were assessed based on CIs and the corresponding P values. Variable importance values were used in the other 4 H2O automatic machine learning classification algorithms to identify predictors (variables with higher importance indexes were screened as predictors, and a 5-fold cross-validation method was used to select important variables by using all data). All statistical analyses were performed using the R software (version 4.1.2). In this study, P<.05 was considered statistically significant.

Measures

Nonuse of Web-Based Medical Records

Web-based medical records are used to organize processes in clinical and outpatient settings and forge doctor-patient communication that establishes mutual understanding and trust. The variable “nonuse of web-based medical records” in this study was calculated based on the following question in the HINTS: “How many times did you access your web-based medical record nonuse were identified based on the results of the H2O automatic machine learning methods.
medical record in the last 12 months?” We used this question to identify users and nonusers of web-based medical records. The respondents who reported accessing their web-based medical records at least once were coded as users, and those who reported accessing their records 0 times were coded as nonusers.

**Demographic and Other Related Variables**

Demographic variables of interest (dichotomized for analysis) included sex (male and female), race and ethnicity (non-Hispanic White and racial and ethnic minority group), education (high school or lower and more than high school), income ranges (<US $20,000 and ≥US $20,000), area (nonmetropolitan and metropolitan), and marital status (married and not married), as well as numerical demographic variables, including age (continuous years) and BMI.

For further analysis, we selected as many variables as possible from the HINTS database to identify their relationship with the use of web-based medical records. Statistical tests were performed on almost all variables in the questionnaire, including chi-square tests for categorical variables and 2-tailed t tests for continuous variables. The variables that passed the significance test were used as potential predictors, as follows (consistent with the question blocks in the questionnaire): 6 variables, such as Confidence in access to health information, in part A (looking for health information); 6 variables, such as Internet use, in part B (using the internet to find information); 2 variables, such as Have regular health providers, in part C (your health care); 3 variables, such as Health provider maintain MR (medical record), in part D (medical records); Care for someone in part E (caregiving): 7 variables, such as General health, in part F (your overall health); 2 variables, such as Notice calorie information, in part G (health and nutrition); 2 variables, such as Exercise days per week, in part H (physical activity and exercise); 3 variables, such as Smoke, in part K (tobacco products; this part of the questionnaire was about the respondents’ consumption of tobacco products); 3 variables, such as Ever tested colon cancer, in part L (cancer screening and awareness); Ever had cancer in part M (your cancer history); 4 variables, such as Everything cause cancer, in part N (beliefs about cancer); and 1 numerical variable, Sitting time per day.

Specific variables and their descriptive statistics are shown in Multimedia Appendix 2, and Multimedia Appendix 1 lists details of some of the aforementioned variables, including demographic variables and variables adjusted for research needs with the readjustment information.

**Machine Learning Methods**

**LR Model**

LR is a generalized linear regression analysis model that is part of supervised learning in machine learning. LR usually uses numerical or categorical independent variables \(x_1, x_2, \ldots, x_n\) to predict the value of the categorical dependent variable \(y\) to determine the probability that \(y\) belongs to a particular category.

\[
p(y = 1 | x_1, x_2, \ldots, x_n) = \left(1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n}\right)^{-1}
\]

The OR expresses the ratio between the probability \(p\) that the dependent variable \(y = 1\) and the probability \(1 - p\) that the dependent variable \(y = 0\). The OR is related to the interpretability of LR. When \(x_i\) is increased by 1, the odds become the original \(e^{\beta_i}\) times.

\[
\logit(p) = \ln(p/(1-p)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n
\]

In the aforementioned formula, \(\beta_1, \beta_2, \ldots, \beta_n\) are the coefficients that measure the contribution of the independent variables \(x_1, x_2, \ldots, x_n\) to \(y\). If the coefficient \(\beta\) is positive, \(e^{\beta} > 1\) and the factor have a direct correlation with \(y\), whereas if \(\beta\) is negative, \(e^{\beta}\) is between 0 and 1.

**H2O’s Auto-GLM**

Generalized linear models (GLMs) were proposed and published by Nelder and Wedderburn [15] in 1972. It is a modeling method that can solve the problem that ordinary linear regression models cannot handle discrete dependent variables. GLM is an extension of the linear model and establishes the relationship between the mathematical expectation of the response variable and the linear combination of predictor variables through a link function. In this study, 5-fold cross-validation was adopted on the data set to select the hyperparameters of the model, and the selection range of the regularization parameter was \((0, 1)\). Ridge regression (\(\alpha=0\)) was used in the regression of the GLM for the variable selection and final classification. The importance of the variable was judged according to the “absolute value of the normalization coefficient” indicator; the larger the value, the greater the importance of the variable. This study used the “h2o.glm” function in the “h2o” package to build a GLM for the classification of web-based medical record use.

H2O is an open-source, in-memory, distributed, fast, and scalable machine learning and predictive analytics platform that allows users to build machine learning models on data and avoid the tedious process of manual hyperparameter tuning. H2O supports traditional (or “Cartesian”) grid searches. In a Cartesian grid search, users specify a set of values for each hyperparameter that they want to search, and H2O trains a model for every combination of the hyperparameter values. This means that, if we have 3 hyperparameters and specify 5, 10, and 2 values for each, the grid will contain a total of \(5 \times 10 \times 2 = 100\) models. After the grid search is complete, the user can query the grid object and sort the models by a specific performance metric (eg, “AUC”) and select the locally best model within the specified parameter range.

**H2O’s Automatic Random Forest**

The random forest is a multivariate statistical technique that considers an ensemble (forest) of trees for efficiency and predictive power [16]. Random forest uses a bagging technique (bootstrap aggregation) to select resamples randomly and choose a random sample of variables at each tree node as the training data set for model calibration. As the random selection of the training data set may affect the model’s results, a large set of trees is applied to guarantee model stability. In this study, the
selection range of the number of trees was between 100 and 500, the selection range of the number of variables in the variable selection set at the node of the tree was approximately $p^{0.5}$ ($p$ is the total number of variables), and the maximum tree depth was selected from 10 to 30. When selecting variables, the parameters of the final model selected under 5-fold cross-validation were as follows: the number of trees was 150, the number of variables contained in the variable selection set at the node of the tree was 7, and the maximum depth of the tree was 10. The importance of the variable was judged according to the “mean decrease gini” indicator, where the larger the value, the greater the importance of the variable. When fitting the model, the final parameters were as follows: the number of trees was 300, the number of variables included in the variable selection set at the node of the tree was 2, and the maximum depth of the tree was 10. Model-fitting processes were implemented using the “h2o.randomForest” function in the “h2o” package. Parameter tuning was implemented using the “h2o.grid” function in the “h2o” package by grid searching for parameters.

**H2O’s Auto–Deep Learning**

The concept of deep learning originates from the study of artificial neural networks, and a multilayer perceptron with multiple hidden layers is a basic deep learning structure. Deep learning algorithms try to identify potential relationships in a data set by mimicking human brain functions. Similar to the human brain structure, deep learning models consist of neurons in complex and nonlinear forms. Deep learning models have 3 basic types of layers: input, hidden, and output layers. Each neuron in the current layer is connected to the input signal of each neuron in the previous layer. In each connection process, the signal from the previous layer is multiplied by a weight, and a bias is added and then passed through a nonlinear activation function through multiple composites of simple nonlinear functions to achieve a complex input space–to–output space map. In this study, the input values were observations of 49 variables, and the output value was the probability of the use of web-based medical records. When training the model, the number of hidden layers was 2 to 3; the number of nodes in the first layer was between 100 and 200; the number of nodes in the second layer was between 50 and 100; and the number of nodes in the third layer was 5, if any. The activation function was selected from the rectifier and rectifier with dropout; dropout ratio defaults to 0.5. The deep learning model chosen using 5-fold cross-validation to be applied for selecting variables contained 3 hidden layers, each with 100, 50, and 5 nodes. When training the model, a $50\%$ random dropout of the nodes was set to prevent overfitting. The variable importance of the model was measured using the combination of absolute values of the coefficients. The final model for classification contained 3 hidden layers, each with 200, 50, and 5 nodes with $50\%$ random dropout, which provided the highest mean AUC in the test set of the model in 5-fold cross-validation. In this study, we used the “h2o.deeplearning” function in the “h2o” package to realize the deep learning algorithm.

**H2O’s Auto-GBM**

The gradient boosting machine (GBM) algorithm is a type of boosting algorithm. GBM is a model that trains decision trees sequentially. Each decision tree is based on the errors of the previous tree. The core idea is to generate various weak learners in series, and the goal of each weak learner is to fit the negative gradient of the loss function of the previous accumulated model. After adding the weak learner, it enables the accumulated model loss to decrease along the negative gradient direction. It uses different weights to linearly combine the basic learners to ensure that learners with greater performance can obtain larger weights. The most commonly used base learners are tree models. Variable importance is determined by calculating the relative influence of each variable: whether that variable is selected to split during the tree-building process and how much the squared error improves or decreases as a result. We used 5-fold cross-validation in both the variable selection and classification model-fitting procedures, which would provide indicators for the selection of optimal hyperparameters. In the process of using a grid search for hyperparameter optimization, it is necessary to train the models under different hyperparameter specifications and evaluate the goodness of fit of the models under these specifications through 5-fold cross-validation. The number of trees ranged from 100 to 500, and the final parameter was set to 300 in both the variable selection and classification model-fitting procedures, which promises a balance in the training and test set errors in 5-fold cross-validation. In addition, the selection of the learning rate ranged from 0.01 to 0.10, and the final parameter was set to 0.01. The maximum depth was between 10 and 30, and the final parameter was set to 10 in both the variable selection and classification model-fitting procedures, which implements a trade-off between model bias and model variance. This study used the “h2o.gbm” function in the “h2o” package to build a GBM for the use of web-based medical record classification problems. Parameter tuning was implemented using the “h2o.grid” function in the “h2o” package by grid searching for parameters.

**Results**

**Descriptive Statistics**

The merged data sets from HINTS cycles 3 and 4 yielded a sample of 9072 respondents, including 5409 (59.62%) nonusers and 3663 (40.38%) users of web-based medical records. Multimedia Appendix 2 presents the frequencies and proportions of the variables. The chi-square test of categorical variables and the 2-tailed $t$ test of continuous variables showed significant differences in some variables between nonusers and users of web-based medical records ($P<.05$). Among the categorical variables, respondents who chose the following options comprised a significantly higher proportion ($P<.05$) of the group not using web-based medical records: “male,” “trust information about health or medical topics from religious organizations and leaders,” “have no drink,” and “smoke more.” For example, in this group, male individuals accounted for 45.23% (2196/4855) of the respondents, whereas in the group using web-based medical records, the percentage decreased to 38.65% (1331/3444). The same was true for the aforementioned variables: “trust information about health or medical topics from
religious organizations and leaders” (1501/4934, 30.42% vs 824/3573, 23.06%), “have no drink” (2623/4692, 55.9% vs 1575/3421, 46.04%), and “smoke more” (824/3573, 39.38% vs 1264/3630, 34.82%). However, those choosing “Non-Hispanic White” (3492/4885, 71.48% vs 2686/3488, 77.01%), “have higher education level” (3569/4741, 75.28% vs 3042/3349, 90.83%), “in marriage” (2481/5198, 47.73% vs 2252/3600, 62.56%), “ever looked for information about cancer” (3841/5379, 71.41% vs 3501/3650, 95.92%), “use electronic” (3752/5346, 70.18% vs 2930/3625, 80.83%), “caring someone” (723/5216, 13.86% vs 656/3604, 18.20%), “general health relative good” (4343/5331, 81.47% vs 3169/3625, 87.42%), “high confidence in ability to take good care of own health” (5036/5337, 94.36% vs 3503/3632, 96.45%), and others were significantly higher (P<.05) in the group using web-based medical records. Among the numeric variables, mean values of age were significantly higher (P<.05) in the group not using web-based medical records, whereas time spent sitting was significantly higher (P<.05) in the group using web-based medical records. The 2-tailed t test of continuous variables also showed no significant difference (P>.05) in some variables between individuals who had and had not used web-based medical records. In other words, the proportions of these variables were similar between the 2 groups. As for BMI, the average value in both groups was approximately 28.5 (SD 0.1).

**Machine Learning Model Results**

As shown in Tables 1 and Table 2, a total of 29 predictors of nonuse of web-based medical records variables (Age, Sitting time per day, BMI, Confidence in access to health information, education, Electronic means use, Ever tested colon cancer, Everything cause cancer, Number of visits to health provider, Have electronic device, income, Obesity affects cancer onset, Social media use, Little interest, Marital status, Offered access to MR by health provider, Offered access to MR by health insurer, Health provider maintain MR, race, Have regular health providers, Seek cancer information, Shared health information, Smoke, Exercise days per week, Strength training days per week, Trust doctor, Trust religious organizations, Internet use, and Electronic wearable device use) were selected in all the 5 algorithms, and 7 variables (Age, Electronic means use, Number of visits to health provider, Offered access to MR by health provider, Offered access to MR by health insurer, Health provider maintain MR, and Internet use) were selected simultaneously in the 5 algorithms.
Table 1. Predictors of nonuse of web-based medical records (MRs) using the logistic regression algorithm.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>OR^a (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race (reference: non-Hispanic White)</td>
<td>1.04 (0.90-1.19)</td>
</tr>
<tr>
<td>Education (reference: high school or lower)</td>
<td>0.32 (0.27-0.37)</td>
</tr>
<tr>
<td>Income (&lt;US $20,000)</td>
<td>0.35 (0.29-0.43)</td>
</tr>
<tr>
<td>Marital status (reference: not married)</td>
<td>0.60 (0.54-0.68)</td>
</tr>
<tr>
<td>Trust doctor(^b) (reference: low_level)</td>
<td>0.73 (0.53-0.99)</td>
</tr>
<tr>
<td>Trust religious organization(^b) (reference: low_level)</td>
<td>1.31 (1.15-1.50)</td>
</tr>
<tr>
<td>Internet use (reference: no)</td>
<td>0.12 (0.09-0.16)</td>
</tr>
<tr>
<td>Electronic means use (reference: no)</td>
<td>0.05 (0.03-0.07)</td>
</tr>
<tr>
<td>Electronic wearable device use (reference: no)</td>
<td>0.40 (0.35-0.45)</td>
</tr>
<tr>
<td>Shared health information (reference: N/A(^c))</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>0.71 (0.58-0.87)</td>
</tr>
<tr>
<td>Yes</td>
<td>0.26 (0.20-0.33)</td>
</tr>
<tr>
<td>Social media use (reference: no)</td>
<td>0.37 (0.32-0.43)</td>
</tr>
<tr>
<td>Have regular health providers(^d) (reference: no)</td>
<td>0.36 (0.32-0.42)</td>
</tr>
<tr>
<td>Number of visits to health provider(^e) (reference: none)</td>
<td></td>
</tr>
<tr>
<td>1 time</td>
<td>0.24 (0.18-0.33)</td>
</tr>
<tr>
<td>2 times</td>
<td>0.19 (0.15-0.26)</td>
</tr>
<tr>
<td>3 times</td>
<td>0.15 (0.12-0.21)</td>
</tr>
<tr>
<td>4 times</td>
<td>0.14 (0.11-0.19)</td>
</tr>
<tr>
<td>5-9 times</td>
<td>0.10 (0.08-0.14)</td>
</tr>
<tr>
<td>≥10 times</td>
<td>0.10 (0.08-0.14)</td>
</tr>
<tr>
<td>Health provider maintain MR (reference: no)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0.11 (0.06-0.22)</td>
</tr>
<tr>
<td>Don’t know</td>
<td>1.14 (0.56-2.30)</td>
</tr>
<tr>
<td>Offered access to MR by health provider(^f) (reference: no)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0.03 (0.03-0.04)</td>
</tr>
<tr>
<td>Don’t know</td>
<td>0.83 (0.57-1.22)</td>
</tr>
<tr>
<td>Offered access to MR by health insurer(^f) (reference: no)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0.21 (0.19-0.25)</td>
</tr>
<tr>
<td>Don’t know</td>
<td>0.82 (0.71-0.95)</td>
</tr>
<tr>
<td>Strength training days per week (reference: none)</td>
<td></td>
</tr>
<tr>
<td>1-3 days per week</td>
<td>0.63 (0.56-0.71)</td>
</tr>
<tr>
<td>4-7 days per week</td>
<td>0.84 (0.70-1.01)</td>
</tr>
<tr>
<td>Ever tested colon cancer (reference: no)</td>
<td>0.74 (0.66-0.83)</td>
</tr>
<tr>
<td>Age</td>
<td>1.01 (1.00-1.01)</td>
</tr>
<tr>
<td>BMI</td>
<td>0.99 (0.99-1.00)</td>
</tr>
</tbody>
</table>

\(^a\)OR: odds ratio.

\(^b\)In general, how much would you trust information about cancer from a doctor/government health agencies/charitable organizations/religious organizations and leaders? (Supplement to the variable-related questions in the survey).

\(^c\)N/A: not applicable.

\(^d\)Not including psychiatrists and other mental health professionals, is there a particular doctor, nurse, or other health professional that you see most often? (Supplement to the variable-related questions in the survey).
In the past 12 months, not counting times you went to an emergency room, how many times did you go to a doctor, nurse, or other health professional to get care for yourself? (Supplement to the variable-related questions in the survey).

Have you ever been offered online access to your medical records by your health care provider/health insurer? (Supplement to the variable-related questions in the survey).

Table 1 shows significant predictors of nonuse of web-based medical records in LR ($P < .05$). The variables in LR were screened using 2 methods: backward selection and stepwise regression. The results obtained using the 2 variable selection methods were consistent, and 20 significant variables were finally selected.

The results of LR showed that sociodemographic indicators, such as age, BMI, education, marital status, income, and race, significantly affected the nonuse of web-based medical records, whereas sex and area had no significant effect on the prediction of nonuse of web-based medical records. On the basis of sociodemographic variables, people who were relatively older (OR 1.01, 95% CI 1.00-1.01), had a relatively lower BMI (OR 0.99, 95% CI 0.99-1.00), had relatively lower education (higher education OR 0.32, 95% CI 0.27-0.37), were not married (married OR 0.60, 95% CI 0.54-0.68), had a lower income (higher income OR 0.35, 95% CI 0.29-0.43), and belonged to racial and ethnic minority groups (OR 1.04, 95% CI 0.9-1.19) were more likely to not use web-based medical records. People who did not often access the internet or send and receive emails (OR 0.12, 95% CI 0.09-0.16), had not used a computer or smartphone to inquire about medical information in the past 12 months (OR 0.05, 95% CI 0.03-0.07), did not often use a wearable device to track health (OR 0.40, 95% CI 0.35-0.45), did not share health information from an electronic monitoring device or smartphone with a health professional in the previous 12 months, and people who had not used social media in the last 12 months (OR 0.37, 95% CI 0.32-0.43) were more inclined not to use web-based medical records. Moreover, those who were less concerned about their own health were more likely to not use web-based medical records. People who were more likely to not use web-based medical records tended to be those who did not see a particular doctor or health care professional frequently (OR 0.36, 95% CI 0.32-0.42); had not gone to a doctor, nurse, or other health professional to receive care in the last 12 months; did not have doctors or other health care providers maintain their medical records in a computerized system (OR 0.11, 95% CI 0.06-0.22); were not offered web-based access to their medical records by the health care provider (OR 0.03, 95% CI 0.03-0.04); were not offered web-based access to their medical records by the health insurer (OR 0.21, 95% CI 0.19-0.25); and did not perform leisure-time physical activities specifically designed to strengthen muscles. People who strongly trusted information about health or medical topics from religious organizations and leaders (OR 1.31, 95% CI 1.15-1.50), trusted information about health or medical topics from a doctor only a little (OR 0.73, 95% CI 0.53-0.99), and did not check for colon cancer (OR 0.74, 95% CI 0.66-0.83) were more inclined to not use web-based medical records.
Table 2. Predictors of web-based medical record (MR) nonuse built using automatic generalized linear model (auto-GLM), automatic random forest, auto–deep learning, and automatic gradient boosting machine (auto-GBM).

<table>
<thead>
<tr>
<th>Model and predictor</th>
<th>Importance scores</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Auto-GLM</strong></td>
<td></td>
</tr>
<tr>
<td>Offered access to MR by health provider$^a$</td>
<td>100</td>
</tr>
<tr>
<td>Electronic means use</td>
<td>47.41</td>
</tr>
<tr>
<td>Health provider maintain MR</td>
<td>22.72</td>
</tr>
<tr>
<td>Number of visits to health provider$^b$</td>
<td>21.63</td>
</tr>
<tr>
<td>Age</td>
<td>17.11</td>
</tr>
<tr>
<td>Offered access to MR by health insurer$^a$</td>
<td>14.94</td>
</tr>
<tr>
<td>Electronic wearable device use</td>
<td>13.91</td>
</tr>
<tr>
<td>Have regular health providers$^c$</td>
<td>13.10</td>
</tr>
<tr>
<td>Shared health information</td>
<td>12.06</td>
</tr>
<tr>
<td>Internet use</td>
<td>10.97</td>
</tr>
<tr>
<td>Ever tested colon cancer</td>
<td>9.64</td>
</tr>
<tr>
<td>Social media use</td>
<td>9.30</td>
</tr>
<tr>
<td>Income</td>
<td>6.45</td>
</tr>
<tr>
<td>Education</td>
<td>6.27</td>
</tr>
<tr>
<td>Race</td>
<td>6.23</td>
</tr>
<tr>
<td><strong>Automatic random forest</strong></td>
<td></td>
</tr>
<tr>
<td>Offered access to MR by health provider</td>
<td>100</td>
</tr>
<tr>
<td>Electronic means use</td>
<td>19.15</td>
</tr>
<tr>
<td>Offered access to MR by health insurer</td>
<td>19.02</td>
</tr>
<tr>
<td>Health provider maintain MR</td>
<td>16.03</td>
</tr>
<tr>
<td>Number of visits to health provider</td>
<td>10.20</td>
</tr>
<tr>
<td>Internet use</td>
<td>6.88</td>
</tr>
<tr>
<td>Have regular health providers</td>
<td>5.97</td>
</tr>
<tr>
<td>Age</td>
<td>5.43</td>
</tr>
<tr>
<td>Shared health information</td>
<td>4.71</td>
</tr>
<tr>
<td>Electronic wearable device use</td>
<td>4.58</td>
</tr>
<tr>
<td>Sitting time per day</td>
<td>4.29</td>
</tr>
<tr>
<td>BMI</td>
<td>4.15</td>
</tr>
<tr>
<td>Have electronic device</td>
<td>4.13</td>
</tr>
<tr>
<td>Education</td>
<td>2.96</td>
</tr>
<tr>
<td>Sitting time per day</td>
<td>2.83</td>
</tr>
<tr>
<td><strong>Auto–deep learning</strong></td>
<td></td>
</tr>
<tr>
<td>Offered access to MR by health provider</td>
<td>100</td>
</tr>
<tr>
<td>Electronic means use</td>
<td>51.18</td>
</tr>
<tr>
<td>Health provider maintain MR</td>
<td>40.19</td>
</tr>
<tr>
<td>Internet use</td>
<td>34.94</td>
</tr>
<tr>
<td>Offered access to MR by health insurer</td>
<td>34.41</td>
</tr>
<tr>
<td>Number of visits to health provider</td>
<td>33.24</td>
</tr>
<tr>
<td>Little interest$^d$</td>
<td>33.02</td>
</tr>
</tbody>
</table>
The essence of this study is a binary classification problem of judging whether individuals have used web-based medical records based on a set of inputs, such as education and income. Therefore, we evaluated the 5 methods using a series of evaluation metrics commonly used for classification algorithms. Table 3 presents the accuracy, precision, recall, $F_1$-score, and AUC values for the 5 machine learning methods on the validation and test sets. Accuracy is a metric of a classification model that measures the percentage of correct classification accounts for the total number of classifications. Precision is the proportion of correctly predicted positives to all predicted positives, whereas recall is the proportion of correctly predicted positives to all actual positives. The $F_1$-score is the harmonic mean of precision and recall. From the results of the verification set, LR had 3 indicators that performed best, with an accuracy of 83.35%, a recall of 88.97%, and an $F_1$-score of 83.59%.

However, the performance of LR on the test set was inferior to that of machine learning methods. In the test set, auto-GLM had the highest accuracy (82.49%), auto-GBM had the highest precision (79.73%), and automatic random forest had the highest recall (87.15%) and AUC (82.87%). AUC is not affected by the classification threshold and data distribution and, thus, reflects the overall classification power of the model. Automatic random forest had the highest AUC in both the validation (88.52%) and test (82.87%) sets. Therefore, in general, we were more inclined to choose the automatic random forest as the optimal model for predicting web-based medical record nonuse. Equations 3 to 6 provide the evaluation formulas for the machine learning models.

$$\text{Accuracy} = \frac{TP \ [True \ Positive] + TN \ [True \ Negative]}{TP + TN + FP \ [False \ Positive] + FN \ [False \ Negative]} \ (3)$$
AUCd, %  82.29  81.8  88.46  82.72  88.52  82.87  87.54  81.56  87.68  79.57

\(^a\)LR: logistic regression.  
\(^b\)Auto-GLM: automatic generalized linear model.  
\(^c\)Auto-GBM: automatic gradient boosting machine.  
\(^d\)AUC: area under the curve.

### Discussion

#### Principal Findings

Effective communication is key for delivering high-quality health care services [1]. Ensuring good communication by recording, processing, and sharing health information with patients is integral to the health care process. At present, the development of information and communications technology has allowed for the realization of web-based medical records, but because of some situations, web-based medical records are still not fully popular among patients. On the basis of demographic and individual behavioral characteristics, this study explored the predictors of web-based medical record nonuse.

We conducted a comprehensive assessment of the effects of individual sociodemographic characteristics, lifestyle, and behavioral habits on nonuse of web-based medical records. Although generalizing the factors that influence individuals’ nonuse of web-based medical records was difficult, based on the survey data, this study conducted in a data-rich variable selection environment demonstrated that an individual’s nonuse of web-based medical records is related to their sociodemographic characteristics, such as age, lifestyle, behavioral habits, and attention to health problems.

To date, numerous studies on the use of web-based medical records by patients have been based on surveys wherein data were gathered using a semistructured interview approach [17]. Using data from the National Cancer Institute 2019 to 2020 HINTS database, we applied the following question—“How many times did you access your web-based medical record in the last 12 months?”—to determine whether a person used their web-based medical records. The reason for avoiding web-based medical records is related to their nonuse. The use of electronic equipment such as mobile phones and computers, generally do not access web-based medical records, whereas people who are not accustomed to using electronic devices, such as smoking, electronic device use, and internet use, also significantly affected the nonuse of web-based medical records. Finally, individuals’ health status and their level of health concern were also associated with the nonuse of web-based medical records. Of course, some regions or units do not provide web-based access to medical records, or these are not maintained by health care providers, which may directly limit the use of web-based medical records.

On the basis of some of the conclusions of this study, recommendations can be made to promote the widespread use of web-based medical records. The use of electronic equipment is also a factor affecting the use of web-based medical records. People who are not accustomed to using electronic devices, such as mobile phones and computers, generally do not access their web-based medical records. The reason for avoiding web-based medical records may not be the disadvantage of web-based medical records itself but the resistance to electronic products, which is more common in older adults. However, older adults are more likely to become sick, so web-based medical records for this group are also a direction that needs special attention and development, such as building an interface statistical method, and the latter 4 algorithms are part of H2O’s automatic parameterization methods. A total of 29 influencing variables concerning the use of web-based medical records were selected based on coefficient significance in the LR model and the variable importance indicators in the other 4 methods. Many well-established determinants were also identified as proof of concept for our analytical approach, such as sociodemographic characteristics [9]. Although nonlinear and ensemble algorithms exhibit better predictive performance than traditional parametric models, they are less interpretable [18]. Therefore, predictors determined by such algorithms should be evaluated in conjunction with relevant research evidence.

This study showed that sociodemographic indicators, such as age, BMI, race, marital status, education, and income, significantly affected the nonuse of web-based medical records, whereas sex and area did not significantly affect the prediction of the nonuse of web-based medical records. In addition, predictors involving personal lifestyle and behavioral habits, such as smoking, electronic device use, and internet use, also played essential roles in predicting the nonuse of web-based medical records. Finally, individuals’ health status and their level of health concern were also associated with the nonuse of web-based medical records. Of course, some regions or units do not provide web-based access to medical records, or these are not maintained by health care providers, which may directly limit the use of web-based medical records.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>LRa Validation</th>
<th>Test</th>
<th>Auto-GLMb Validation</th>
<th>Test</th>
<th>Automatic random forest Validation</th>
<th>Test</th>
<th>Auto–deep learning Validation</th>
<th>Test</th>
<th>Auto-GBM Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy, %</td>
<td>83.35</td>
<td>81.47</td>
<td>82.17</td>
<td>82.49</td>
<td>82.48</td>
<td>82.38</td>
<td>82.57</td>
<td>80.93</td>
<td>81.32</td>
<td>79.69</td>
</tr>
<tr>
<td>Precision, %</td>
<td>78.86</td>
<td>76.26</td>
<td>85.09</td>
<td>77.73</td>
<td>86.1</td>
<td>76.43</td>
<td>89.52</td>
<td>74.3</td>
<td>85.38</td>
<td>79.73</td>
</tr>
<tr>
<td>Recall, %</td>
<td>88.97</td>
<td>86.7</td>
<td>79.76</td>
<td>84.81</td>
<td>79.04</td>
<td>87.15</td>
<td>75.01</td>
<td>87.15</td>
<td>77.78</td>
<td>76.96</td>
</tr>
<tr>
<td>( F_1 )-score, %</td>
<td>83.59</td>
<td>82.01</td>
<td>82.34</td>
<td>81.11</td>
<td>82.3</td>
<td>81.44</td>
<td>81.57</td>
<td>80.22</td>
<td>81.31</td>
<td>78.32</td>
</tr>
<tr>
<td>AUCd, %</td>
<td>82.29</td>
<td>81.8</td>
<td>88.46</td>
<td>82.72</td>
<td>88.52</td>
<td>82.87</td>
<td>87.54</td>
<td>81.56</td>
<td>87.68</td>
<td>79.57</td>
</tr>
</tbody>
</table>

\(^a\)LR: logistic regression.  
\(^b\)Auto-GLM: automatic generalized linear model.  
\(^c\)Auto-GBM: automatic gradient boosting machine.  
\(^d\)AUC: area under the curve.
that is friendly to the older adult population and keeping the internet-accessible interface simple and clear. Web-based medical records are often used by people with more related health problems. A study by the Office of the National Coordinator of Health Information Technology found that individuals may not realize the value of accessing their web-based medical records until they have a medical need. Given that the patient record request process can be time consuming, it may be more beneficial to have access to a person’s data in advance of an urgent health need. Therefore, popularizing health knowledge to the public and increasing the public’s attention to health information can increase the public’s demand for health-related information to a certain extent, thereby promoting the use of web-based medical records.

This research can provide a theoretical basis for predicting individual web-based medical record use. On the basis of the predictors of people not using web-based medical records selected by machine learning algorithms, individuals who do not use such records can be identified in advance, and use of web-based medical records can be promoted among them. Thus, this would provide more effective doctor-patient communication and better health care services.

Limitations
This study has some limitations. First, this study explored the influencing factors of nonuse of web-based medical records and discussed the correlation between each influencing factor and the target variable but did not involve the study of the influence path and influence mechanism. We considered conducting a causal analysis, but the cross-sectional survey design of the HINTS prevented us from making or testing causal claims. Second, the relationships between the selected variables were not studied further, and the screened important predictors may have certain collinearity. Third, the data were obtained using a self-report questionnaire. Therefore, we did not obtain detailed information on the nonuse of web-based medical records, and self-report bias may have affected the results. Finally, patient access to web-based medical records varies from country to country, and cultural background also has a strong impact on medical services. This study used HINTS data from the United States, and the conclusions may not be generalizable to other countries.

Comparison With Prior Work
The use of web-based medical records can enhance patient participation and co-operation in disease treatment and enhance doctor-patient communication to promote disease treatment. This is evidenced in the study by Stewart et al [19], whose research took patients with diabetes as the research object and found that patient portals support engagement by facilitating patient access to their health information and facilitating patient-provider communication. With the advancement of internet technology and the popularization of electronic products, the use of web-based medical records has become more convenient, but its penetration rate is still not high. Therefore, it has become a research hot spot to explore the characteristics and differences between users and nonusers of web-based medical records and identify the influencing factors of low use rate.

There are many studies based on HINTS data, such as that by Anthony et al [20], who used data from the 2017 HINTS to estimate 2 separate multivariable LR models to predict the factors associated with not having been offered access and those associated with not using a portal. On the basis of the 2017 to 2018 HINTS data, Patel and Johnson [11] used descriptive statistics and hypothesis testing to assess individuals’ access, viewing, and use of their web-based medical records and the use of smartphone health apps and other electronic devices. Trivedi et al [12] used multivariable LR analyses to examine the association between sociodemographic and health care–related factors on being offered access to web-based medical records and accessing web-based medical records and cited reasons for not accessing web-based medical records. Hong et al [21] used LR to investigate the trend of patient portal use in the general population and the barriers to adoption. The aforementioned studies are all based on traditional statistical methods such as LR [12,20,21] and hypothesis testing related to association analysis [11], and the selection of influencing factors and the conclusions drawn only involve demographic variables and some variables of interest, whereas our research combines traditional statistical methods and machine learning methods to find predictors of web-based medical record use based on variable-rich environments. Furthermore, the results of machine learning methods also provide variable importance scores and rankings, which have also not been covered in previous studies. To the best of our knowledge, this is the first study to apply a range of H2O’s automatic machine learning algorithms to a nationally representative sample for optimizing the classification of web-based medical record nonuse. Compared with previous studies, we found that personal lifestyle and behavioral habits as well as individuals’ health status and their level of health concern significantly affect web-based medical record nonuse.

Conclusions
Using data from the National Cancer Institute 2019 to 2020 HINTS database, this study applied 5 machine learning algorithms—LR (linear), auto-GLM, automatic random forest, auto–deep learning, and auto-GBM—to identify and investigate the factors that affect whether individuals use web-based medical records. Using these 5 models, 29 variables were identified as crucial predictors of nonuse of web-based medical records. When monitoring web-based medical record use trends, research should consider social factors such as age, education, BMI, and marital status, as well as personal lifestyle and behavioral habits, including smoking, use of electronic devices and the internet, patients’ personal health status, and their level of health concern. The use of electronic medical records can be targeted to specific patient groups, allowing more people to benefit from their usefulness.

The main contributions of this study are as follows: (1) using authoritative data, potential predictors were selected based on a data-rich environment, involving more comprehensive variables and avoiding unnecessary subjectivity, and (2) the key parameters of the machine learning methods considerably influenced the accuracy of the model. In this study, H2O’s automatic parameter selection methods were introduced to optimize the key parameters of the model. Compared with
traditional machine learning algorithms, the H2O automatic machine learning methods effectively improved model performance.

Acknowledgments
This research was funded by the Youth Project of National Social Science Fund of China (grant 21CTJ008).

Conflicts of Interest
None declared.

Multimedia Appendix 1
Variables extracted from the Health Information National Trends Survey database for research.

Multimedia Appendix 2
Distribution of characteristics of variables in the Health Information National Trends Survey database (N=9072).

References


**Abbreviations**

- **AUC**: area under the curve
- **auto-GBM**: automatic gradient boosting machine
- **auto-GLM**: automatic generalized linear model
- **GBM**: gradient boosting machine
- **GLM**: generalized linear model
- **HINTS**: Health Information National Trends Survey
- **LR**: logistic regression
- **OR**: odds ratio

©Yang Chen, Xuejiao Liu, Lei Gao, Miao Zhu, Ben-Chang Shia, Mingchih Chen, Linglong Ye, Lei Qin. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.06.2023. 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.
Identifying Risk Factors Associated With Lower Back Pain in Electronic Medical Record Free Text: Deep Learning Approach Using Clinical Note Annotations

Aman Jaiswal¹, BTech; Alan Katz², MBChB, MSc; Marcello Nesca², BCom, BA, MSc; Evangelos Milios¹, EE, SM, PhD

¹

Corresponding Author:
Alan Katz, MBChB, MSc

Abstract

Background: Lower back pain is a common weakening condition that affects a large population. It is a leading cause of disability and lost productivity, and the associated medical costs and lost wages place a substantial burden on individuals and society. Recent advances in artificial intelligence and natural language processing have opened new opportunities for the identification and management of risk factors for lower back pain. In this paper, we propose and train a deep learning model on a data set of clinical notes that have been annotated with relevant risk factors, and we evaluate the model’s performance in identifying risk factors in new clinical notes.

Objective: The primary objective is to develop a novel deep learning approach to detect risk factors for underlying disease in patients presenting with lower back pain in clinical encounter notes. The secondary objective is to propose solutions to potential challenges of using deep learning and natural language processing techniques for identifying risk factors in electronic medical record free text and make practical recommendations for future research in this area.

Methods: We manually annotated clinical notes for the presence of six risk factors for severe underlying disease in patients presenting with lower back pain. Data were highly imbalanced, with only 12% (n=296) of the annotated notes having at least one risk factor. To address imbalanced data, a combination of semantic textual similarity and regular expressions was used to further capture notes for annotation. Further analysis was conducted to study the impact of downsampling, binary formulation of multi-label classification, and unsupervised pretraining on classification performance.

Results: Of 2749 labeled clinical notes, 347 exhibited at least one risk factor, while 2402 exhibited none. The initial analysis shows that downsampling the training set to equalize the ratio of clinical notes with and without risk factors improved the macro–area under the receiver operating characteristic curve (AUROC) by 2%. The Bidirectional Encoder Representations from Transformers (BERT) model improved the macro-AUROC by 15% over the traditional machine learning baseline. In experiment 2, the proposed BERT–convolutional neural network (CNN) model for longer texts improved (4% macro-AUROC) over the BERT baseline, and the multitask models are more stable for minority classes. In experiment 3, domain adaptation of BERT-CNN using masked language modeling improved the macro-AUROC by 2%.

Conclusions: Primary care clinical notes are likely to require manipulation to perform meaningful free-text analysis. The application of BERT models for multi-label classification on downsampled annotated clinical notes is useful in detecting risk factors suggesting an indication for imaging for patients with lower back pain.

(JMIR Med Inform 2023;11:e45105) doi:10.2196/45105

KEYWORDS

machine learning; lower back pain; natural language processing; semantic textual similarity; electronic medical records; risk factors; deep learning

Introduction

Lower back pain (LBP) is recognized as a common disability worldwide [1-3]. While there is no agreed-upon definition of LBP, in a systematic review, it was primarily defined through routinely collected electronic health data, which include International Classification of Diseases, Ninth Revision (ICD-9) and International Statistical Classification of Diseases, Tenth Revision (ICD-10) codes [4]. One estimate of the burden of LBP is that 13% of adults in the United States live with LBP, while in Canada, among those living with chronic pain, 50.9% identified the location of their pain in the upper or lower back.
In a systematic review [4], the mean prevalence of LBP among the studies collected ranged between 1.4% and 15.6%. While the burden of LBP remains high, it is important to understand the indicators for possible serious underlying causes that require imaging, also known as “risk factors” [5]. According to Choosing Wisely Canada, risk factors may include [6]:

- A history of cancer
- Unexplained weight loss
- A recent infection
- Fever
- Loss of bowel or bladder control
- Abnormal reflexes or the loss of muscle power in the legs

Radiological (diagnostic) imaging includes procedures such as x-rays, computed tomography scans, or magnetic resonance imaging scans. Recommendations from clinical practice guidelines state that, unless risk factors are present, radiological imaging is not needed for patients with LBP [5,7]. Moreover, ordering radiological imaging when it is unnecessary puts the patient at risk for radiation exposure and other negative consequences [5,6]. Despite these recommendations, patients with LBP are frequently subjected to unnecessary imaging [8].

The data for this study in clinical practice uses electronic medical records (EMRs). The widespread use of this IT has introduced the feasibility of analyzing large numbers of clinical notes without having to manually access paper charts and perform the analyses using automated approaches such as natural language processing (NLP) [9]. The Canadian Primary Care Sentinel Surveillance Network [10] routinely extracts clinical information such as clinical encounter notes, note type, and the date of the notes from primary care clinical practices with the permission of the providers. Applying NLP methods to EMR data makes it possible to detect LBP risk factors and understand the use of imaging in this common clinical presentation.

Since the introduction of transformers in 2019 [11], which are deep language models that can be fine-tuned for specific tasks, deep language models have achieved a significant milestone in natural language understanding. The transfer learning paradigm of unsupervised pretraining and fine-tuning [12] using Bidirectional Encoder Representations from Transformers (BERT) has reduced the requirement for large labeled data sets to achieve state-of-the-art analytic performance. Previous research [13] has explored the use of topic models and deep neural networks to automatically distinguish acute LBP episodes using free-text clinical notes.

Methods

The following steps were undertaken to achieve our goal: preparation of EMR data, EMR annotation process, addressing imbalanced data, and application of the proposed model.

Preparation of EMR Data

We accessed a random sample of deidentified EMR data, and using the regular expressions created in SAS (SAS Institute), we identified a cohort of patients with any indication of LBP. Notes were further filtered by note type to only include provider-generated clinical notes. The data were then split randomly into three files. Ethics approval for the study was provided by the University of Manitoba Health Research Ethics Board and the Health Information Privacy Committee.

EMR Annotation Process

Six medical students reviewed the EMR notes to identify the six LBP risk factors in accordance to Choosing Wisely Canada. They worked in teams of two to validate the application of the inclusion and exclusion criteria, each note being annotated by two students. The inclusion criteria listed in Textbox 1 were the presence of specific clinical notes suggestive of at least one of the six risk factors indicating the need for imaging. The exclusion criteria were the presence of clinical conditions that could lead to symptoms that may be confused with any of the underlying conditions represented by the six risk factors and clinical notes that do not represent relevant visits.

[2,3]
Inclusion and exclusion criteria for risk factors.

**Inclusion criteria**
- Lower extremities for loss of muscle function
- Positive straight leg test
- Nerve impingement
- Sciatica, but need to confirm radiculopathy
- Incontinence related to a nerve issue
- If back pain has improved
- Follow-up discussions of imaging results
- Saddle anesthesia
- Notes that do not specify upper vs lower back pain

**Exclusion criteria**
- HIV is not a relevant infection (regardless of viral load and strain/location)
- Urinary symptoms other than incontinence are neither risk factors nor symptoms of relevant infection
- Shingles as an infection if it is a lumbar dermatome
- Nocturnal enuresis
- Degenerative diseases or osteoarthritis with an indication of back pain
- Copy/pasted imaging results onto the electronic medical record note
- Notes that mention previous or resolved back pain
- Well child/adolescent visit

An experienced clinician (AK) arbitrated any disagreements between student annotators. This supported the inclusion of correctly labeled records in the classification model. For the annotation process, we used Microsoft Forms (Microsoft Corporation), which enabled us to collect the relevant data in a systematic and organized manner. Specifically, the output from Microsoft Forms was linked to a secure CSV file containing the clinical notes, using a unique identifier to facilitate data merging and subsequent analysis.

**Addressing Imbalanced Data**

Our data collection process consisted of two rounds. In the first round, we established the initial distribution of risk factors. Analysis of this round revealed an imbalanced distribution of labels, a well-known factor that can impact the performance of deep learning methods [14,15]. Specifically, we observed an imbalance in both the infrequent occurrence of individual risk factors and the high frequency of the “null class,” which denotes the absence of risk factors.

To address this imbalance, we adopted a 2-pronged approach. First, we collected additional clinical notes specifically targeting minority risk factors. Second, we downsampled the majority of notes with “null class.” Notably, the initial data set lacked any clinical notes for unexplained weight loss. Table 1 depicts the distribution of risk factors after the first labeling round, revealing that only 12% (n=296) of the 2487 annotated notes exhibited any risk factors.

<table>
<thead>
<tr>
<th>Risk factors</th>
<th>Annotations (round 1), n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>26</td>
</tr>
<tr>
<td>Weight</td>
<td>0</td>
</tr>
<tr>
<td>Fever</td>
<td>8</td>
</tr>
<tr>
<td>Infection</td>
<td>8</td>
</tr>
<tr>
<td>Bowel</td>
<td>9</td>
</tr>
<tr>
<td>Abreflex</td>
<td>233</td>
</tr>
</tbody>
</table>

*Zero notes exhibit the unexplained weight loss risk factor.
Acquiring More Notes to Annotate

Prior studies have explored methods for addressing the challenge of obtaining sufficient data for training [16]. To acquire clinical notes for labeling that are more likely to exhibit a minority risk factor, we used unsupervised semantic textual similarity (STS). It is a ranking task where given a text query and a list of clinical notes, the STS model ranks the clinical notes that are semantically like the query. We trained two unsupervised STS models, Transformers and Sequential Denoising Auto-Encoder (TSDAE) [17] and Simple Contrastive Learning of Sentence Embeddings (SimCSE) [18], implemented using the SentenceTransformer Python library [19]. To rank the unlabeled clinical notes (ie, 55,000 notes with any LBP indication), we formed the queries using rationales, collected as part of the first labeling round. Here, we refer to “rationale” as an extracted snippet or text from the clinical note the annotators highlighted as evidence for a risk factor.

Figure 1 illustrates the STS sampling process with numbered steps. First, we group the clinical notes based on the exhibited risk factors. We then concatenate the rationales for each group of clinical notes to form queries and rank the unlabeled clinical notes using the unsupervised STS models. If the rationales were unavailable from the first labeling round (eg, “weight loss”), we used risk factor definition or custom text as the query. We selected the top K notes from the ranked clinical notes, where “K” is set within the 10-50 range. We further filtered noisy outputs using phrases such as “has fever,” “has back pain,” and “lost weight.” Finally, we iterated the process for each risk factor and provided the selected notes for the second labeling round.
This approach helped maximize annotations for clinical notes that exhibited risk factors. Table 2 depicts the complete distribution of risk factors after both rounds of labeling. Of the 262 annotated clinical notes in the second round, 19.5% (n=51) of the clinical notes exhibited risk factors, in contrast to 12% (n=296) in the first round.
Table. Risk factor distribution after both rounds of labeling.\(^a\)

<table>
<thead>
<tr>
<th>Risk factors</th>
<th>Annotations (round 1 + 2), n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>53</td>
</tr>
<tr>
<td>Weight</td>
<td>32</td>
</tr>
<tr>
<td>Fever</td>
<td>17</td>
</tr>
<tr>
<td>Infection</td>
<td>9</td>
</tr>
<tr>
<td>Bowel</td>
<td>9</td>
</tr>
<tr>
<td>Abreflex</td>
<td>236</td>
</tr>
</tbody>
</table>

\(^a\)This includes 2487 notes from the first round and 262 notes from the second round. In the second labeling round, we collected 32 clinical notes for the unexplained weight loss risk factor.

Treating Class Imbalance With Downsampling

Following the second round of labeling, a significant class imbalance was observed in the resulting distribution of labels. Specifically, out of the total 2749 annotated clinical notes, only 347 were labeled as having one or more risk factors, while the remaining 2402 notes were labeled with no risk factor. To mitigate this issue, two common approaches are oversampling the minority class or downsampling the majority class. In a multi-label data set, each instance can be assigned to one or more classes. For instance, in the case of clinical notes, they may have one or more risk factors, making it challenging to oversample the minority class. This is because generating synthetic instances requires randomly selecting a minority clinical note that may have a combination of labels rather than a single label. However, this approach may bias the model toward the minority class and lead to overfitting. Consequently, we opted for downsampling the majority class to balance the class distribution and prevent the model from being biased toward the majority class.

Specifically, a subset of the clinical notes with “no risk factors” was randomly selected to match the number of clinical notes with “any risk factor.” This approach aimed to balance the class distribution and enable the model to learn from both positive and negative examples. To assess the effectiveness of the downsampling strategy, we conducted a comparative analysis of the model’s performance with and without downsampling.

Application of Proposed Model

Transformer-based BERT [11] models can be fine-tuned for detecting risk factors in clinical notes using a small labeled data set. The requirement for large labeled data sets is eased with models that are pretrained on large clinical text. In this work, we used BlueBERT [20] as our back-end model that is pretrained on PubMed abstracts and clinical notes from the Medical Information Mart for Intensive Care (MIMIC-III) data set [21]. However, BERT models are limited to a maximum input length of 512 tokens. The length of clinical notes in our data set ranges from 7 to 1400 tokens with 8% (n=221) of the notes having more than 512 tokens. To overcome this limitation, we propose a novel architecture called BERT–convolutional neural network (CNN) that chunks the inputs and processes them using convolution layers. The proposed chunking method is illustrated in Figure 2.
Figure 2. BERT input chunking: a clinical note is first separated into chunks of 512 tokens. Each chunk is then independently processed by the BERT-based back-end model. The chunk embedding is obtained by averaging the token embeddings from the last five layers of BERT. Finally, all the chunk embeddings are concatenated and processed using convolution layers, as defined by Kim [22]. Note: The sample clinical note does not belong to the real data set. BERT: Bidirectional Encoder Representations from Transformers.

Experimental Setup

The study used a repeated 2-fold cross-validation approach with two repetitions to improve the estimated performance of the machine learning models. As the data set was multi-label, we adopted the iterative stratification method [23,24] provided by the scikit-multilearn library [25] to generate stratified splits for the folds. This ensured that every split had a similar distribution of risk factors. The 2-fold cross-validation was repeated twice, resulting in a total of four runs. Wherever applicable, we implemented the downsampling technique (as described earlier) on the training set. Our results are reported in terms of the area under the receiver operating characteristic curve (AUROC) of individual risk factors and their macroaverage across the folds. Table 3 reports the frequency of positive risk factors in each split of the folds.
Table. Frequency of positive risk factors in train-test splits. We report the approximate counts of each risk factor across folds. Note: the counts do not include the clinical notes with no risk factors, which are approximately 1198 and 1195 for the train and test split, respectively.

<table>
<thead>
<tr>
<th>Positive risk factors</th>
<th>Train split (n=1374 notes), n</th>
<th>Test split (n=1375 notes), n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td>Weight</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Fever</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Infection</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Bowel</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Abreflex</td>
<td>118</td>
<td>118</td>
</tr>
</tbody>
</table>

Ethics Approval
The study received ethics approval from the Health Research Ethics Board of the University of Manitoba (study number HS20263; review number H2016:408).

Results
Overview
In this section, we report the analysis of the data collection and classification performance of the transformer-based models with different configurations, including traditional machine learning and BERT-based baselines. The transformer-based models were trained for 10 epochs each, with a learning rate ranging from 5e-05 to 6e-5. Unless specified otherwise, all the BERT-based models use BlueBERT [20] as the back end.

Data Collection Analysis
Each annotation was added to the clinical note level independently. These notes are associated with patient- and site-level information, allowing for further analysis based on the patient and site as the unit of analysis. Table 4 presents an analysis of the LBP characteristics reported in the collected data, using notes, patient, and site ID as the units of analysis. This enables a multilevel analysis of the reported characteristics, providing a detailed understanding of their distribution across various units of analysis.

Table. Lower back pain characteristics gathered from collected data, with notes, patient, and site ID each serving as the units of analysis.

<table>
<thead>
<tr>
<th>Unit of analysis</th>
<th>Values, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notes (N=2749)</td>
<td></td>
</tr>
<tr>
<td>History of cancer</td>
<td>53 (1.9)</td>
</tr>
<tr>
<td>Signs of fever</td>
<td>17 (0.6)</td>
</tr>
<tr>
<td>Unexplained weight loss</td>
<td>32 (1.2)</td>
</tr>
<tr>
<td>Recent infection</td>
<td>9 (0.3)</td>
</tr>
<tr>
<td>Loss of bowel or bladder control</td>
<td>9 (0.3)</td>
</tr>
<tr>
<td>Abnormal reflexes</td>
<td>236 (8.6)</td>
</tr>
<tr>
<td>Patients (N=1943)</td>
<td></td>
</tr>
<tr>
<td>History of cancer</td>
<td>40 (2.1)</td>
</tr>
<tr>
<td>Signs of fever</td>
<td>17 (0.9)</td>
</tr>
<tr>
<td>Unexplained weight loss</td>
<td>32 (1.6)</td>
</tr>
<tr>
<td>Recent infection</td>
<td>9 (0.5)</td>
</tr>
<tr>
<td>Loss of bowel or bladder control</td>
<td>8 (0.4)</td>
</tr>
<tr>
<td>Abnormal reflexes</td>
<td>201 (10.3)</td>
</tr>
<tr>
<td>Site ID (N=22)</td>
<td></td>
</tr>
<tr>
<td>History of cancer</td>
<td>12 (55)</td>
</tr>
<tr>
<td>Signs of fever</td>
<td>11 (50)</td>
</tr>
<tr>
<td>Unexplained weight loss</td>
<td>12 (55)</td>
</tr>
<tr>
<td>Recent infection</td>
<td>5 (23)</td>
</tr>
<tr>
<td>Loss of bowel or bladder control</td>
<td>7 (32)</td>
</tr>
<tr>
<td>Abnormal reflexes</td>
<td>13 (59)</td>
</tr>
</tbody>
</table>
A total of 2749 clinical notes were annotated to collect information on risk factors for LBP. The most reported risk factor was “abnormal reflexes,” with 236 annotations, followed by “history of cancer” with 53 annotations. Out of the 1943 patients covered by the annotation process, only 40 were labeled with a “history of cancer,” accounting for 2.1% (n=40) of the total patients. More than 10% of patients were reported with “abnormal reflexes,” while “recent infection” and “loss of bowel control” were reported in only 9 and 8 patients, respectively.

The analysis of clinical sites associated with the clinical notes revealed that 12 of 22 sites reported at least two risk factors, with “recent infection” and “loss of bowel or bladder control” being the least commonly reported risk factors, mentioned in only 5 and 7 clinical sites, respectively. These findings indicate that “abnormal reflexes” is the most reported characteristic of LBP across all units of analysis, with “history of cancer,” “unexplained weight loss,” and “signs of fever” being reported less frequently. The frequency of “loss of bowel or bladder control” and “recent infection” was relatively low across all units of analysis, indicating that these characteristics may not be as common as others in cases of LBP. The distribution of these characteristics varies across different units of analysis, which highlights the importance of examining LBP characteristics at multiple levels.

**Performance With and Without Downsampling**

In our initial analysis, we compared the impact of downsampling the training set, as described earlier, on the average and label-wise performance of the models. Figure 3 displays the results of this comparison. We also included a tf-idf (term frequency–inverse document frequency) + logistic regression model trained with a multi-output classifier [26] as a baseline, which was the best-performing baseline (among 7 candidates, including k-nearest neighbors, naive Bayes, random forest, and models from the scikit-multilearn Library [25]). On average, the BERT models performed 15% better than the baseline. Downsampling the training set improved performance by 2% for BERT-Multi models and reduced the SD as reflected by the error bars for minority labels (eg, “bowel” and “fever”). Downsampling of the majority class (ie, “No Risk factor notes”) also helped stabilize the performance of the models, as indicated by the smaller error bars. We used the downsampled training set for further analysis.
Performance With BERTCNN and Independent Binary Classifiers

Using the downsampled training set for all the models, we compared the performance of four different models chosen by architecture (BERT, BERTCNN) and task formulation (multitask learning, binary classification). Figure 4 shows the results. The comparison of BERT and BERTCNN highlights the importance of not truncating longer inputs. The comparison of the proposed model (BERTCNN) with their binary variants helps in understanding the trade-off between parameter efficiency and performance. The average AUROC of all the models are comparable, with BERTCNN-Multi performing 4% better than BERT-Multi. The multitask BERT and BERTCNN models match the performance of their binary alternative with six times fewer parameters. When sufficient positive samples are present for a risk factor (eg, abreflex), all the models perform comparably with a low SD. When the samples are insufficient (eg, “infection” and “bowel”), the binary models have high SD (indicated by the error bars), as few-samples BERT fine-tuning is known to be unstable [27]. In such cases, the multitask models generally produce more stable results, with the BERTCNN-Multi performing 9% better than BERT-Multi. In general, the BERTCNN model can benefit from the extra context...
found in the complete clinical note to improve prediction performance.

**Figure 4.** BERT-Multi, BERT-Binary, BERTCNN-Multi, and BERTCNN-Binary trained on the downsampled training data. The AUROC for each risk factor and their macroaverage are reported, with the SDs reflected in the error bars. AUROC: area under the receiver operating characteristic curve; BERT: Bidirectional Encoder Representations from Transformers; BERTCNN: Bidirectional Encoder Representations from Transformers–convolutional neural network.

**Performance With Domain Adaptation Using Unsupervised Training**

The best-performing model can further benefit from pretraining [28] the underlying transformer model using the clinical notes. In this analysis, we investigate the effect of domain adaptation using pretraining on classification performance. We used BERTCNN and further pretrained the back-end model (BlueBERT [20]) with the complete corpus of relevant clinical notes (N=57,000) for 3 epochs. Two choices for pretraining the BERT architecture were considered: masked language modeling (MLM; BERTCNN-MLM-Multi) [12] and causal language modeling (CLM; BERTCNN-CLM-Multi) [29]. In addition, we also report results of the recent transformers-based model for long text in the clinical domain, called clinical-longformer [30,31], which was pretrained on clinical notes from the MIMIC-III data set [21]. Our results, shown in Figure 5, indicate that the MLM method performed 2% better than no domain adaptation and improved the performance for “cancer” by 5%. The longformer model further improves performance over MLM
by 2%. It is worth noting that while the performance improvement of domain adaptation using MLM [32] is not significant, it is comparable to that of the already pretrained BlueBERT [20] and clinical-longformer [30,31], which were pretrained on a much larger corpus of over 2 million notes.

**Figure 5.** Effect of domain adaptation using MLM, CLM, and comparison with the clinical-longformer model. The AUROC for each risk factor and their macroaverage are reported, with the SDs reflected in the error bars. AUROC: area under the receiver operating characteristic curve; BERTCNN: Bidirectional Encoder Representations from Transformers–convolutional neural network; CLM: causal language modeling; MLM: masked language modeling.

**Discussion**

The analysis of electronic clinical notes using machine learning techniques provides the opportunity to explore and evaluate clinical care, previously not possible when clinical experts had to read each clinical record. NLP of clinical records is still a relatively new research endeavor that is rapidly evolving. This study encountered and addressed several challenges that are likely to be common in the analysis of clinical notes. For example, the initially collected data were imbalanced, with most clinical notes having no risk factors for cancer, suggesting the need for further investigation of LBP. By sampling additional clinical notes from the unlabeled pool using unsupervised semantic matching techniques for a limited second round of labeling, we captured 7.5% more clinical notes with at least one risk factor. Strategic resampling can decrease bias in multi-label
data sets, which substantially helps in classification performance. The analysis comparing multitask learning and binary classification suggests we can match the performance of independent binary classifiers and produce more stable results while using a fraction of the learned parameters required for binary classifiers. This study demonstrates the value of domain adaptation as an additional technique to improve the classification results of transformer-based models and improve clinical free-text classification using unsupervised methods.

A strength of this study is the comparison of different models and approaches using a random sample of real clinical notes. We compared the BERT-based model, which does not truncate longer clinical notes and uses the complete context to make predictions, to the more commonly used truncated note model. The extensive empirical analysis on the impact of different modeling choices, including comparisons of multitask and single-task learning, resampling of data, and domain adaptation using unsupervised methods for the detection of LBP risk factors in clinical notes, provides guidance for future analysis of clinical text data.

While the low number of samples for certain risk factors in the test set is a limitation, this was addressed in reporting the AUROC for each individual risk factor, including their macroaverage for each model, and using the repeated k-fold cross-validation approach for better estimation of performance. Future research will involve linking the outcomes of imaging studies to the identification of risk factors in this data set. It is anticipated that patients without risk factors would have normal imaging, while those with risk factors should be more likely to have abnormal imaging suggestive of disease requiring further treatment. These analyses will need to address the imbalance in the data, as a minority of patients have undergone imaging.

Deep learning models, specifically BERT-based models, are suitable for capturing and detecting risk factors for LBP in clinical notes. Semantic matching techniques are effective during data collection in providing minority samples for labeling and improving data set distribution. The proposed method BERTCNN can be successfully applied for clinical notes that may be longer than the input limit of BERT-based models. Detecting risk factors in clinical notes is better formulated as multitask learning, which is more efficient and provides stable results. Furthermore, transformer-based models are successfully adopted for clinical text using transfer learning and MLM.

Acknowledgments
The authors acknowledge the clinicians and patients whose data were accessed for this study through the Manitoba Primary Care Research Network, a node of the Canadian Sentinel Surveillance Network. The authors also thank medical student annotators Elvina Mukhamedshina, Gem Newman, JaeYeon Park, Mehrin Ahmed, Sue Zhang, and Will Siemens.

Conflicts of Interest
None declared.

References


Abbreviations

AUROC: area under the receiver operating characteristic curve
BERT: Bidirectional Encoder Representations from Transformers
BERTCNN: Bidirectional Encoder Representations from Transformers–convolutional neural network
CLM: causal language modeling
EMR: electronic medical record
ICD-10: International Statistical Classification of Diseases, Tenth Revision
ICD-9: International Classification of Diseases, Ninth Revision
LBP: lower back pain
MIMIC-III: Medical Information Mart for Intensive Care
MLM: masked language modeling
NLP: natural language processing
SimCSE: Simple Contrastive Learning of Sentence Embeddings
STS: semantic textual similarity
tf-idf: term frequency–inverse document frequency
TSDAE: Transformers and Sequential Denoising Auto-Encoder
Near Real-time Natural Language Processing for the Extraction of Abdominal Aortic Aneurysm Diagnoses From Radiology Reports: Algorithm Development and Validation Study

Simon Gaviria-Valencia, MD; Sean P Murphy, BS; Vinod C Kaggal, MS; Robert D McBane II, MD; Thom W Rooke, MD; Rajeev Chaudhry, MBBS, MPH; Mateo Alzate-Aguirre, MD; Adelaide M Arruda-Olson, MD, PhD

1Divisions of Preventive Cardiology and Cardiovascular Ultrasound, Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, United States
2Advanced Analytics Services Unit (Natural Language Processing), Department of Information Technology, Mayo Clinic, Rochester, MN, United States
3Enterprise Technology Services (Natural Language Processing), Department of Information Technology, Mayo Clinic, Rochester, MN, United States
4Gonda Vascular Center, Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, United States
5Department of Internal Medicine, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Adelaide M Arruda-Olson, MD, PhD
Divisions of Preventive Cardiology and Cardiovascular Ultrasound
Department of Cardiovascular Medicine
Mayo Clinic
200 First Street SW
Rochester, MN, 55905
United States
Phone: 1 507 266 6853
Email: arrudaolson.adelaide@mayo.edu

Abstract

Background: Management of abdominal aortic aneurysms (AAAs) requires serial imaging surveillance to evaluate the aneurysm dimension. Natural language processing (NLP) has been previously developed to retrospectively identify patients with AAA from electronic health records (EHRs). However, there are no reported studies that use NLP to identify patients with AAA in near real-time from radiology reports.

Objective: This study aims to develop and validate a rule-based NLP algorithm for near real-time automatic extraction of AAA diagnosis from radiology reports for case identification.

Methods: The AAA-NLP algorithm was developed and deployed to an EHR big data infrastructure for near real-time processing of radiology reports from May 1, 2019, to September 2020. NLP extracted named entities for AAA case identification and classified subjects as cases and controls. The reference standard to assess algorithm performance was a manual review of processed radiology reports by trained physicians following standardized criteria. Reviewers were blinded to the diagnosis of each subject. The AAA-NLP algorithm was refined in 3 successive iterations. For each iteration, the AAA-NLP algorithm was modified based on performance compared to the reference standard.

Results: A total of 360 reports were reviewed, of which 120 radiology reports were randomly selected for each iteration. At each iteration, the AAA-NLP algorithm performance improved. The algorithm identified AAA cases in near real-time with high positive predictive value (0.98), sensitivity (0.95), specificity (0.98), F1 score (0.97), and accuracy (0.97).

Conclusions: Implementation of NLP for accurate identification of AAA cases from radiology reports with high performance in near real time is feasible. This NLP technique will support automated input for patient care and clinical decision support tools for the management of patients with AAA.

(JMIR Med Inform 2023;11:e40964) doi:10.2196/40964
KEYWORDS
abdominal aortic aneurysm; algorithm; big data; electronic health record; medical records; natural language processing; radiology reports; radiology

Introduction
Worldwide prevalence rates of abdominal aortic aneurysms (AAAs) range from 1.6% to 3.3% for men older than 60 years [1]. Assessment of AAA may be performed by a variety of imaging tests, including ultrasound (US), computerized tomography (CT), and magnetic resonance imaging (MRI). In the United States, the prevalence of AAA has been reported as 2.8% among 9457 individuals screened by US [2]. Moreover, screening for early identification decreases the risk of aneurysm-related death and morbidity [1,3]. A prior study has shown that 4.5 ruptured AAA per 10,000 person-years were likely to have been prevented by screening, with an estimated 54 life-years gained per year of screening in a population of 23,000 men at risk [4].

The interpretation of imaging examinations is routinely reported in radiology reports as narrative text in electronic health records (EHRs) [5]. The automated extraction of information from narrative text can be accomplished by natural language processing (NLP) [6-8]. Prior studies have demonstrated high accuracy, sensitivity, specificity, and positive predictive value (PPV) of NLP for extraction of clinical concepts from narrative text in radiology reports [9-12]. Moreover, NLP is useful in cohort ascertainment for epidemiologic studies, query-based case retrieval, clinical decision support (CDS), quality assessment of radiologic practices, and diagnostic surveillance [5].

A previous retrospective cohort study from our institution developed a rule-based NLP algorithm for retrospective retrieval of AAA cases from radiology reports, which performed with high accuracy [12]. However, to the best of our knowledge, no prior study has demonstrated the use of NLP to identify AAA cases from radiology reports processed in near real-time. Hence, we tested the hypothesis that a rule-based NLP algorithm will extract AAA diagnosis from radiology reports in near real-time with high accuracy.

Methods

Study Settings
This study used Mayo Clinic radiology reports from May 1, 2019, to September 30, 2020.

Study Design
A rule-based AAA-NLP algorithm was developed for information extraction of AAA diagnosis automatically from radiology reports, including CT abdomen pelvis without intravenous (IV) contrast, CT chest abdomen pelvis angiogram with IV contrast, US abdomen complete, US aorta iliac arteries bilateral with doppler, MRI abdomen with and without IV contrast, and MRI pelvis with and without IV contrast. The rule-based NLP algorithm was developed using MedTagger and deployed in the institutional near real-time big data infrastructure to process relevant radiology reports. MedTagger is an open-source NLP tool that has been previously used in various clinical NLP applications [13]. MedTagger enables section identification, extraction of concepts, sentences, and word tokenization [14,15]. The AAA-NLP algorithm had 2 main components composed of text processing and report classification. AAA-relevant concepts were used to classify all reports (Figure 1).

A custom lexicon for AAA was identified by the study team through a manual review of radiology reports. Subsequently, this lexicon was mapped to corresponding concepts and their synonyms in the Unified Medical Language System Metathesaurus. The lexicon used for AAA identification included aorta abdominal aneurysm, aortic aneurysm abdominal, AAA, aneurysm abdominal aorta, and infrarenal aortic aneurysm. Each radiology report was then processed in near real-time by NLP. The AAA-NLP algorithm extracted both the lexicon and the contextual information of assertions, including negations or confirmations, from each radiology report. Textbox 1 displays the rules used by the NLP algorithm. The AAA-NLP algorithm classified subjects as AAA cases and controls without AAA.

To enable validation, the NLP output generated by near real-time processing of radiology reports was retrieved from the digital infrastructure by the information technology team and converted to a human-readable format for annotation. This annotation was performed by 2 trained physicians following written guidelines for standardization. The annotators were blinded to the diagnosis of each subject and to the results of the other annotator. In the written guidelines, AAA was defined as an aortic aneurysm diameter ≥3 cm by imaging as recommended by clinical practice guidelines [16].

The annotators reviewed the output from 120 processed radiology reports in 3 different training sets for iterative validation cycles to refine the algorithm. A total of 360 reports were reviewed. After abstracting and classifying the radiology reports, the information was entered and stored in a digital data set. Reports with a diagnosis of AAA were categorized as “case”; if there was no evidence of AAA or if an alternate diagnosis other than AAA was reported, the report was categorized as “control.” A board-certified cardiologist verified the information and resolved discrepancies in patient classification.
**Figure 1.** Study design. AAA: abdominal aortic aneurysm; EHR: electronic health record; NLP: natural language processing.

**Textbox 1.** Abdominal aortic aneurysm (AAA)–natural language processing rule and examples of text span.

<table>
<thead>
<tr>
<th>Rule (any token + keyword for AAA + any token)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples of confirmatory assertions</td>
</tr>
<tr>
<td>• Suprarenal aortic abdominal aneurysm which measures up to 5.2 cm</td>
</tr>
<tr>
<td>• Fusiform infrarenal abdominal aortic aneurysm terminating proximal to the aortobiiliac bifurcation, 56 mm, previously 56 mm</td>
</tr>
<tr>
<td>• There is a 5.7×5.1 cm infrarenal aortic aneurysm measured on image 175 of series 4</td>
</tr>
<tr>
<td>Examples of negated assertions</td>
</tr>
<tr>
<td>• Negative for abdominal aortic aneurysm or dissection</td>
</tr>
<tr>
<td>• Abdominal aortic aneurysm is absent</td>
</tr>
<tr>
<td>• Negative for thoracic or abdominal aortic aneurysm, dissection, penetrating atherosclerotic ulcer or intramural hematoma</td>
</tr>
</tbody>
</table>

**Statistical Analysis**

The information extracted by the AAA-NLP algorithm from radiology reports in near real-time was compared to the reference standard manual review of radiology reports following written guidelines for standardization to calculate PPV, sensitivity, specificity, and F1 score. The formula to calculate F1 score was given as follows: \(2 \times \frac{(\text{PPV} \times \text{sensitivity})}{(\text{PPV} + \text{sensitivity})}\) [5].

**Ethics Approval**

This project was approved by the Mayo Clinic Institutional Review Board (approval number 21-006950).

**Results**

Reports of 295 patients were validated in 3 different iterations. The data set for each iteration contained 120 reports, but 46 (16%) patients had more than one report. The reasons for more than one report for the same patient were imaging tests performed before and after repair procedures or surveillance for serial assessment of AAA (Table 1). There were no discrepancies regarding AAA diagnosis between 2 or more imaging reports from the same patient. Table 1 shows the distribution of demographic characteristics across AAA cases and controls. Cases and controls had similar ages in each of the iterative validation cycles, and most patients were Caucasian. AAA cases were more likely to have a history of smoking.

For evaluation of the AAA-NLP algorithm performance, 120 processed reports from each iteration were randomly selected. A total of 360 processed reports were reviewed by 2 physicians blinded to AAA diagnosis. There was 100% agreement for interactions 1 and 3. For interaction 2, the annotators disagreed on 1 report yielding a kappa coefficient of 92%. The disagreement was resolved by a board-certified cardiologist, creating the reference standard for comparison. The number of reports classified by the reference standard as true positives, false positives, true negatives, and false negatives in each iteration is shown in Table 2.
Table 1. Clinical characteristics and radiology report information.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Iteration 1</th>
<th>Iteration 2</th>
<th>Iteration 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case (n=31)</td>
<td>Control (n=52)</td>
<td>Case (n=44)</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>78.6 (11.1)</td>
<td>74.4 (12.4)</td>
<td>70.3 (8.4)</td>
</tr>
<tr>
<td>Male sex, n (%)</td>
<td>26 (84)</td>
<td>21 (40)</td>
<td>34 (77)</td>
</tr>
<tr>
<td>Caucasian, n (%)</td>
<td>31 (100)</td>
<td>52 (100)</td>
<td>42 (95)</td>
</tr>
<tr>
<td>Comorbidities, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>24 (77)</td>
<td>39 (75)</td>
<td>31 (70)</td>
</tr>
<tr>
<td>Hyperlipidemia</td>
<td>21 (68)</td>
<td>22 (42)</td>
<td>29 (66)</td>
</tr>
<tr>
<td>Smoking history</td>
<td>29 (94)</td>
<td>24 (46)</td>
<td>35 (80)</td>
</tr>
<tr>
<td>DM*</td>
<td>9 (29)</td>
<td>7 (13)</td>
<td>10 (23)</td>
</tr>
<tr>
<td>PADb</td>
<td>4 (13)</td>
<td>4 (8)</td>
<td>5 (11)</td>
</tr>
<tr>
<td>CADc</td>
<td>16 (52)</td>
<td>7 (13)</td>
<td>18 (41)</td>
</tr>
<tr>
<td>Radiology reports</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients with ≥2 reports, n</td>
<td>18</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>AAAd diameter (cm), mean (SD)</td>
<td>4.6 (1.08)</td>
<td>N/Ae</td>
<td>4.8 (1.3)</td>
</tr>
<tr>
<td>Reports after AAA repair, n</td>
<td>2</td>
<td>N/A</td>
<td>9</td>
</tr>
</tbody>
</table>

aDM: diabetes mellitus.
bPAD: peripheral artery disease.
cCAD: coronary artery disease.
dAAA: abdominal aortic aneurysm.
eN/A: not applicable.

Table 2. Classification of abdominal aortic aneurysm from radiology reports during iterative validation.

<table>
<thead>
<tr>
<th></th>
<th>Iteration 1</th>
<th>Iteration 2</th>
<th>Iteration 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predicted case</td>
<td>Predicted control</td>
<td>Total</td>
</tr>
<tr>
<td>Actual case</td>
<td>TP* 59</td>
<td>FNb 6</td>
<td>65</td>
</tr>
<tr>
<td>Actual control</td>
<td>FPc 1</td>
<td>TNd 54</td>
<td>55</td>
</tr>
<tr>
<td>Total</td>
<td>60</td>
<td>60</td>
<td>120</td>
</tr>
</tbody>
</table>

aTP: true positive.
bFN: false negative.
cFP: false positive.
dTN: true negative.

Radiology reports are composed of multiple sections. Figure 2 shows an example of a deidentified radiology report with all sections.

During the first iteration implementation, section ID number was used and section detection was challenging. For the second iteration, the algorithm was revised to include section header names for the filter criteria and solve sentence boundary issues. For the third iteration, section detection was implemented based on section names from our complete corpus using the frequency of normalized text with the tool lexical variant generation of the National Library of Medicine [17]. In a separate experiment, 203 additional radiology reports were reviewed by the annotators for evaluation of report section extraction, which resulted in accuracy of 0.96.

During this iterative refinement process, the report sections termed “reason for exam,” “referral diagnosis,” “exam type,” and “signed by” (Figure 2) were excluded, resulting in enhanced NLP algorithm performance. The report sections selected for processing were findings and impressions. During each iteration, the algorithm performance further improved. The performance metrics of the iterations are summarized in Table 3.

During the last iteration, 3 false negatives and 1 false positive contributed to the error analysis. False negatives were due to the complex nature of narrative text in these reports (ie, no significant interval changes in appearances of a partially
thrombosed infrarenal AAA measuring 42×40 mm, extending to the level of aortic bifurcation and proximal common iliac arteries; no signs of rupture or impending rupture of the known infrarenal AAA; and no slightly increased size of fusiform infrarenal AAA). Additionally, the false positive was due to a typographical error, which was the report of a patient with an aorta diameter of 2.7 cm labeled as AAA, which does not meet the criteria for AAA (≥3.0 cm).

**Figure 2.** Example of deidentified radiology report with all sections. In this figure, section names are displayed in blue font. AAA: abdominal aortic aneurysm.

<table>
<thead>
<tr>
<th>Performance metric</th>
<th>Iteration 1 (n=120)</th>
<th>Iteration 2 (n=120)</th>
<th>Iteration 3 (n=120)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.91</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>PPV*</td>
<td>0.98</td>
<td>0.93</td>
<td>0.98</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.98</td>
<td>0.94</td>
<td>0.98</td>
</tr>
<tr>
<td>F1 score</td>
<td>0.94</td>
<td>0.95</td>
<td>0.97</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.94</td>
<td>0.95</td>
<td>0.97</td>
</tr>
</tbody>
</table>

*PPV: positive predictive value.

**Discussion**

**Overview**

In this study, a novel rule-based NLP algorithm was developed for the extraction of AAA diagnosis from radiology reports and prospectively deployed in the institutional big data infrastructure for near real-time processing. Compared to the reference standard of manual review of radiology reports, the AAA-NLP algorithm extracted AAA diagnosis in near real time with high sensitivity, PPV, F1 score, specificity, and accuracy.

To the best of our knowledge, this study is the first to describe the use of NLP algorithms prospectively to extract AAA diagnosis in near real time from radiology reports. Clinicians, information technologists, and informaticians collaborated to refine the algorithm to improve performance. In previous studies, billing codes were used to find AAA cases [18,19]. However, in those studies, the cohorts were limited to patients with AAA who underwent procedures for aneurysm repair or had a history of ruptured AAA [18,19]. No prior studies using billing codes algorithms retrieved a broader spectrum of AAA diagnosis while also including patients presenting with uncomplicated AAA (ie, patients who did not undergo prior repair or who had not previously presented with ruptured AAA). In contrast, in this study, NLP automatically extracted AAA diagnosis from radiology reports prospectively and regardless of prior repair or rupture, thereby expanding the scope of computational approaches to include the detection of AAA cases prior to rupture or repair.

A radiology report consists of free text, organized into standard sections [5]. The American College of Radiology has published guidelines with recommendations for the use of sections for narrative (free text) entry in radiology reports [20]. NLP techniques enable the automatic extraction of information from narrative text [6-8]. Moreover, information extracted by NLP can be used to populate CDS systems automatically without the need for manual data entry and be better aligned with existing workflows such that radiologists can spend time interpreting images rather than filling out forms.
NLP is a computational methodology used for electronic phenotyping to extract meaningful clinical information from text fields [6,7,21]. In this study, we used NLP to process radiology text reports. The previous NLP algorithm used to find cases of AAA from radiology reports [12] was designed for retrospective cohort identification, whereas this report describes the prospective implementation of an NLP algorithm for input to a patient-specific CDS system for near real-time processing of radiology reports. Near real-time processing requires <3 milliseconds to process a document after a radiologist releases a report to the EHR [22]. The AAA-NLP implementation described in this study was developed within the existing digital infrastructure and can be used in clinical practice immediately without the need to retrain the algorithm. Additionally, the previously described algorithm [12] did not identify document sections in the radiology reports. By selecting specific sections for NLP information extraction, improvement in NLP performance was observed, as shown in the Results section. In the future, transformer-based NLP models [23,24] may be trained to interpret nuanced language, and ablation experiments [25] could be used to further evaluate these models.

The use of NLP algorithms has advantages compared to other methods. In comparison, the use of check box forms in radiology reports may require the development of new workflows [26,27]. The use of check box forms also requires the radiologist to direct attention away from the imaging interpretation process [26,27]. Manual entry of summaries of radiology findings in a check box can increase reporting time with decreased radiologist productivity [26,27]. Check box use could also result in the loss of important and clinically relevant descriptive information available only in the radiology narrative reports.

The rule-based AAA-NLP algorithm described in this study shows accurate detection of a broad spectrum of AAA cases prospectively in near real time from radiology reports, regardless of the presence of prior rupture or repair. This methodology will also potentially generate input for CDS to assist providers in managing patients with AAA by displaying the relevant information automatically at the point of care and in near real time for CDS tools. It will also support the automatic identification of cohorts for research purposes (eg, cohorts for clinical trials) and quality projects, and will support a learning health care system. NLP has been previously used for the identification of peripheral arterial disease and critical limb ischemia from narrative clinical notes of EHRs [21,28]. Therefore, it will also be possible to develop NLP algorithms for the identification of AAA cases from clinical notes in near real time.

In efforts to develop a learning health care system, Mayo Clinic has developed a robust big data–empowered clinical NLP infrastructure that enables near-real-time NLP processing for the delivery of relevant information to the point of care via CDS [22]. Accordingly, we have deployed the AAA-NLP algorithm described herein to this digital infrastructure for translation to clinical practice. Importantly, the near real-time identification of patients with AAA by NLP responds to the American Heart Association scientific statement, which recommends the implementation of technologies to extract clinical information in real time that will promptly provide synopses of the information extracted [29].

Limitations

This NLP algorithm was developed, tested, and implemented in a single tertiary medical center. Future studies should evaluate this algorithm at other institutions to demonstrate portability. A robust institutional digital infrastructure is required for the execution of near real-time processing of radiology reports [22]. Hence, the absence of adequate digital infrastructure may limit porting of this algorithm. For implementation, the analysis of radiology report architecture to enable the selection of document types and document sections may also be necessary for portability. Another potential challenge for porting this algorithm to other EHRs is differences in lexicons used for the extraction of the AAA concept across institutions. In mitigation, for this NLP algorithm, each lexicon was mapped to corresponding concepts and synonyms in the publicly available Unified Medical Language System Metathesaurus for standardization.

The algorithm was developed for the extraction of AAA diagnosis but not for the extraction of iliac artery or thoracic aortic aneurysms. Future studies should create and validate NLP algorithms for the extraction of thoracic and iliac artery aneurysms. The clinical criteria for AAA diagnosis involve a minimum diameter, but this NLP algorithm did not interpret the reported diameter. This is an area for future improvement in the algorithm, as clinical criteria for AAA may change over time. In this study, most patients were Caucasian. This was likely related to the ethnic distribution of communities in the Midwest, where this study was conducted [30,31]. Additionally, prior studies have reported a higher prevalence of AAA among Caucasians compared to other races [31,32]. There were differences in comorbidities of patients included in the 3 iterations. However, the NLP was developed for the extraction of the diagnosis of AAA and not developed for the extraction of associated patient comorbidities. The differences in patient comorbidities did not influence NLP performance for the extraction of AAA from radiology reports.

Conclusions

Implementation of NLP for prospective identification of AAA cases from radiology reports in near real time with high performance is feasible. This near real-time NLP technique described will potentially be helpful for the generation of automated input for CDS tools to assist clinicians in the management of patients with AAA, quality improvement projects, and research (automated identification of cohorts).

Acknowledgments

The authors would like to thank Kara M Firzlaff for secretarial support and Christopher G Scott for statistical analysis. This study was funded by the Mayo Clinic K2R award.
Data Availability
The data sets generated and analyzed during this study are not publicly available because participants in this study did not agree for their data to be shared publicly but are available from the corresponding author on reasonable request.

Conflicts of Interest
None declared.

References


Abbreviations

AAA: abdominal aortic aneurysm
CDS: clinical decision support
CT: computerized tomography
EHR: electronic health record
IV: intravenous
MRI: magnetic resonance imaging
NLP: natural language processing
PPV: positive predictive value
US: ultrasound
Deep Learning Approach for Negation and Speculation Detection for Automated Important Finding Flagging and Extraction in Radiology Report: Internal Validation and Technique Comparison Study

Kung-Hsun Weng¹, MSc, MD; Chung-Feng Liu², PhD; Chia-Jung Chen³, MSc

¹Department of Medical Imaging, Chi Mei Medical Center, Chiali, Tainan, Taiwan
²Department of Medical Research, Chi Mei Medical Center, Tainan, Taiwan
³Department of Information Systems, Chi Mei Medical Center, Tainan, Taiwan

Corresponding Author:
Chia-Jung Chen, MSc
Department of Information Systems
Chi Mei Medical Center
No.901, Zhonghua Rd.
Yongkang Dist.
Tainan, 71004
Taiwan
Phone: 886 6 2812811 ext 52069
Email: carolchen@mail.chimei.org.tw

Abstract

Background: Negation and speculation unrelated to abnormal findings can lead to false-positive alarms for automatic radiology report highlighting or flagging by laboratory information systems.

Objective: This internal validation study evaluated the performance of natural language processing methods (NegEx, NegBio, NegBERT, and transformers).

Methods: We annotated all negative and speculative statements unrelated to abnormal findings in reports. In experiment 1, we fine-tuned several transformer models (ALBERT [A Lite Bidirectional Encoder Representations from Transformers], BERT [Bidirectional Encoder Representations from Transformers], DeBERTa [Decoding-Enhanced BERT With Disentangled Attention], DistilBERT [Distilled version of BERT], ELECTRA [Efficiently Learning an Encoder That Classifies Token Replacements Accurately], ERNIE [Enhanced Representation through Knowledge Integration], RoBERTa [Robustly Optimized BERT Pretraining Approach], SpanBERT, and XLNet) and compared their performance using precision, recall, accuracy, and F₁-scores. In experiment 2, we compared the best model from experiment 1 with 3 established negation and speculation-detection algorithms (NegEx, NegBio, and NegBERT).

Results: Our study collected 6000 radiology reports from 3 branches of the Chi Mei Hospital, covering multiple imaging modalities and body parts. A total of 15.01% (105,755/704,512) of words and 39.45% (4529/11,480) of important diagnostic keywords occurred in negative or speculative statements unrelated to abnormal findings. In experiment 1, all models achieved an accuracy of >0.98 and F₁-score of >0.90 on the test data set. ALBERT exhibited the best performance (accuracy=0.991; F₁-score=0.958). In experiment 2, ALBERT outperformed the optimized NegEx, NegBio, and NegBERT methods in terms of overall performance (accuracy=0.996; F₁-score=0.991), in the prediction of whether diagnostic keywords occur in speculative statements unrelated to abnormal findings, and in the improvement of the performance of keyword extraction (accuracy=0.996; F₁-score=0.997).

Conclusions: The ALBERT deep learning method showed the best performance. Our results represent a significant advancement in the clinical applications of computer-aided notification systems.

(JMIR Med Inform 2023;11:e46348) doi:10.2196/46348

https://medinform.jmir.org/2023/1/e46348
KEYWORDS
radiology report; natural language processing; negation; deep learning; transfer learning; supervised learning; validation study; Bidirectional Encoder Representations from Transformers; BERT; clinical application; radiology

Introduction

Background
Timely and effective communication of test results is essential in modern medicine. To promptly address patients’ problems, hospitals must ensure that the test results are completed without delay and that clinicians are aware of substantial abnormal findings. Delayed or failed communication of important findings by the department performing the test and the clinical team can increase the risk of adverse patient events and result in medical malpractice and compensation, especially for potentially life-threatening and important diagnoses [1].

Although radiology reports are the primary method of communication between radiology and clinical departments, the fact that a radiologist produces a report does not necessarily mean that the clinician reads it entirely. Ignácio et al [2] showed that only 55.7% of clinicians read the entire report thoroughly. Reda et al [3] showed that >40% of clinicians read only the conclusions or only read the conclusions in detail. More than 30% of clinicians have made preventable medical errors because they did not read radiology reports carefully. Even if the radiologist has made the correct diagnosis in the report, the clinician may still miss it.

To address these communication issues, current radiology guidelines [4] now require radiologists to go beyond report completion and use additional communication methods for reports with significant findings, including flagging or alerting the report, e-mailing, or direct verbal communication via telephone. Natural language processing can also automatically extract data from radiology reports, for example, automatically extracting important diagnoses, follow-up data, or management recommendations or automatically identifying reports that require specific action [5]. These methods can help to identify important information in radiology reports or reports that need to be read in detail to alert clinicians.

In addition, the laboratory information system (LIS) used in hospitals today can automatically highlight abnormalities found in tests and display them differently to ensure that clinicians do not miss important findings, such as using different colors or special symbols [6]. For example, in our hospital, if a patient has undergone a routine blood test and some of the blood cell counts are abnormal, the LIS will automatically display the results on the computer screen in a unique color for the abnormal values and a typical color for the others. The LIS also displays important keywords (e.g., nodules) within radiology reports in different colors.

However, because most radiology reports are freely typed by radiologists in an unstructured manner, both techniques encounter challenges. Negative and speculative statements are significant problems.

Radiologists can use negative statements to communicate the absence of specific diagnoses and provide a clearer picture of the patient’s condition. For example, the statement “No definite CT evidence of aortic dissection” informs the clinician that the patient’s condition is not related to aortic dissection.

The diagnoses in the speculative statements may or may not be related to the actual abnormal findings. The radiology report may contain speculative statements in the presence of an imaging finding of uncertain significance that requires further investigation, for example, “RUL lung nodule. Lung cancer should be suspected.” In such cases, the diagnoses (lung cancer) in the speculative statements are related to abnormal findings. Even if the radiologist finds no problems with the study, the radiology report may still contain speculative statements to prevent potential medicolegal issues. Disclaimer (eg, “10%–15% of cases of breast cancer are missed on mammograms” [7]) or statement of limitations (eg, “non-enhanced images, small lesion may be obscured”) are common examples. In such cases, the diagnoses (breast cancer or lesion) in the speculative statements are unrelated to the actual diagnoses.

A notification system that does not distinguish whether diagnostic Information is contained in negative or speculative statements unrelated to abnormal findings and annotates or extracts all of them to “alert” the clinician may generate excessive false alarms. Excessive false alarms can overload the clinician’s senses and lead to the “cry wolf” phenomenon, causing alarm fatigue. Consequently, clinicians may delay detection or even ignore truly valuable alerts, posing a risk to patients, especially if the percentage of false alarms is high [8].

This study aimed to address the potential analytical inaccuracies resulting from negative and speculative statements in radiology reports and to facilitate the use of unstructured reports by hospital information systems.

Prior Work
Current studies have adopted various approaches to detect negation and speculation, including rule-based, machine learning–based, and deep learning–based approaches [9-17].

The rule-based approach relies on experts to define the rules that are understandable to humans. NegEx, proposed by Chapman et al [18]; NegFinder, proposed by Mutalik et al [19]; NegHunter, proposed by Gindl et al [20]; and NegExpanzer, proposed by Aronow et al [21], are regular expression-based approaches. Regular expression-based methods have limitations, such as the inability to capture the syntactic structure and the possibility of misinterpreting the scope of the negative and speculative statements. For example, “No change of tumor” may be misinterpreted as both “No change” and “No tumor.”

Methods such as DEEPEN (Dependency Parser Negation), proposed by Mehrabi et al [22], and NegBio, proposed by Peng et al [23], analyze the syntactic structure based on grammar. These methods are more accurate than regular expression-based approaches in limiting the scope of negative and speculative statements and reducing false positives because these methods consider the dependency relationship between words. However,
these methods have certain limitations. For example, errors in the analysis may occur if the grammar of the text deviates from
typical norms, such as the presence of long noun phrases [23].
When analyzing text, most of these methods [18-20,22,23] split
the text into sentences that are analyzed independently. The
algorithms and expert-defined rules only consider a single
sentence at once and do not consider both the preceding and
following contexts.

With the advancement of artificial intelligence, machine learning
techniques have been applied to detect negation and speculation.
For example, Medlock et al [24] proposed a weakly supervised
learning-based approach to predict the labels of training samples
for machine learning training and used the trained models to
detect speculation in biomedical texts. Rokach et al [25]
compared several machine learning approaches, including the
Hidden Markov Model, Conditional Random Field (CRF),
decision tree, and AdaBoost, cascaded decision tree classifiers
and with and without the Longest Common Sequence. They found
that the cascaded decision tree with the Longest Common
Sequence performed best. Morante et al proposed k-nearest
neighbor algorithm-based [26] and meta-learning-based
approaches [27]. Ou et al [28] compared rule-based and support
vector machine-based machine learning methods and obtained
better performance of machine learning methods.

Later studies began investigating deep learning-based
approaches and achieved better results than previous non-deep
learning approaches. Qian et al [17] were the first to propose a
deep learning method for negation and speculation detection
using a convolutional neural network-based model by using
the relative position of tokens and path features from syntactic
trees as features.

By contrast, recurrent neural networks and their derivatives,
such as Long Short-Term Memory (LSTM), are suitable for
processing sequential data. These architectures can incorporate
dependencies on preceding and following elements, making
them particularly useful for natural language processing tasks,
and have achieved good results in recognizing negations and
speculations. For example, in a study by Fancellu et al [14], a
Bidirectional LSTM (BiLSTM)-based model was applied, and
it demonstrated better performance than other methods on the
Sherlock data set. Lazib et al [9] compared methods, including
LSTM, BiLSTM, Gated Recurrent Unit, and CRF, and showed
that the recurrent neural network-based architecture performed
the best. Gautam et al [15] compared several LSTM-based
models and obtained the best performance using 2-layer
encoders and decoders with dropouts. Taylor et al [10] applied
the BiLSTM-based model to the analysis of negation in
an LSTM-based approach and investigated the effect of
expert-provided negation cues on the detection performance of
the negation scopes. Sykes et al [12] compared the methods
based on BiLSTM and feedforward neural networks and
rule-based methods, including pyCorText, NegBio, and EdIE-R,
for negation detection in radiology reports. The BiLSTM-based
approach outperformed other approaches.

BERT (Bidirectional Encoder Representations from
Transformers) [29], proposed by Google in 2018, is a pretrained,
transformer-based model that is effective for negation detection.
Khandelwal et al [16] developed NegBERT and, in another
study [13], used a multitasking approach with BERT, XLNet,
and RoBERTa (Robustly Optimized BERT Pretraining
Approach) for negation and speculation detection, with
improved results on BioScope and Simon Fraser University
review data sets compared with the control methods. Zavala et
al [30] proposed a system based on BiLSTM with CRF and
fine-tuned BERT; evaluated the methods on English and Spanish
clinical, biomedical, and review text; and showed improved
performance compared with previous methods. They also found
that pretrained word embedding, especially contextualized
embedding, helped to understand the biomedical text.

Numerous variants of BERT have been developed to improve
performance and simplify the model. ALBERT (A Lite BERT)
[31] reduces the model parameters and improves the
performance through parameter sharing and matrix
decomposition. DistilBERT (Distilled version of BERT) [32]
uses knowledge distillation to reduce the size and improve the
inference speed while retaining most of the language
understanding. XLNet [33] implements autoregressive training
while preserving the advantages of autoencoding models and
outperforms BERT on 20 tasks. RoBERTa [34] improves the
training method to outperform BERT and XLNet. ERNIE
(Enhanced Representation through Knowledge Integration) [35]
uses an alternative masking method to outperform BERT in
Chinese tasks. SpanBERT [36] extends BERT with span-based
masking and an additional training objective, resulting in a
better performance on span-based tasks. DeBERTa
(Decoding-Enhanced BERT With Disentangled Attention) [37]
improves BERT and RoBERTa with decoupled attention,
improved mask encoder, and virtual adversarial training and
outperforms RoBERTa-Large on the Multigene Natural
Language Inference, Stanford Question Answering Data set,
and Reading Comprehension data set from examinations tasks
and humans on the SuperGLUE task. ELECTRA (Efficiently
Learning an Encoder That Classifies Token Replacements
Accurately) [38] outperforms BERT with a new pretraining
task, Replaced Token Detection, and performs similarly to
RoBERTa and XLNet with one-fourth the computation.

**Contribution of This Work**

This study has implications for optimizing the performance of
hospital information systems in managing unstructured
electronic medical records. The key findings and results of this
study are as follows.

First, we found that fine-tuned general-purpose transformer
models could outperform NegEx, NegBio, and NegBERT, which
are explicitly designed for negation and speculation detection.
We identified sources of error in the latter 3 methods and
suggested potential improvements.

Second, we found that transformer, unlike NegEx and NegBio,
demonstrated the ability to perform multisentence contextual
analysis and further granular classification of speculative
statements as related or unrelated to abnormal findings. This
capability can improve information filtering in hospital
information systems to eliminate nondiagnostically relevant
information.
Finally, in contrast to other studies using BERT [16,39], we found that using a lightweight transformer model and learning the cues and scopes of negative and speculative sentences in a single step can perform well.

Methods

Ethics Approval
The Chi Mei Hospital Institutional Review Board reviewed and approved this study (11105-J02). This study is a retrospective analysis study using deidentified electronic medical records, thus obviating the requirement for obtaining informed consent from the individuals. Figure 1 shows the flow diagram of the study.

Inclusion and Exclusion Criteria
The inclusion criteria for this study were radiological examinations performed in the 3 branches of our institution between 2012 and 2022, with the reports being written in English language and the type of examination being x-ray, special radiology, computed tomography (CT), magnetic resonance imaging (MRI), or ultrasound. We included cases that met all criteria. The exclusion criteria were Chinese reports and patients aged <20 years at the time of examination. We excluded cases that met any of the exclusion criteria. Samples were collected using 2 independent keyword searches in a search engine targeting radiology reports that met the inclusion criteria but not the exclusion criteria.

Data

Overview
The training and development data set consisted of 5000 radiology reports randomly selected from a keyword search using the terms “fracture,” “dissection,” “infarct,” “pneumothorax,” “extravasation,” “thrombosis,” or “pneumoperitoneum.” The test data set consisted of 1000 reports selected from a keyword search using the terms “tumor,” “consolidation,” “pulmonary TB,” “metastasis,” or “bleeding.” Keywords were selected from our institution’s list of important keywords and randomly assigned to the data sets. These keywords are referred to as “important keywords” in the study. The samples in the training and development and test data sets were mutually exclusive with no overlap.

The training and development data set was automatically partitioned into training and development data sets in a 9:1 ratio for model training. The training, development, and test data sets ratio was 9:1:2, with 4500, 500, and 1000 radiology reports, respectively.

In this study, each word or token was assigned to one of the 2 categories, as shown in Table 1: “Positive statements, or speculative statements potentially related to abnormal findings” (category 0) and “negative statements, or speculative statements not related to abnormal findings” (category 1). We combined speculative statements unrelated to abnormal findings with negative statements as a single class because of their limited representation. The rationale for category 1 is that the information conveyed is not relevant to abnormal findings and should not trigger highlights or alerts. A token is the minimum output unit of the transformer-based model’s tokenizer.

All radiology reports included in the study were deidentified by removing identifying information such as medical record...
number, application number, examination date, ordering department, and examination time. A radiologist with 12 years of experience (KHW) reviewed the reports and annotated all negative and speculative statements unrelated to abnormal findings using the open-source Doccano [40] software. The annotation served as the gold standard for subsequent analysis.

Table 1. Classification of words and tokens in this study.

<table>
<thead>
<tr>
<th>Typea and subtype</th>
<th>Example</th>
<th>Categoryb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>Liver laceration at S6 without active contrast extravasation</td>
<td>1</td>
</tr>
<tr>
<td>Speculative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unrelated to abnormal findings</td>
<td>No CT evidence of large infarct. Suggest MRI to exclude hyperacute infarct if indicated</td>
<td>1</td>
</tr>
<tr>
<td>Potentially related to abnormal findings</td>
<td>Rt cerebellum acute infarct cannot be ruled out.</td>
<td>0</td>
</tr>
<tr>
<td>Positive</td>
<td>Rt cerebellum acute infarct</td>
<td>0</td>
</tr>
</tbody>
</table>

aType refers to the type of statement.
bToken category in the italicized text if italicization is used. All texts without italics were classified as category 0. Category 0: positive statements or speculative statements potentially related to abnormal findings. Category 1: negative statements or speculative statements not related to abnormal findings.

Included Negations

This study included all statements in which the radiologist explicitly denied a diagnosis or a finding. Our data included morphological negation and sentential negation, which are common forms of negative statements in English text [22]. Morphological negation involves using prefixes, such as “un-” or “ir-,” to modify certain words to express negation. Sentential negation involves using negative words, such as “no” or “without,” to negate part of the statement. In addition, radiologists at the authors’ hospital often use unique symbols or abbreviations, such as “(−)” or “[−].”

Included Speculations

In cases where the imaging study is inconclusive but there is still the possibility of a significant abnormality, the information system should notify the clinician and allow the clinician to make the final decision. Therefore, for the task of speculation detection, our focus was limited to speculative statements that were unrelated to abnormal findings. Meanwhile, we treated speculative statements that may correlate with actual abnormal findings as equivalent to positive statements.

After reviewing the samples, we identified 2 scenarios in which speculative statements could be confidently determined to be unrelated to abnormal findings. First, the radiologist explicitly stated that there was no relevant abnormality. Second, the radiologist stated that certain diagnoses could not be evaluated owing to study limitations. In all the other scenarios, speculative statements may be associated with abnormal findings.

In the following 3 examples, we classify the diagnoses or findings written in italics as speculative statements unrelated to abnormal findings. The actual test results were normal or unrelated to these diagnoses or findings.

1. No CT evidence of pulmonary embolism. Suggest V/Q scan to exclude small branch embolism if indicated.
2. No CT evidence of large infarct. Suggest MRI to exclude hyperacute infarct if indicated.
3. Liver tumor cannot be excluded by noncontrast CT.

In the following 2 examples, the diagnoses or findings written in italics are speculative statements considered potentially related to actual abnormal findings:

1. Equivocal filling defect in RLL segmental pulmonary artery. Suggest V/Q scan to exclude small branch embolism if indicated.
2. Rt cerebellum acute infarct cannot be ruled out.

Design of the Experiments

We conducted 2 experiments to evaluate the ability of general all-purpose pretrained deep learning models and existing negation and speculation-detection algorithms to identify negation and speculation in real-world radiology reports.

In experiment 1 (Figure 2), we fine-tuned several transformer-based models using our training and validation data sets. We performed token category prediction (category 0 or 1) for all tokens in the training, validation, and test data sets.

In experiment 2 (Figure 3), we compared 3 negation and speculation-detection algorithms that performed well on public data sets with the best model from experiment 1. The algorithms evaluated were NegEx, NegBio, which has predefined expert rules and open-source implementation, and NegBERT, whose training code is available. We then performed category prediction (category 0 or 1) for all words that matched a given “important keyword” in the test data set. We also analyzed the sources of errors. In addition, we compared the performance of keyword extraction in positive and speculative statements potentially related to abnormal findings before and after applying various algorithms.
Figure 2. Experiment 1. X: the original text, ŷ: class predicted by the model; y: the gold standard. Category 0: positive statements or speculative statements potentially related to abnormal findings; category 1: negative statements or speculative statements unrelated to abnormal findings. ALBERT: A Lite Bidirectional Encoder Representations From Transformers; BERT: Bidirectional Encoder Representations From Transformers; DeBERTa: Decoding-Enhanced Bidirectional Encoder Representations From Transformers With Disentangled Attention; DistilBERT: Distilled version of Bidirectional Encoder Representations From Transformers; ELECTRA: Efficiently Learning an Encoder That Classifies Token Replacements Accurately; ERNIE: Enhanced Representation through Knowledge Integration; RoBERTa: Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach; RUL: right upper lobe.

Figure 3. Experiment 2 Note. X: the original text; ŷ: class predicted by the model; y: the gold standard; category 0: positive statements or speculative statements potentially related to abnormal findings; category 1: negative statements or speculative statements unrelated to abnormal findings; bold text: word matching a designated “important keyword.” Exp: experiment.
Modeling in Experiments

The deep learning models used in experiment 1 were ALBERT, BERT, DeBERTa, DistilBERT, ELECTRA, ERNIE, RoBERTa, SpanBERT, and XLNet. All models were fine-tuned based on the pretrained models from Hugging Face.

We used early stopping and used the F1-score as the model evaluation metric. We used the Adam optimizer with a batch size of 16 and weight decay of 0.01. Table 2 lists the parameters of other models. We set all other unspecified parameters to the default values provided by the open-source PyTorch framework. We segmented the texts into blocks of no more than 510 characters before entering the model to avoid truncation.

We adopted a sequence-to-sequence approach for the training. The training program input the report text in the training and development data set into the model using the corresponding tokenizer and trained the model. The models predicted the token categories using the radiologist-annotated data as the gold standard. The test data set was not included in the training process.

For the NegEx algorithm, we used the negspaCy pipeline component of the open-source Spacy software [41]. The specific named entity recognition model used was “en_ner_bc5cdr_md.”

Table 2. Deep learning model and training parameters used in this study.

<table>
<thead>
<tr>
<th>Model</th>
<th>Learning rate</th>
<th>Warm-up steps</th>
<th>Adam beta1</th>
<th>Adam beta2</th>
<th>Adam epsilon</th>
<th>FP16</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALBERTb</td>
<td>1 × 10^{-5}</td>
<td>10,000</td>
<td>0.9</td>
<td>0.999</td>
<td>1 × 10^{-8}</td>
<td>False</td>
</tr>
<tr>
<td>BERTc</td>
<td>1 × 10^{-4}</td>
<td>10,000</td>
<td>0.9</td>
<td>0.999</td>
<td>1 × 10^{-8}</td>
<td>False</td>
</tr>
<tr>
<td>DeBERTa d</td>
<td>1 × 10^{-4}</td>
<td>10,000</td>
<td>0.9</td>
<td>0.999</td>
<td>1 × 10^{-6}</td>
<td>True</td>
</tr>
<tr>
<td>DistilBERTe</td>
<td>2 × 10^{-5}</td>
<td>0</td>
<td>0.9</td>
<td>0.999</td>
<td>1 × 10^{-8}</td>
<td>False</td>
</tr>
<tr>
<td>ELECTRAf</td>
<td>1 × 10^{-4}</td>
<td>10,000</td>
<td>0.9</td>
<td>0.999</td>
<td>1 × 10^{-6}</td>
<td>False</td>
</tr>
<tr>
<td>ERNIEg</td>
<td>5 × 10^{-5}</td>
<td>4000</td>
<td>0.9</td>
<td>0.98</td>
<td>1 × 10^{-8}</td>
<td>False</td>
</tr>
<tr>
<td>RoBERTah</td>
<td>1 × 10^{-4}</td>
<td>10,000</td>
<td>0.9</td>
<td>0.999</td>
<td>1 × 10^{-8}</td>
<td>False</td>
</tr>
<tr>
<td>SpanBERT</td>
<td>5 × 10^{-5}</td>
<td>10,000</td>
<td>0.9</td>
<td>0.999</td>
<td>1 × 10^{-8}</td>
<td>False</td>
</tr>
<tr>
<td>XLNet</td>
<td>2 × 10^{-5}</td>
<td>10,000</td>
<td>0.9</td>
<td>0.999</td>
<td>1 × 10^{-6}</td>
<td>False</td>
</tr>
</tbody>
</table>

aFP16: half-precision floating-point format. 
bALBERT: A Lite Bidirectional Encoder Representations From Transformers. 
cBERT: Bidirectional Encoder Representations From Transformers. 
dDeBERTa: Decoding-Enhanced Bidirectional Encoder Representations From Transformers With Disentangled Attention. 
eDistilBERT: Distilled version of Bidirectional Encoder Representations from Transformers. 
fELECTRA: Efficiently Learning an Encoder That Classifies Token Replacements Accurately. 
gERNIE: Enhanced Representation through Knowledge Integration. 
hRoBERTa: Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach.

Results

Demographics

The data set included in this study consisted of 6000 radiology reports, including plain radiography reports (2538/6000, 42.3%), CT reports (2163/6000, 36.05%), MRI reports (668/6000, 11.13%), ultrasound reports (483/6000, 8.05%), angiography reports (97/6000, 1.62%), and reports from other types of studies (51/6000, 0.85%). The report was completed by 78 radiology residents and their attending physicians. The training, validation and test data sets were mutually exclusive with no overlap in the samples.

The data set used in this study consisted of 78,901 sentences and 704,512 words. A total of 15.01% (105,755/704,512) of all
the words in the data set, were identified as negative and speculative statements unrelated to abnormal findings. Table 3 presents examples and frequencies of these statements. In this study, we defined a "word" as a contiguous sequence of one or more non–white space characters of maximum length. For example, "(→) metastasis" contains 2 words.

Of all the 16,374 cases of sentential negations identified, 15,568 (95.1%) used "no," "without," "not," or "none" as the first word of the negative statement. Furthermore, of all the 2763 cases of negation using symbols or abbreviations, we observed that 2411 (87.2%) used (→), (_) ( ), or [→] at the beginning, end, or middle of the negated clause. Table 4 presents the frequency and number of occurrences of important keywords, as defined in this study, within negative or speculative statements unrelated to abnormal findings and the total number of occurrences in the study.

### Table 3. Types and numbers of negative and the speculative sentences unrelated to abnormal findings included in this study (N=19,467).

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>Findings, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentential negation</td>
<td>No evidence of aortic dissection</td>
<td>16,374 (84.11)</td>
</tr>
<tr>
<td>Symbols or abbreviations</td>
<td>Metastasis (→)</td>
<td>2762 (14.19)</td>
</tr>
<tr>
<td></td>
<td>Thrombosis: No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DM(^a) HTN(^b)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Anti-HCV(^c) [Negative]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- lung - bone</td>
<td></td>
</tr>
<tr>
<td>Speculative statements not related to abnor-</td>
<td>No CT(^d) evidence of pulmonary embolism. Suggest V/Q(^e) scan to exclude small branch embolism if indicated</td>
<td>196 (1.01)</td>
</tr>
<tr>
<td>mal findings</td>
<td>Metallic artifacts, lesion may be obscured</td>
<td></td>
</tr>
<tr>
<td>Morphological negation</td>
<td>This coronary CT scan is nondiagnostic.</td>
<td>135 (0.69)</td>
</tr>
</tbody>
</table>

\(^a\)DM: diabetes mellitus.  
\(^b\)HTN: hypertension.  
\(^c\)HCV: hepatitis C virus.  
\(^d\)CT: computed tomography.  
\(^e\)V/Q: ventilation and perfusion.

### Table 4. Occurrence and frequency of important keywords defined in this study within negative or the speculative statements unrelated to abnormal findings.

<table>
<thead>
<tr>
<th>Keywords and their overall occurrences (n=11,480)</th>
<th>Occurrences (N+S)(^a), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pneumothorax, n=1288 (11.22%)</td>
<td>976 (75.78)</td>
</tr>
<tr>
<td>Extravasation, n=182 (1.58%)</td>
<td>84 (46.2)</td>
</tr>
<tr>
<td>Fracture, n=2161 (18.82%)</td>
<td>992 (45.90)</td>
</tr>
<tr>
<td>Tumor, n=2698 (23.5%)</td>
<td>1025 (37.99)</td>
</tr>
<tr>
<td>Infarct, n=1364 (11.88%)</td>
<td>514 (37.68)</td>
</tr>
<tr>
<td>Consolidation, n=428 (3.73%)</td>
<td>152 (35.5)</td>
</tr>
<tr>
<td>Pneumoperitoneum, n=63 (0.55%)</td>
<td>19 (30)</td>
</tr>
<tr>
<td>Thrombosis, n=614 (5.35%)</td>
<td>143 (23.3)</td>
</tr>
<tr>
<td>Dissection, n=673 (5.86%)</td>
<td>147 (21.8)</td>
</tr>
<tr>
<td>Metastasis, n=1876 (16.34%)</td>
<td>450 (23.98)</td>
</tr>
<tr>
<td>Bleeding, n=118 (1.03%)</td>
<td>27 (22.9)</td>
</tr>
<tr>
<td>Pulmonary TB(^b), n=15 (0.13%)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

\(^a\)Number of occurrences within negative or speculative statements unrelated to abnormal findings.  
\(^b\)TB: tuberculosis.

### Result of Experiment 1

Table 5 presents the results of experiment 1. The accuracy of all transformer-based models included in this experiment was greater than 0.98 for both the training, validation, and test data sets, with macro \(F_1\)-scores >0.90. The best-performing model, ALBERT, was selected for inclusion in experiment 2.
Table 5. Comparison of deep learning prediction performance.

<table>
<thead>
<tr>
<th></th>
<th>Train and validation data set</th>
<th>Test data set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>ALBERT</td>
<td>0.992</td>
<td>0.990</td>
</tr>
<tr>
<td>BERT</td>
<td>0.980</td>
<td>0.986</td>
</tr>
<tr>
<td>DeBERTa</td>
<td>0.989</td>
<td>0.971</td>
</tr>
<tr>
<td>DistilBERT</td>
<td>0.994</td>
<td>0.990</td>
</tr>
<tr>
<td>ELECTRA</td>
<td>0.982</td>
<td>0.982</td>
</tr>
<tr>
<td>ERNIE</td>
<td>0.987</td>
<td>0.984</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>0.959</td>
<td>0.979</td>
</tr>
<tr>
<td>SpanBERT</td>
<td>0.992</td>
<td>0.992</td>
</tr>
<tr>
<td>XLNet</td>
<td>0.993</td>
<td>0.993</td>
</tr>
</tbody>
</table>

aALBERT: A Lite Bidirectional Encoder Representations From Transformers.

bItalics highlight that the performance of A Lite Bidirectional Encoder Representations From Transformers is the best comparing to the control method across various performance metrics.

cBERT: Bidirectional Encoder Representations From Transformers.

dDeBERTa: Decoding-Enhanced Bidirectional Encoder Representations From Transformers With Disentangled Attention.

eDistilBERT: Distilled version of Bidirectional Encoder Representations from Transformers.

fELECTRA: Efficiently Learning an Encoder That Classifies Token Replacements Accurately.

gERNIE: Enhanced Representation through Knowledge Integration.

hRoBERTa: Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach.

Result of Experiment 2

Before optimization, the performance of NegBio and NegBERT was suboptimal. The F₁-scores for NegEx, NegBio, and NegBERT were 0.889, 0.587, and 0.393, respectively. Our optimization significantly improved the performance of NegBio and NegBERT by increasing their F₁-scores by 0.239 and 0.588, respectively.

Table 6 shows the performance of ALBERT and optimized NegEx, NegBio, and NegBERT. The precision, recall, and F₁-score of our fine-tuned transformer-based model (ALBERT) were better than those of the optimized NegEx, NegBio, and NegBERT.

Table 7 shows the performance evaluation of keyword extraction before and after applying the different negation and speculation-detection algorithms. The ALBERT method resulted in the most significant performance improvement in extracting keywords from positive and speculative statements potentially associated with abnormal findings.
Table 7. Comparison of the performance of keyword extraction in the test data set both before and after applying A Lite Bidirectional Encoder Representations From Transformers (ALBERT) and optimized NegEx, NegBio, and NegBERT.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F₁</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALBERT</td>
<td>0.998ᵃ</td>
<td>0.997ᵃ</td>
<td>0.997ᵃ</td>
<td>0.996ᵃ</td>
</tr>
<tr>
<td>NegEx</td>
<td>0.986</td>
<td>0.959</td>
<td>0.972</td>
<td>0.959</td>
</tr>
<tr>
<td>NegBio</td>
<td>0.934</td>
<td>0.958</td>
<td>0.945</td>
<td>0.917</td>
</tr>
<tr>
<td>NegBERT</td>
<td>0.99</td>
<td>0.998</td>
<td>0.994</td>
<td>0.991</td>
</tr>
<tr>
<td>Baselineᵇ</td>
<td>0.752</td>
<td>1.00</td>
<td>0.859</td>
<td>0.752</td>
</tr>
</tbody>
</table>

ᵃItalics highlight that the performance of ALBERT is the best comparing to the control method (NegEx, NegBio, NegBERT) and baseline (no negation or speculation detection were performed) across various performance metrics.
ᵇAll named entities considered “positive.” No negation or speculation-detection algorithm was applied.

Sources of Errors

Overview

We analyzed the sources of the errors (Table 8). Despite changes in the rules defined by the experts, errors persisted in NegEx and NegBio. We identified the following causes:

Table 8. Analysis of the causes of errors in different methods (after optimization).

<table>
<thead>
<tr>
<th>Method and cause of the wrong predictionᵃ</th>
<th>Counts, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NegBio (n=177)</strong></td>
<td></td>
</tr>
<tr>
<td>Errors in the extraction of named entities</td>
<td>58 (32.8)</td>
</tr>
<tr>
<td>Symbol-related errors</td>
<td>49 (27.7)</td>
</tr>
<tr>
<td>Tokenization error</td>
<td>21 (11.9)</td>
</tr>
<tr>
<td>Errors in the prediction of speculative statements</td>
<td>14 (7.9)</td>
</tr>
<tr>
<td><strong>NegEx (n=87)</strong></td>
<td></td>
</tr>
<tr>
<td>False-positive prediction related to speculative statements</td>
<td>37 (42)</td>
</tr>
<tr>
<td>Trigger word not triggered</td>
<td>21 (24)</td>
</tr>
<tr>
<td>Incorrect scope resolution</td>
<td>16 (18)</td>
</tr>
<tr>
<td>Symbol-related errors</td>
<td>6 (6)</td>
</tr>
<tr>
<td><strong>NegBERT (n=20)</strong></td>
<td></td>
</tr>
<tr>
<td>All false-negative predictions</td>
<td>16 (80)</td>
</tr>
<tr>
<td>All false-positive predictions</td>
<td>4 (20)</td>
</tr>
<tr>
<td>False-positive predictions related to speculative statements unrelated to abnormal findings</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>ALBERTᵇ (n=9)</strong></td>
<td></td>
</tr>
<tr>
<td>All false-positive predictions</td>
<td>5 (55)</td>
</tr>
<tr>
<td>False-positive predictions related to speculative statements unrelated to abnormal findings</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

ᵃThe table only list the most important causes of identifiable error.
ᵇALBERT: A Lite Bidirectional Encoder Representations From Transformers.

Findings of NegEx

First, we found many errors owing to incompatibility between the NegEx method for identifying speculative statements and the study requirements. NegEx made identical predictions for all keywords in the identified speculative statements regardless of their relevance to abnormal findings. However, our study categorized keywords in speculative sentences differently based on their relevance to abnormal findings, leading to discrepancies with NegEx’s results.

Second, the trigger word would only sometimes trigger. For example, in the phrase “1.No evidence of tumor,” the trigger word “No” would not be recognized because it was concatenated with the character “1.” without any intervening space.
Third, errors also occurred owing to the misinterpretation of the scope of negation and speculation, such as misinterpreting “No improvement of the tumor” as “No tumor.”

Fourth, errors occurred in the presence of symbols in radiology reports; for example, the use of special symbols by radiologists that are undefined in the trigger word or the confusion caused by the co-occurrence of special symbols that express a positive and a negative statement: (−) fatty liver and (+) portal vein thrombosis.

**Findings of NegBio**

We identified the following errors when using NegBio:

First, errors occurred in named entity extraction. The named entities in NegBio’s output file might be missing target keywords or had incorrect positions, resulting in incorrect future analyses.

Second, errors occurred when the radiology report contained negations using symbols or abbreviations, such as “metastasis (−).” Our analysis showed that these symbols could lead to unpredictable results in syntactic structure analysis and subsequent analyses.

Third, combining words with numerals or punctuation marks leads to errors in tokenization and subsequent analysis. For example, “1.No” in “1.No obvious acute infarct or brain metastasis” was not correctly parsed as “No.”

Fourth, many errors occurred because NegBio made identical predictions for diagnostic keywords in all speculative sentences, regardless of their relevance to abnormal findings. This behavior was inconsistent with the labeling of this experiment.

**Findings of NegBERT and ALBERT**

We observed the suboptimal performance of NegBERT when applied to corpora from different domains and tasks. The performance of NegBERT trained on the Simon Fraser University review corpus was suboptimal when evaluated on our corpus and task. Retraining NegBERT with our data significantly improved its performance, indicating that the poor performance was primarily due to differences in the training data and labeling.

Our error analysis showed that retrained NegBERT and ALBERT made fewer errors than the other methods in predicting whether words occurred in speculative statements unrelated to abnormal findings. The number of all false-positive predictions by NegBERT and ALBERT was 4 and 5, respectively. Both were lower than the number of false-positive predictions made by NegEx and NegBio for this prediction task, indicating higher specificity. However, because we grouped all negative and speculative statements not related to abnormal findings into the same category, we could not calculate the exact value of specificity. Both models showed 100% sensitivity in identifying important diagnostic keywords in speculative statements unrelated to abnormal findings, with no false-negative predictions.

Owing to the complexity of BERT, we could not further analyze the causes of other errors.

**Discussion**

**Principal Findings**

**Overview**

This study found that 39.45% (4529/11,480) of the important diagnostic keywords occurred in negative or speculative statements unrelated to abnormal findings, posing a challenge for automatic labeling by LISs and information extraction techniques.

Our study proposes a deep learning method that accurately distinguishes whether diagnostic keywords are in negative or speculative statements unrelated to abnormal findings. Our research has revealed the shortcomings of existing methods, including NegEx, NegBio, and NegBERT, while highlighting the advantages of our proposed approach over these methods.

**Limitation of NegEx and NegBio**

We observed common errors in Spacy’s NegEx and NegBio that the expert rule adjustment could not resolve.

First, several vital errors in NegEx and NegBio, including errors related to trigger words in NegEx, tokenization errors in NegBio, and symbol-related errors in NegEx and NegBio, were attributed to interference from punctuation and numerals. For example, in the radiology reports in our sample, English sentences were often combined with numbers and punctuation marks and written as numbered or bulleted lists, such as “1.No evidence of aortic dissection” In addition, using symbols or abbreviations in the form of checklists was also common. For example, “Metastasis (−)” or “Anti-HCV [Negative]” were frequently used. Our results showed that NegEx and NegBio could not handle this issue correctly.

Second, NegEx and NegBio also caused many errors in the analyses where the simultaneous observation of multiple sentences is required. Our data showed that it is often necessary to examine multiple sentences simultaneously to determine whether speculative statements are associated with abnormal findings. For example, in “No CT evidence of large infarct. Suggest MRI to exclude hyperacute infarct if indicated,” without considering the first sentence, which denies the finding of infarct evidence, it cannot be determined that the “hyperacute infarct” in the second sentence is unrelated to the actual findings. NegEx and NegBio, which are designed to analyze sentences in isolation without considering contextual information, cannot meet this requirement.

Our results regarding NegEx are consistent with previous research of Wu et al [42], highlighting the importance of tuning algorithms such as NegEx to achieve optimal performance in different corpora. Our results also confirm that NegEx produces incorrect results owing to improper negation scope resolution [22].

We found that NegBio requires modifying expert-defined rules to improve its performance. Our study is the first to report NegBio’s limited generalizability in real-world radiology reports across all body parts. We also observed problems with the implementation of NegBio.

---

https://medinform.jmir.org/2023/1/e46348
Limitation of NegBERT

Our experiment showed a significant improvement in NegBERT’s performance after retraining on our hospital data set. The difference in the training data and annotations is likely the reason for the initial poor performance of NegBERT.

This observation is consistent with previous findings that deep learning models such as BERT tend to perform poorly on out-of-domain corpora. For example, a study by Miller et al [39] using RoBERTa for negation detection on both in-domain and out-of-domain corpora observed \( F_1 \)-scores of 0.95 and 0.583, respectively. Our experiment supports this result and shows that the drop in \( F_1 \)-scores can be even worse depending on the corpus and task.

Advantages of ALBERT and BERT Transformer

We performed a comparison between the ALBERT and NegBERT methods and made the following key observations.

First, learning the negation cue and scope in 2 steps provides a limited performance improvement. Our method takes a different approach from NegBERT and traditional negation recognition studies in that our model learn the entire part of the sentence containing both the cue and scope in the same step without explicitly telling the model which word is the “cue” of the negation or speculation. However, the performance was still better than that of the retrained NegBERT. The study by Sergeeva et al [11] based on LSTM suggests that the deep learning method can learn negation cue information to some extent automatically, with performance comparable with that of automatic cue prediction algorithms. Our results show that BERT might have a similar capability. Our results suggest that providing additional cue information through expert annotation may not significantly improve performance compared with other factors, such as model selection, hyperparameter optimization, and training techniques.

Second, our results show that the model size and complexity do not necessarily correlate with improved performance. In our study, the fine-tuned ALBERT model outperformed larger and more complex models, including BERT and XLNet used by NegBERT, as well as RoBERTa used in the study by Miller et al [39]. The use of lightweight models, such as ALBERT, may have practical advantages, including reduced computational resource requirements and training time, compared with BERT [31].

In our study, ALBERT and retrained NegBERT outperformed NegEx and NegBio in terms of the number of false-positive predictions and specificity while maintaining 100% sensitivity in predicting whether keywords occurred in speculative sentences unrelated to abnormal findings. This task required multisentence context analysis of our data set, and our results suggest that BERT can look at multiple sentences simultaneously. The attention mechanism is a reasonable explanation for this phenomenon.

Comparison With Prior Work

Our study fine-tuned the ALBERT model using a more comprehensive data set that included a broader range of imaging modalities and subspecialties than previous studies. Table 9 shows the best performances and corresponding data sets used in previous studies that detected whether named entities occurred in negation and speculation in radiology reports. The range of imaging modalities and subspecialties represented in the radiology reports in these studies was limited, such as chest x-ray reports only in the study by Peng et al [23] or brain CT and MRI reports only in the studies by Grivas et al [43] and Sykes et al [12]. We hypothesized that including a more diverse set of examination and imaging subspecialties in the data results in a more representative sample of the report content and improves the model’s generalizability. Our results support this hypothesis, as the ALBERT model showed only a 0.034 decrease in its \( F_1 \)-score on an unseen test data set with different disease types and inputs from different physicians.

Our experiments also address a more difficult speculation-detection task than previous studies; however, ALBERT still demonstrates good performance. This distinction requires the ability of the algorithm to consider multiple sentences simultaneously in our data set. To the best of our knowledge, our study is the first to propose a distinction between speculative sentences related and unrelated to abnormal findings based on the application scenario to facilitate more precise filtering and the first study to highlight the impact of the lack of multisentence analysis in negation detection algorithms.
# Table 9. Comparison of best performances between studies distinguishing whether named entities occurred in negation or speculation.

<table>
<thead>
<tr>
<th>Study</th>
<th>Algorithm</th>
<th>(P^a)</th>
<th>(R^b)</th>
<th>(F_1)</th>
<th>Best Type</th>
<th>(N^d)</th>
<th>Task Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our study</td>
<td>ALBERT(^e)</td>
<td>0.991</td>
<td>0.992</td>
<td>0.991</td>
<td>Test data set</td>
<td>6000</td>
<td>ND(^h)+S(^i)</td>
</tr>
<tr>
<td>Sykes et al [12]</td>
<td>BiLSTM(^j)</td>
<td>0.973</td>
<td>0.981</td>
<td>0.977</td>
<td>ESS(^k)</td>
<td>630</td>
<td>ND+S(^l)</td>
</tr>
<tr>
<td>Peng et al [23]</td>
<td>NegBio</td>
<td>0.944</td>
<td>0.944</td>
<td>0.944</td>
<td>Chest x-ray</td>
<td>900</td>
<td>ND+S(^m)</td>
</tr>
<tr>
<td>Grivas et al [43]</td>
<td>Edie-R</td>
<td>0.925</td>
<td>0.943</td>
<td>0.934</td>
<td>ESS (^k)</td>
<td>630</td>
<td>Brain CT and MRI</td>
</tr>
</tbody>
</table>

\(^a\)P: precision.  
\(^b\)R: recall.  
\(^c\)Name of the best-performing data set. Other data sets are not included.  
\(^d\)Number of samples in the best-performing data set; other data sets are not included.  
\(^e\)Task performed in the study.  
\(^f\)Types of radiologic studies included in the study.  
\(^g\)ALBERT: A Lite Bidirectional Encoder Representations From Transformers.  
\(^h\)ND: negation detection.  
\(^i\)S*: detection of speculation unrelated to abnormal findings.  
\(^j\)BiLSTM: Bidirectional Long Short-Term Memory.  
\(^k\)ESS: Edinburgh Stroke Study.  
\(^l\)S: speculation detection.  
\(^m\)CT: computed tomography.  
\(^n\)MRI: magnetic resonance imaging.

## Implication in Clinical Practice
We found problems with NegEx and NegBio in that modifying expert-defined rules could not be solved, including difficulties with numbers and punctuation, implementation-specific challenges, and the design constraint of observing only a single sentence at a time; thus, NegEx and NegBio should be used cautiously or avoided in such situations to prevent errors. On the basis of our data, we also found that NegBio and NegBERT have limitations in generalizability, making them inappropriate for use without training or modeling.

Our results indicate that BERT is more suitable than NegEx and NegBio for tasks involving multisentence context analysis, similar to the experiment conducted in this study. NegEx and NegBio were designed for single-sentence analysis because they segmented the text into independent sentences. This approach limits the ability to incorporate contextual information from other sentences into the analysis. While NegEx and NegBio can perform binary classification of words in sentences as speculative or not, they lack the capacity for further granular differentiation based on contextual information.

We found that the training process of the transformers did not require 2 separate learning phases for cue and scope. Our findings could reduce the workload of expert annotation in clinical applications, as the explicit annotation of cues in a separate step requires additional work. This hypothesis needs further testing in future studies.

Our results show that deep learning models outperform non–deep learning methods, and lightweight models such as ALBERT can achieve superior performance and outperform other transformer-based models. However, fine-tuning based on the specific domain corpus and task is still essential regardless of the model used.

## Limitations
The data were obtained from 3 internal branches of a single institution and not from publicly available data sets. In addition, the speculation-detection task differed from previous studies in this area. The comparability of the performance with that of previous studies may be limited. If open data using the same annotation methodology become available, subsequent research could verify our findings by implementing the same model on the open data set.

Our study optimized the control methods (NegEx and NegBio), but we cannot exclude the possibility of further performance improvement by modifying or adding expert rules. However, this highlights the limitations of an expert rule–based approach, which requires experts not only to detect negations and speculations but also to summarize and modify rules manually. Moreover, expert rules cannot resolve the algorithmic design or implementation constraints.

To prevent the deep learning model from training failure, we combined negative statements with speculative statements unrelated to abnormal findings in the same category because of the low proportion of the latter. As a result, we cannot separately evaluate the model’s performance on negative and speculative sentences unrelated to abnormal findings or accurately quantify the latter’s performance. Nevertheless, metrics such as the number of false-positive predictions can still be used to compare the performance between methods.

## Conclusions
Manual free-text reporting remains the norm in radiology worldwide, hampering the ability to perform computer-assisted...
analyses. The presence of information irrelevant to the actual findings poses a significant challenge to the implementation of automatic radiology report highlighting, flagging, or information extraction.

Previous research on negation and speculation detection in radiology has aimed to identify all instances. Our study advances this by targeting only speculative statements unrelated to abnormal findings and improving the discrimination of relevant information using BERT’s multisentence contextual analysis capabilities.

Lightweight transformer models, such as ALBERT, can outperform NegEx, NegBio, and NegBERT on more complex and diverse real-world radiology reports. Despite achieving good results on public data sets, NegBio and NegBERT demonstrated different performances on more complicated real-world radiology reports.

Our research has potential applications in academia and clinical practice. Future studies may consider including lightweight models such as ALBERT. In clinical practice, our method achieved high performance. It can help algorithms such as keyword highlighting in hospital information systems to identify passages of potentially important information without false alarms, improving physician efficiency and health care quality. Our results also apply to radiology report information retrieval, such as search engines, in which negative and speculative statements unrelated to abnormalities can lead to incorrect results.

Acknowledgments
This research received no specific grants from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contributions
KHW proposed the research topic and experimental design and completed the recruitment, data analysis, computer programming, and writing of the entire paper.
CFL contributed to research design improvement and manuscript proofreading.
CJC performed big medical data exporting and cleaning and manuscript proofreading.

Conflicts of Interest
None declared.

References


Abbreviations

ALBERT: A Lite Bidirectional Encoder Representations From Transformers
BERT: Bidirectional Encoder Representations From Transformers
BiLSTM: Bidirectional Long Short-Term Memory
CRF: Conditional Random Field
CT: computed tomography
DeBERTa: Decoding-Enhanced Bidirectional Encoder Representations From Transformers With Disentangled Attention
DEEPEN: Dependency Parser Negation
DistilBERT: Distilled version of Bidirectional Encoder Representations From Transformers
ELECTRA: Efficiently Learning an Encoder That Classifies Token Replacements Accurately
ERNIE: Enhanced Representation through Knowledge Integration
LIS: laboratory information system
LSTM: Long Short-Term Memory
MRI: magnetic resonance imaging
RoBERTa: Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach
Understanding Views Around the Creation of a Consented, Donated Databank of Clinical Free Text to Develop and Train Natural Language Processing Models for Research: Focus Group Interviews With Stakeholders

Natalie K Fitzpatrick¹, MSc; Richard Dobson², PhD; Angus Roberts², PhD; Kerina Jones³, PhD; Anoop D Shah¹,⁴, PhD; Goran Nenadic⁵, PhD; Elizabeth Ford⁶, PhD

¹Institute of Health Informatics, University College London, London, United Kingdom
²Department of Biostatistics and Health Informatics, King’s College London, London, United Kingdom
³Department of Population Data Science, Swansea University Medical School, Swansea, United Kingdom
⁴University College London Hospitals NHS Foundation Trust, London, United Kingdom
⁵Department of Computer Science, University of Manchester, Manchester, United Kingdom
⁶Brighton and Sussex Medical School, Brighton, United Kingdom

Corresponding Author:
Natalie K Fitzpatrick, MSc
Institute of Health Informatics
University College London
222 Euston Road
London, NW1 2DA
United Kingdom
Phone: 44 7808032697
Email: n.fitzpatrick@ucl.ac.uk

Abstract

Background: Information stored within electronic health records is often recorded as unstructured text. Special computerized natural language processing (NLP) tools are needed to process this text; however, complex governance arrangements make such data in the National Health Service hard to access, and therefore, it is difficult to use for research in improving NLP methods. The creation of a donated databank of clinical free text could provide an important opportunity for researchers to develop NLP methods and tools and may circumvent delays in accessing the data needed to train the models. However, to date, there has been little or no engagement with stakeholders on the acceptability and design considerations of establishing a free-text databank for this purpose.

Objective: This study aimed to ascertain stakeholder views around the creation of a consented, donated databank of clinical free text to help create, train, and evaluate NLP for clinical research and to inform the potential next steps for adopting a partner-led approach to establish a national, funded databank of free text for use by the research community.

Methods: Web-based in-depth focus group interviews were conducted with 4 stakeholder groups (patients and members of the public, clinicians, information governance leads and research ethics members, and NLP researchers).

Results: All stakeholder groups were strongly in favor of the databank and saw great value in creating an environment where NLP tools can be tested and trained to improve their accuracy. Participants highlighted a range of complex issues for consideration as the databank is developed, including communicating the intended purpose, the approach to access and safeguarding the data, who should have access, and how to fund the databank. Participants recommended that a small-scale, gradual approach be adopted to start to gather donations and encouraged further engagement with stakeholders to develop a road map and set of standards for the databank.

Conclusions: These findings provide a clear mandate to begin developing the databank and a framework for stakeholder expectations, which we would aim to meet with the databank delivery.

(JMIR Med Inform 2023;11:e45534) doi:10.2196/45534
KEYWORDS
consent; databank; electronic health records; free text; governance; natural language processing; public involvement; unstructured text

Introduction

Background
Electronic health records (EHRs) contain a rich narrative of the patient journey and have huge potential for research [1]. However, research using EHRs is typically limited to the structured data (such as numerical values and diagnoses coded using a controlled vocabulary), despite a large proportion of the information in EHRs being in the form of unstructured (free) text. The analysis of free text at scale requires specialized tools and methods (natural language processing [NLP]) to “read,” process, and structure the information before it can be used at scale for research purposes.

NLP of clinical text has many potential benefits, both for individual care and improving health services [2]. These include (1) to facilitate the process of clinical coding [3], which is the process by which clinical coding staff in hospitals assign codes from a specific terminology (eg, International Classification of Diseases, 10th revision [ICD-10]) [4] to patient episodes for reimbursement; (2) to facilitate structured recording of diagnoses in clinical care using Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [5], which is currently not done consistently [6]; and (3) to enable research using information in EHRs, which are currently not coded. Compared with manual review of free text, automated analysis is much faster and enables a much larger amount of text to be analyzed, enabling larger and more representative patient samples to be used for research.

The Challenge
Tools and platforms for text access and analysis such asCogStack (developed by a consortia of scientists at King’s College London, King’s College Hospital, South London and the Maudsley, Guy’s and St Thomas’s hospital, University College London, and University College London Hospitals National Health Service [NHS] Foundation Trust and some members of the CogStack open-source community) [7] have been developed and installed at some NHS sites with great success, but overall, access to free text for researchers is still currently difficult. Ideally, free text needs to be brought out of the NHS environment so expert computer scientists working in university research or other non-NHS environments can use the data to train their computer algorithms to extract the important clinical information. In the United Kingdom, the application of NLP for health care text research is largely limited to within large NHS hospital trusts with academic affiliations and in-house NLP expertise owing to complex governance requirements arising from increased concerns around the potential risk of reidentifying patients. In Scotland, a successful model adopted by groups including the Health Informatics Centre at the University of Dundee [8], DataLoch [9], and the national electronic Data Research and Innovation Service [10] in collaboration with the University of Edinburgh Clinical NLP Research Group [11] involves the provision of data for research through secure trusted research environments outside NHS or university settings. Nind et al [12] describe an approach for extracting, linking, deidentifying, and hosting clinical imaging data within a controlled secure environment as a resource for national and international research. This model provides a potential alternative approach to hosting the databank outside of an NHS or university setting and allows for timely and secure access to data; however, the governance framework is complex. In the United Kingdom, before medical data can be shared outside of the NHS environment, identifiers such as names of patients, family members, and health care professionals; addresses; and dates of birth, which can occur anywhere in the text, first need to be removed—a process known as “deidentification” [13]. Even when deidentified, there remains a risk that some identifiable information may have been missed, third parties might be identified, or the narrative may be too revealing.

Routinely collected health data are legally accessed for secondary purposes such as research by 2 lawful bases under UK data protection law: one is the principle of informed consent from the patient and the other is “task in the public interest” [14]. For processing under the lawful basis of “task in the public interest,” health care data needs to be deidentified or anonymized before it can be shared outside the clinical environment under the General Data Protection Regulation principle of data minimization [15]. This is where governance becomes difficult, as deidentifying free-text clinic notes, letters, and reports is complex and a rapidly evolving field, and the accuracy of the process is hard to assess [2]. Technology exists to automatically redact identifying information so that only deidentified documents are presented to computer scientists to develop NLP [1]. However, the reidentification risk from automatically deidentified text remains unknown. Many independent health research ethics committees (RECs) do not have the specialized technical knowledge needed to evaluate the risks posed to patients by this type of research, and indeed, many researchers and data custodians are not sure of the scrutiny and approvals needed to legally process free text for research. As a consequence, a conservative approach is usually taken, resulting in heavy restrictions on data access [16]. Therefore, there are currently very few health care free-text data sets available to NLP researchers to develop and evaluate their algorithms; 1 example is the Medical Information Mart for Intensive Care (MIMIC) database [17,18] in the United States. MIMIC is based on a selected patient population (critical care patients) from 1 US center and contains structured and unstructured (eg, diagnostic reports and physician notes) deidentified data linked to hospital EHR and mortality data. Data that will contribute to the databank remain to be decided and will follow further consultation with stakeholders, but at the very minimum, it will include unstructured text from primary care and hospital records for a defined population of patients in the United Kingdom.
The Solution

One possibility for breaking down barriers to access to free-text data is to enable access to clinical text via the lawful basis of informed consent. Creating a “donated” databank of clinical free text in which each patient represented has given informed and explicit consent for their data to be used in this way could provide an important and timely opportunity for NLP researchers to develop and train NLP algorithms to process the free text, which can then be used on other data sets in the NHS to conduct clinical research. NLP researchers in universities or other non-NHS settings only need to access a sample of patient free-text data to develop and train the NLP algorithms on the databank, which could then be run on unseen patient free text housed within the NHS for research, avoiding the important privacy issues laid out above.

To test early thinking on the databank, we carried out a series of in-depth focus groups among 4 key stakeholder groups to find out what key stakeholders think about a consented, donated databank of clinical free text to help create, train, and evaluate NLP for clinical research.

Methods

Participant Selection and Inclusion Criteria

Four stakeholder groups were identified based on their potential interest and investment in the databank as follows: (1) patients and members of the public, (2) clinicians (NHS general practitioners [GPs], hospital doctors, and doctors in training), (3) NHS Trust information governance (IG) leads and REC members, and (4) NLP researchers based in universities or NHS hospitals. Participants lived in the United Kingdom and were aged ≥ 18 years. Patients and members of the public were based in the community and had to have some previous knowledge or understanding of the use of free-text health data for research; for example, they may have had attended events or workshops on this topic or had experience participating in advisory committees on the use of free text.

Participant Recruitment

Patients and members of the public were recruited via an advert posted by existing networks, including Health Data Research United Kingdom [19] and the National Institute for Health Research People in Research network [20]. Other stakeholders were identified via existing national networks, contacts, and organizations and were approached directly by email by the research team. In addition, IG leads were identified via the Office of the National Data Guardian [21] and by searching the websites of NHS Trusts and Health Boards. NHS Research Ethics Service committee members were identified by searching the NHS Health Research Authority website [22], and academic NLP researchers were identified via professional networks including the UK health care text analytics network known as Healthex [23].

Potential participants were invited by the research team by email to participate in 1 of the 4 relevant stakeholder focus groups. Before deciding whether to take part, participants were asked to read a study information sheet and return a completed expression of interest form recording basic demographic information including age category (deciles), gender identity, and ethnicity. To understand participants’ views before taking part in the study, they were also asked to indicate how comfortable they might feel about donating their health data for the purposes of the databank outlined in the participant information sheet from 1 of the following categories: very comfortable, somewhat comfortable, not sure, somewhat uncomfortable, or very uncomfortable. Invites were then sent a consent form by email, which they were asked to complete and sign. Patient and public members were offered a modest financial incentive for participating in the study in line with National Institute for Health Research guidance for recognizing public participation in research [24].

Focus Groups

Focus groups were conducted on the web on Zoom between March 24 and 31, 2022, and lasted for 90 minutes. A deliberative approach was used where focus groups began with a short presentation on the donated databank by a member of the research team, tailored to each stakeholder group, followed by a question-and-answer session so discussions could be fully informed. The proposed model presented to participants in the prediscussion presentations was of an opt-in approach where people would consent to donate their data to the databank. The facilitator did not direct discussions to confirm whether donated data would be identified or not so that participants could freely share their views around both scenarios.

The team employed a third-party organization with considerable experience in conducting focus groups on the topic of health data to facilitate the groups. Discussions were framed around 4 key questions: (1) Is having a donated free-text databank a good idea? (2) How best could the risks of holding donated, consented potentially identifiable data be managed? (3) What do you think about consent, and how it could be managed? and (4) Who should be allowed access, how should a databank be housed, and for what purposes? Multimedia Appendix 1 presents the questions asked of each stakeholder group.

Discussions were audio recorded, transcribed, and analyzed using thematic analysis.

Ethics Approval

The study was approved by the University College London REC (0976/002) and complies with the COREQ (Consolidated Criteria for Reporting Qualitative Research) [25] checklist for reporting qualitative studies.

Results

Overview

A total of 61 participants took part in the focus groups including patients and members of the public (24 participants), clinicians (10 participants), NHS Trust IG leads and REC members (14 participants), and NLP researchers (13 participants).

In total, 75% (46/61) of the participants recorded their demographic information on their expression of interest form. Of those, 54% (25/46) were female, 52% (24/46) were aged between 31 and 50 years, and 73% (33/45) were White. Overall, most participants (30/46, 66%) were either very comfortable or
somewhat comfortable donating their data to the databank, 28% (13/46) were not sure whether they would be willing to donate their data, and only 6% (3/46) of all participants felt either very or somewhat uncomfortable donating their data, all of whom were patients and public members, although the numbers were small (3 patients and public members out of 17; Table 1).

Key findings of the study are summarized in Multimedia Appendix 2.

Table 1. Participant demographic information and views around sharing their own data captured before participating in the focus groups.

<table>
<thead>
<tr>
<th></th>
<th>All participants (n=46^a^), n (%)</th>
<th>Patients and public members (n=17^a^), n (%)</th>
<th>Clinicians (n=10^a^), n (%)</th>
<th>Information governance leads and research ethics committee members (n=14^a^), n (%)</th>
<th>Natural language processing researchers (n=5^a^), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>25 (54)</td>
<td>10 (59)</td>
<td>3 (30)</td>
<td>9 (64)</td>
<td>3 (60)</td>
</tr>
<tr>
<td>Male</td>
<td>21 (46)</td>
<td>7 (41)</td>
<td>7 (70)</td>
<td>5 (36)</td>
<td>2 (40)</td>
</tr>
<tr>
<td>Intersex</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Other</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Age group (years)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤30</td>
<td>3 (7)</td>
<td>2 (12)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>1 (20)</td>
</tr>
<tr>
<td>31-50</td>
<td>24 (52)</td>
<td>4 (24)</td>
<td>8 (80)</td>
<td>8 (57)</td>
<td>4 (80)</td>
</tr>
<tr>
<td>51-65</td>
<td>12 (26)</td>
<td>6 (35)</td>
<td>2 (20)</td>
<td>4 (29)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>&gt;65</td>
<td>7 (15)</td>
<td>5 (29)</td>
<td>0 (0)</td>
<td>2 (14)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian or Asian British</td>
<td>7 (16)</td>
<td>3 (18)</td>
<td>3 (30)</td>
<td>0 (0)</td>
<td>1 (20)</td>
</tr>
<tr>
<td>Black, Black British, or African</td>
<td>2 (4)</td>
<td>2 (11)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Mixed or multiple ethnic groups</td>
<td>1 (2)</td>
<td>1 (6)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>White</td>
<td>33 (73)</td>
<td>11 (65)</td>
<td>6 (60)</td>
<td>13 (100)</td>
<td>3 (60)</td>
</tr>
<tr>
<td>Other ethnic group</td>
<td>2 (5)</td>
<td>0 (0)</td>
<td>1 (10)</td>
<td>0 (0)</td>
<td>1 (20)</td>
</tr>
<tr>
<td><strong>Views around donating data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I would feel very comfortable donating my data</td>
<td>15 (33)</td>
<td>8 (47)</td>
<td>4 (40)</td>
<td>3 (21)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>I would feel somewhat comfortable donating my data</td>
<td>15 (33)</td>
<td>5 (29)</td>
<td>3 (30)</td>
<td>6 (43)</td>
<td>1 (20)</td>
</tr>
<tr>
<td>Not sure</td>
<td>13 (28)</td>
<td>1 (6)</td>
<td>3 (30)</td>
<td>5 (36)</td>
<td>4 (80)</td>
</tr>
<tr>
<td>I would feel somewhat uncomfortable donating my data</td>
<td>2 (4)</td>
<td>2 (12)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>I would feel very uncomfortable donating my data</td>
<td>1 (2)</td>
<td>1 (6)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

^a^Data are based on 75% (46/61) of the participants who returned this information. Within stakeholder groups, data were returned as follows: patients or public members, 71% (17/24); clinicians, 100% (10/10); information governance leads or research ethics committee members, 100% (14/14); and natural language processing researchers, 38% (5/13).

Is Having a Donated Free-Text Databank a Good Idea?

**Perceived Benefits and Challenges**

Participants were very enthusiastic about the databank and its intended purpose and saw great value in establishing a platform for development and testing NLP tools to improve their accuracy. Many participants across groups articulated the benefits of producing trustworthy tools to unlock the rich data available in free text, which extended beyond improving NLP methods, including expediting access to, and use of, NHS data by speeding up permissions; accelerating development of NLP tools; and improving health and care leading to better outcomes for patients. The NLP researcher group highlighted its potential value as a training resource to teach and onboard researchers and help familiarize them with free-text data. NLP researchers and clinician groups both welcomed the opportunity to access free-text data for a UK-based population, which would be more appropriate for developing NLP tools on UK health care data sources, moving away from a reliance on US-based data such as MIMIC III or the recently released MIMIC IV [17,18]:

*...I'd love to see it come to fruition. I think it would be an absolute gold mine.* [IG lead and REC member]

*Don’t let this racehorse designed by [a] committee become a camel, just get something out. I think...*
anything is better than what's currently offered, which is nothing. [NLP researcher]

Several participants in the patient and public group felt that increased access to patient data as a result of the databank may prompt clinicians to improve the quality of their free-text data entry, as they will be more conscious of its wider use:

Very much a great idea. So, the MIMIC dataset I've worked with a lot has been really transformative for clinical NLP research in the US. But the MIMIC dataset has some serious issues, in terms of the kinds of data that are included, the representativeness of the sample, and so on, and so forth. So having something that can be created, as a research-specific resource like this, and created with more intentionality, and more design, as to what should be going into it. I think is a really, really incredibly valuable thing to do. [NLP researcher]

Despite strong support, participants in all groups advised that clarity around the purpose of the databank; how it will be used; and by whom, both now and in the future, will be essential to its success. Patient and public participants felt this should be made clear in the consent process. Many participants, but patients or public members in particular, expressed concerns around how inaccurate recording, the lack of up-to-date data, or subjective data based on a physician’s own impressions may threaten the aims of the databank, citing frequent instances of errors in their own health records. IG leads and REC members and patients or public participants felt that the accuracy of data must be improved before people trust the outcomes of the databank, and patients or public participants were keen for easier access to their own EHR so they can amend inaccuracies or missing information. However, other participants did not feel data accuracy would be a key factor in the success of the databank, and NLP researchers suggested the databank could provide a unique opportunity for investigating the effect of inaccurate or subjective training data on NLP research findings.

Biases owing to missing data or the lack of generalizability was a considerable concern among all groups. Many felt that donations will be more likely to come from White, middle-class populations and less likely to include people with rare diseases or whose records contain sensitive information. This was seen as most likely to affect data in mental health and social care settings:

There may be intrinsic biases in the actual data that’s getting selected because certain classes of patients, whether it’s by demographics, such as race or income, have more trust in what this is trying to achieve and, therefore, people with less trust won’t actually consent to their data being used and we know, for instance, that that can be quite heavily in race, in the UK, on health data, for instance, and health services and the provision of health services. So, that may introduce biases in the data set. [NLP researcher]

I’m not saying it would be necessarily unprofessional, but there could be things that may have been written 10 years ago and that maybe wouldn’t be written now. Is that going to affect your data sets? So, I guess it’s really about the bias that might be there within the unstructured data and whether you’re proliferating that bias by collecting them and then training algorithms. [NLP researcher]

Many participants saw artificial intelligence (AI) as playing a key role in health care in the future but were concerned about how AI tools are developed and perform in general. These concerns, which included questions around how tools “decide” which words to analyze, the possibility of scan reports being misread or missing key data, and the accuracy of annotation, are key to feeding into the communication plans for the databank.

Data Privacy and Use

Fears around data privacy for both patients and clinicians were raised. Participants discussed complexities in relation to free text, which might act as a barrier to data donation, and how clinicians may be uncomfortable that their identity and views are shared with researchers, for example, where GPs’ personal views around a patient’s health are recorded. Participants advised that such fears might be mitigated to some extent by ensuring robust data security, governance, and transparency around the “data pipeline”—that is, what the databank will be used for and by whom, for example, whether there is any commercial benefit. IG leads and REC members and patients or public participants in particular discussed challenges in articulating how tools may be used in the future and by whom, as technology evolves and society’s views around acceptable and ethical use of their data may change over time. Participants thought that building scenarios for future use into the consent process is therefore important:

My only worry or concerns would be the way that technology develops way into the future and therefore, algorithms, as a result. And it could be ethical now but maybe it would be less ethical in the future. [Patient or public member]

Types of Data to Include in the Databank and Data Linkages

All participants felt that it is very important that the databank benefit from the inclusion of data from a range of sources to reflect the whole systems approach of the health service and a more integrated care system of the future.

I don’t see how, at this stage, you can start to select what you want to look at because you don’t know what you want to look at. If the purpose of this is to develop algorithms for extracting useful, contextual information then you want the original data, which is being used to train the algorithm, to be as broad as possible. [IG lead and REC member]

We are talking about a holistic approach, so that vision should be, I totally agree, whole patient records. And if we are going to use the computers, and train the algorithm, the whole purpose is looking at the wider picture, and bringing it together. [Clinician]
Along with primary and secondary care data, some saw value in training NLP tools based on sources that are less commonly used for research, including social services and housing data, to encourage more research in these fields. Perceived advantages of including a broad range of data include introducing a more holistic view of a person’s health needs and including training data based on the different styles of free-text data held in different settings. Some participants discussed the advantages of introducing the databank in phases to ensure timely access and build trust, for example, by starting with 1 health condition such as diabetes or mental health or geographical location.

Such phasing would need to be carefully considered. If condition-specific phasing were to be used, then the imperfections of the diagnostic codes used to select data for a phase would need to be addressed. Solutions might be to use existing NLP applications to identify conditions, accepting that this would have its own limitations, or choosing broader categories of data such as data from specific medical specialties or units:

> Our health is based in our experience, and what starts in primary, might end up in secondary. You can’t divide them into two separate things. [Patient or public member]

There were mixed views around linking free-text data to other forms of data. NLP researchers particularly valued the opportunity to link the data to other sources, especially mental health data where rich narrative adds important additional detail for research, and some felt that the inclusion of coded data could help verify the accuracy of the free-text data more efficiently. However, some questioned the value of such linkages to train NLP tools and felt linkages, for example, to administrative data, may increase the risk of reidentification and would be resource intensive to manage:

> In psychiatry, it tends to be that a lot of the information is locked behind this clinical free text, and they tend to be, by comparison to other medical specialties, a bit more verbose, a bit more narrative, in nature. And so, the ability to incorporate that, and the metadata that’s required to have that, needs to be built in. I think, from the very, very start this has to be extensible, because I’d love to, for example, be able to look at GP notes, and see how they translate over into secondary care, but there are decisions about how this may be structured, early on, that could make that more difficult. So I think that’s something that we need to build in from the start. [NLP researcher]

IG leads and REC participants discussed the potential for linkages resulting in “scope drift,” which could lead to the use of data outside of its original purpose. If this occurs, the databank should clearly communicate its wider remit. Scope drift was also a concern for linkages to patient-generated data such as wearable and monitoring devices. Some questioned the accuracy of these data sets and whether this may lead to the development of inaccurate NLP tools, although potential advantages including signaling support for greater use of patient data and “future-proofing” development of NLP tools by including these data, given the inevitable advancement and incorporation of these devices in health care, were also discussed:

> This is where, as an IG person, I always start getting nervous and we’re having conversations in our population health because everybody starts saying, “The police, it would be really good if we can put an algorithm together to identify individuals that might start being people that will cause domestic harm.” So, that’s an area that is always a bit nervous to say, “Well, what is that trying to establish,” especially with this project. Because what’s the purpose of having a linked dataset and then trying to do modelling on trying to extract free-text data? I’m not quite sure of the purpose of those two things together. [IG lead and REC member]

Managing Risks of Holding Donated, Consented Identifiable Data

Data Privacy and Deidentification

The groups did not discuss a preference for whether data should be left in an identifiable form or should be deidentified; rather, discussions focused broadly on the issue of how to minimize the risk of reidentification, indicating that participants expected data to be deidentified. In particular, the patient or public group highlighted the importance of deidentification alongside data security, including robust data storage and management practices to mitigate risks to data privacy. All participants recognized that eliminating risk completely is unrealistic, but introducing steps to reduce the likelihood of reidentification by using birth year, partial postcode, or a sample of the notes rather than the whole record was discussed. Some clinician participants were conscious of potential legal implications for themselves and supported the removal of clinician details as well. Participants were mindful that the process of deidentification should be carried out on a case-by-case basis. NLP researcher members suggested involving data controllers in agreeing with the approach to deidentification, given their expert knowledge of the data and availability of resources such as data dictionaries to aid the deidentification process. Although rare diseases were regarded as posing a particular risk to reidentification, patients or public participants who themselves have been diagnosed with a rare condition were keen that this should not act as a barrier to the much-needed research and proposed that clear explanations about how data would be used, by whom, and what the data protection issues are might offset concerns. Alongside deidentification, it was felt that following the UK Caldicott Principles, which helps ensure confidential and appropriate use of people’s data [26], adopting strong data security measures to protect against hacks and ensuring that only legitimate, vetted people have access to only the data they need were viewed as key to managing privacy concerns. Patients or public group members in particular voiced the importance of articulating these safeguards to reassure potential donors, and IG leads and REC members suggested that because deidentification is both difficult to define and achieve, there should be an emphasis on defining the purpose of accessing the data and the methods of safeguarding the data:
Is it a pseudonym? Is it a number? Is it aggregated? What level of anonymity is there when we’re discussing this? What’s proposed, or are there different levels (to de-id)? [Patient or public member]

If you’re doing a decent level of de-identification so you’re getting masking rates of 90% or something, the risk is going to be very minimal, particularly if you’re providing samples of notes rather than an entire record level. So if you’re looking at medication as your concept for annotation, then you don’t need stuff that doesn’t contain that data, or is unlikely to. So you can start to pick how you pull your sample notes from a record. I think there’s quite a lot you can do to continue to reduce and reduce and reduce that risk, but you won’t eliminate it. [IG lead and REC member]

Raising Awareness of the Databank

Future consultation with stakeholders, including clinicians, was viewed as essential by all groups, who were keen that engagement be continued to help develop the scope, consent model, and communication plan for the databank. Several participants advised careful planning on how to explain the databank. Patients or public participants expressed the importance of terminology when communicating plans, for example, to clarify what is meant by free text and what is in a health care record that might help allay people’s fears around what they are agreeing to donate:

I want some power. I don’t want to be a passive recipient of this whole data process, which is what’s happened to a lot of us regarding data so far. [Patient or public member]

Participants suggested that a targeted, small-scale approach be adopted in the early stages of raising awareness of the databank and starting to gather donations, working with trusted organizations who use health data for research. The involvement of GP practice staff in communicating the databank was seen as important, but clinician participants in particular were skeptical, saying that this was both impractical given their already stretched resources and unnecessary given patients already have the right to agree to their health records being shared. One alternative discussed was to recruit GPs or practice staff who were willing to take on this role. Some ideas for how to reach potential donors included making posters and leaflets available in GP surgeries with a QR code linked to a website about the databank, working with trusted organizations that support the use of health data for research such as HealthWise [27] and use MY data [28], and identifying community-based “public champions” to advocate the benefits and safety of the databank:

I would share it with patients but I’m constantly limited by time—something else to include in the consultation, so it has to be done through a different delivery mechanism than face to face. [Clinician]

Managing Consent and Offering Choice

Participants expressed the importance of working with trusted NHS and research organizations and providing accessible consent including web-based consent, the availability of audio consent for people with low literacy levels, and translation into multiple languages. The information sheet should clearly set out the intended purpose of the databank and focus on providing reassurance regarding use and data security. If the approach was to include deidentified data, participants wanted this made clear in the supporting materials, possibly aided by including examples of “dummy” data to show what types of data are excluded in the deidentification process. Patients or public members thought that offering people the opportunity to view the databank before consenting might help them understand what information would be included:

People should know what they’re getting involved in. Maybe seeing the database before they opt-in, and also, how far back will the data be taken from… [Patient or public member]

Having an actual example of a fake letter or fake clinical notes with a lot of identifiable data and what, actually, will go to the databank in front of you so you see that they did remove all information about you talking about your kids or how your neighbour is annoying you because she is very noisy, all that kind of thing. Maybe, they would see it, that would make a lot of reassurance. [NLP researcher]

Patients or public participants felt that offering choice over which data to donate and which type of organization can access the databank might increase donation rates, but others thought this may be complex and resource intensive to achieve and may encourage withholding of sensitive information. The IG leads and REC members group felt that offering choice with a promise to withhold sensitive information is likely to be unachievable and, therefore, undermine trust.

Who Should Be Allowed Access, How Should a Databank Be Housed, and for What Purposes?

Overview

Participants were in favor of access to the databank by both university- and NHS-based NLP researchers. Establishing a “road map” of the types of organizations with which the databank will work was suggested as helpful. Many participants favored a defined approach to access in the early stages of developing the databank, whereby access should be limited to NHS and university-based NLP researchers, and developing a set of standards could ensure that the use of the databank remains in line with its intended purpose. However, the IG leads and REC members group felt that it may be more appropriate to define the types of organizations that can access the databank based on a “compliance” model in which users should show that they can meet a defined level of capability and accountability. Standards should include an assessment of the applicant organization’s motivation for using the databank and their reputation, although the road map should include a well-thought-out vetting process to include applicants such as start-ups who may not have a proven track record in trustworthy access to data:

If there were particular requirements around use of the data, commitments to not attempt to re-identify,
similar things like that, I think in my view both academic organisations, research students and also research organisations, with the right safeguards, even start-ups, I feel if they can meet a certain level of capability and accountability, then I feel that should be the bar rather than defining the type of organisation. [IG lead and REC member]

Participants had mixed views on the use of the databank by commercial organizations including technology and pharmaceutical companies. Patients or public members were generally against the use of their data by these organizations for the purposes of the databank, although other groups felt that these partnerships may be beneficial owing to commercial organizations’ considerable expertise and resources and that they should be allowed access if they could show that they meet the databank’s standards. The primary consideration of commercial access was to ensure adequate return and public benefit. Participants discussed the need for assurances around how tools will be fed back into the public sector, for example, whether the NHS would have discounted access to any tools that were developed as a result of the databank. Access by charities was viewed as potentially problematic, as charities are less regulated and often have a campaign focus, which may result in data use being less scrutinized or controlled and the creation of NLP tools biased toward certain outcomes. Government organizations, insurance companies, and lawyers were deemed unsuitable.

Several participants agreed that existing models of good practice about access such as the Secure Anonymised Information Linkage (SAIL) Databank [29] in Wales should be incorporated and that learning from the Centre for Data Ethics and Innovation Public Attitudes Survey on Data and AI [30] about who is trusted with data and in what circumstances should inform the road map.

Fee-Based Model for Use of the Databank

Clinician and NLP researcher groups were asked about how the databank should be funded. Both groups welcomed government funding to help develop the databank, but a fee-based model was viewed as more sustainable in terms of supporting the management of access and oversight, ensuring data quality, and allowing data to be updated over time. Charging a not-for-profit fee was viewed as realistic, and participants favored the development of a tiered costing model with different levels of access and cost depending on the user, their reason for access, and the volume of data requested. Suggestions for tiered access included providing an institute-wide membership fee to enable access for anyone working within the organization and discounted rates or free access for those who contribute to data donation, maintenance of the databank, or data cleaning or other data quality control. It was suggested that the development of the cost model should be informed by existing approaches such as the Linguistic Data Consortium [31], which supports NLP research by creating and sharing resources, and the UK Data Service [32], a large repository of economic, social, and health data sets for research and teaching.

Data Gathering, Management, and Housing

Participants felt that with health care services under notable pressure, a robust and sustainable costing plan to support the transfer of data into the databank is essential. For example, GP practices may expect to be paid to carry out data extraction for the databank; therefore, costs should be factored into the costing model.

Services are already stretched and I know in some of my practices if you come to them and ask them for data, whether patient’s consented or not, they’re going to tell you to go take a high jump. I’m not going to spend my time extracting that for you or printing that off and sending it to you because I haven’t got the time or capacity. [Clinician]

The participants articulated 5 considerations that should be built into the way the databank is managed. The considerations were that the databank should be (1) accessible: access should be easier than the current process of applying for and accessing data from data providers; (2) up-to-date: the databank should be supplemented with new data so NLP tools are trained on current information; (3) controlled: robust technical controls should be put in place as the primary mechanism for managing data, including allowing and revoking access; (4) tracked: a mechanism should be in place to allow the use of data to be tracked in real time to see what happens to the data after they are accessed; and (5) transparent: transparency around how the databank is being used to track public benefit, for example, publishing details on a website, was seen as essential.

There was no clear consensus on the best approach to housing the databank. The options discussed included housing it within the NHS (perceived benefits were trustworthiness and existing robust policies around data breaches or data misuse; disadvantages were possible lack of technical infrastructure or resources to be able to manage it effectively) or a university setting. Different models considered included adopting a partnership approach in which the databank could be housed within a university but be governed by the NHS. Storing data within a secure environment such as a trusted research environment where data cannot be removed was felt to be appropriate, and existing models of good practice should be drawn upon, for example, Genomics England [33] and Health Data Research United Kingdom [19].

Over sight and Management

The integrity of the databank was closely linked to the approach to oversight and clear communication of how gatekeeping will take place. Most participants favored the establishment of an oversight committee to consider the range and types of data collected, review applications for access and use, ensure transparency around use including what the NLP tools being developed will be used for, and monitor and review data safeguarding. The committee should be independent and consist of people with diverse demographics, backgrounds, and expertise, including experts in the use of data, data controllers, and lay representatives. Patients or public participants discussed the importance of ensuring that the application process to join the oversight committee was accessible and did not put off
potential new applicants by overemphasizing a requirement for previous experience, as is often the case.

**Discussion**

**Principal Findings**

This study set out to test early stakeholder thinking around the acceptability and design considerations for the creation of a consented donated databank of clinical free text to develop and test NLP methods and tools. Understanding the details to inform establishment of such a databank was highlighted as a key recommendation in a recent position paper on the development of data governance standards for using clinical free-text data in health research [1]. All stakeholder groups voiced strong support and a pressing need for a free-text databank for the purposes set out to them. Participants highlighted a range of complex issues for consideration as the databank is developed, but there was a plea, particularly among the NLP researcher group, to move with haste to design something that works without becoming overburdened with the many complexities. One suggested approach was to develop the databank in phases, with the initial phase focusing on a specific health condition or type of health data, to test out whether people are willing to donate their data for this purpose and how it would work. Although not raised by any of the groups, a sensible starting point for the databank may be to exploit existing cohorts such as Generation Scotland or the UK Biobank where participants have already consented to share their data for research and can be easily contacted to invite them to donate their data to the databank. Participants stressed the importance of ongoing engagement and involvement with stakeholder communities in the development and operation of the databank. This position encapsulates the widely held view of the importance of transparency to increase the general lack of awareness about how patient data are used, by whom, and for what purposes [16].

Future proofing the databank at an early stage was viewed as important to take into account how uses of the databank and advances in technology might change over time. Participants also highlighted the importance of ensuring that plans for the provision and maintenance of the databank are sustainable in the future. There was a general agreement among stakeholder groups that the databank should draw upon a range of data sources to ensure that NLP tools reflect an integrated care system in the future, although there were mixed views about the benefits of linking to other data sources. The types of data to be included in the databank (eg. structured data in the GP or hospital record that may improve the performance of existing models owing to the inclusion of text-based features [34,35] or linkage to other data outside the NHS EHR, eg. national registry, mortality, or administrative data) should reflect stakeholder views on acceptability and practicalities, including cost, especially at the start. These issues should be explored in more detail in the next phase of the study. Participants agreed that the way the databank is created will be crucial to its success. The importance of communicating the intended purpose and approach to accessing the databank and the proposed mechanisms for safeguarding the data came up at numerous points during discussions with all groups. Learning from existing examples of good practice around access to data, data security, and how to fund the databank was also deemed crucial.

During the discussions, participants were not directed by the facilitator as to whether data stored in the databank would remain identifiable or not to encourage a broad conversation around anonymity in relation to the databank. Although it was made clear that donations would be based on explicit donor consent, issues of trust and maintaining patient confidentiality featured strongly in the discussions, particularly among the patient and public and IG and REC groups who were aware that public awareness of the risk of reidentification might act as a barrier to data donation, particularly among people with rare diseases who may be easier to identify. Interestingly, only patients and public participants expressed concern about sharing their own data with a databank (Table 1), although the numbers were small, with only 18% (3/17) of the patients or public members declaring that they would be very or somewhat uncomfortable donating their data. This finding is likely to reflect both a clearer understanding of the potential benefits of the databank among other stakeholder groups, given their expertise in this area, and the complexity of what is required to manage and mitigate the risks of data breaches and uphold privacy, which other stakeholder groups may understand more clearly. More work is needed to engage with patients and public members in this area to develop strategies for clear and widespread articulation of the benefits of the donated databank. Given the nature of the data and the challenges of removing personal identifiers, a realistic approach to deidentification will need to be adopted and made transparent and should be supported by robust processes to protect against risks. The deidentification approach may need to be dynamic, depending on the type of data and health condition, as some identifying data might be essential to the research study, for example, if the databank will be used to develop and test deidentification tools. It will also be worth exploring options to replace identifiers with random replacement identifiers to enable this type of work and remain mindful of the advances in generating synthetic data. Further exploration of stakeholder views to understand if stakeholders expect data to be stored in an identifiable form or if they want it to be deidentified is warranted. Although stringent steps will be adopted to minimize the risk of reidentification of patients by deidentifying the data and ensuring strict controls over who can access the data and how data will be made available, for example, within a trusted research environment, risk cannot be eliminated completely. The model for the databank should balance strong governance and security measures to ensure that access is not unnecessarily burdensome or complex. Learning from existing models will be the focus of the next phase of the development of the databank.

Another clear theme throughout the discussions was the need to develop a carefully planned and strong communication plan to build trust. It was suggested that distinct key messages be prepared depending on the stakeholders’ interests in the databank. For example, there should be targeted communication with clinicians regarding their data privacy and with data controllers regarding data security. Communication should...
incorporate relevant background information (eg, to counter the lack of awareness of what data are contained within EHRs) and address the context for the databank, which drives donations and the involvement of GP practices and other data providers.

Ideally, linguistic features in NLP should be representative of the entire population to ensure that the findings are not biased and are representative across patient groups. However, it is also important that the databank reflects what happens in the real world, despite the potential limitations owing to bias. The participants discussed the potential for bias and its impact on the databank (“garbage in, garbage out”) in 2 areas: first, potential biases in the data because of inaccuracies or missing data, and second, potential biases in the data because of donations that lacked demographic variation. Biases in the annotation process [36] and other potential biases have not been discussed. Further in-depth consideration of how to avoid biases and the potential consequences for the trained models is needed when developing the databank and should be addressed clearly in the communication plan.

Participants expressed mixed views about the impact of training NLP models on inaccurate or missing data or data that are not up-to-date. Patients or public participants were concerned that the quality of data would affect the quality of the algorithms that will be developed. Patients or public participants were able to highlight numerous examples of inaccurate or missing data in their own health records which they found concerning and they expressed concern around how the potential lack of accurate or up-to-date data might impact on trust in the databank. Additional engagement with stakeholders, in particular patients and public participants, should be carried out to tease out and address major questions or lack of understanding about the impact of accuracy, subjectivity, and representativeness of data when training NLP tools so that people have a better understanding of what the databank can achieve.

Communication should include efforts to make clear that NLP development is not interested in whether data are accurate or true, as it is simply trying to learn the linguistic properties of the data and how they relate to target concepts defined by NLP researchers and annotators.

Notably, participants were acutely aware of, and made reference throughout the discussions to, the important role AI will play in health care in the future but raised concerns about how AI tools are developed and perform and what might be ethical in the future. Embedding ethical approaches in developing data-driven technologies for AI and understanding public trust is high on the UK governments’, researchers’, and other stakeholders’ agendas. For example, the NHS England Accelerated Access Collaborative [37] is committed to working with patients to ensure that AI innovations reflect the priorities of the end users and support innovators to embed public involvement in the development of AI technologies. The Centre for Data Ethics and Innovation, which is responsible for monitoring public attitudes toward data and AI over time, recently published findings from its second “Public Attitudes to Data and AI (PADAI) Tracker Survey” [30]. Findings from our study reflect similar views to those identified in this survey: for example, data security and privacy remain major concerns, people expect strong governance overseen by experts, and trust is strongly linked to the level of trust in the organizations that are accessing the data. Adhering to best practices around ethical AI principles and frameworks and anchoring public involvement in the development of the databank should be a priority to build trust, and developers of the databank should engage with leaders in the field to ensure this is embedded in plans for the databank, for example, the NHS England AI Ethics Initiative [38]. To keep up with the research and development in AI applied to clinical settings that is happening in the United States (made possible by data sets such as MIMIC III and IV), the UK government should channel resources into funding such a databank to harness rapid advances of AI technology and support long-term investment in the AI ecosystem in the future.

Limitations

This study has limitations. The sample size was relatively small, and there was a lack of diversity, particularly from younger and older participants and people of color. Therefore, the participants’ views are unlikely to be representative of the UK general population. Engagement with more diverse groups and stakeholders who were not included in this work, for example, data controllers, is essential when planning next steps for the databank in the future. Thematic analysis does not allow views to be quantified, so we were unable to report how many participants felt a particular way. Furthermore, the aim of the focus groups was not to produce a specific set of recommendations. Rather, our findings provide useful insights into initial thinking, and the recommendations presented in this study (Textbox 1) therefore reflect a set of potential suggestions and advice based on the views of participants generally. Web-based discussions were limited to 90 minutes to ensure that the length of the focus groups was manageable for participants, which meant some topics could not be explored in depth. The research team therefore agreed in advance how to limit questions to ensure topics deemed the most relevant to particular stakeholders were covered within the time frame. Opportunities to explore topics in more detail will be sought in the next phase of the study.
Textbox 1. Proposed recommendations and suggestions for setting up the databank.

General approach

- Recommendation 1: Stakeholders should be involved throughout the development, implementation, and maintenance of the databank, including development of the scope, consent model, and communication plan.
- Recommendation 2: The databank should draw on the existing successful examples that can offer helpful models for consent, governance, data housing, and data security and be governed by an oversight committee.

Scope and phasing of the databank

- Recommendation 3: The databank should have a clearly defined purpose and take into account how natural language processing (NLP) researchers may wish to use it in the future.
- Recommendation 4: Development of the databank should be based on a small-scale, gradual approach to starting to gather donations to establish proof of concept and interest in donating to the databank. This might involve gathering data for 1 health condition (eg, diabetes) or location (eg, a mental health National Health Service [NHS] Trust) before moving to include others.

Channels to reach potential donors

- Recommendation 5: Reaching potential donors and publicizing the databank should include trusted individuals, networks, and organizations that support research using health data.
- Recommendation 6: Innovative ways to reach out to minority groups such as identifying public community champions who can advise and reassure others about the benefits and safety of the databank should be explored.

Consent

- Recommendation 7: Ensure the consent process is simple and accessible. Consent should be collected electronically, and information should link to a relevant NHS research ethics committee website and be offered in multiple formats and languages.
- Recommendation 8: The focus of the consent information sheet should be to provide reassurance around use; data security; and, if appropriate, deidentification. It should clearly define the purpose of the databank, provide a clear explanation of what data people are being asked to donate, and describe examples of scenarios for future use.
- Recommendation 9: Opportunities for showing potential participants their own personal health record before consenting should be explored.

Communication

- Recommendation 10: A clear and comprehensive communication plan should be carefully planned and developed with targeted messages for the different stakeholders (eg, clinicians regarding their data privacy and data controllers regarding data security).
- Recommendation 11: Communication should cover the following key elements clearly to build trust in the databank: predonation involvement (eg, possibility for participants to see their personal data and amend errors before donating); general aspects around data (what is free text? and what data are in a health care record?); content (what data are to be donated?); purpose (what the free-text data will be used for and by whom?); different contexts that NLP tools will be used in (eg, will data be used largely for commercial benefit?); and, crucially, the public benefit that NLP tools trained on the data could bring.

Pathways to databank access

- Recommendation 12: The foremost consideration for access should be to ensure public benefit and that benefits of data use are shared equitably.
- Recommendation 13: A “road map” should be developed to include the types of organizations the databank will work with, based on a compliance model where users should show they can meet a defined level of capability and accountability. The road map should include a set of standards and approach to “due diligence” to ensure databank use is in line with its intended purpose. Access could be granted based on an organization’s ability to meet, and commitment to comply with, the standards and an assessment of the applicant organization’s reputation and motivation for using the databank, rather than limiting which types of organization should be allowed access. The road map should incorporate recent learning on who is trusted with data and in what circumstances.
- Recommendation 14: Although development of the databank is in its infancy, it may be prudent to limit access to a small group such as NLP researchers linked to the NHS and UK universities.
- Recommendation 15: Access should be easier than the current process of applying for and accessing data from data providers.

Cost model for the databank

- Recommendation 16: A clear, transparent, and not-for-profit fee-based model should be developed that ensures sustainability of the databank over time. Fees should be used to maintain the database, support and manage access and standards, support oversight, ensure quality of data, and support updating the databank with new data over time.
- Recommendation 17: A tiered access model should include different levels of fees depending on the user, reason for use, and volume of data required (eg, access to a portion of the data set or all of it). Discounted or free access should be considered, for example, discounted access for organizations that contribute to data donation or where the databank would be used for teaching purposes. “In kind” arrangements could be considered for organizations that collaborate on improving quality of data (eg, cleaning data for access).
Comparison With Prior Work

Although the focus of this study centered on creation of the databank that has not, to the authors’ knowledge, been previously explored, there were several areas where themes overlapped with previous research on attitudes toward the use of free-text data for research, which has been discussed particularly among patients and the public [16,39]. Although several benefits highlighted by participants in this study related specifically to the databank, wider benefits discussed included the potential for improving health and care leading to better outcomes for patients, which mirrored benefits identified by other UK research studies that used clinical free text [2]. Despite acknowledging a broad range of potential benefits, participants raised a number of concerns, particularly around how AI tools are developed and perform in general, the effect of possible biases, privacy risks, and reidentification. Previous research on potential harms of the use of free-text data for research has shown that the public harbors similar concerns around the use of free-text data for research generally, despite no evidence of these harms actually taking place following data breaches [40]. The issue of trust was raised several times, as was the importance of clear communication and a transparent approach to help build trust. Participants in this study felt that trust is strongly linked to the level of trust in the organizations that access the data, which echoes findings from other studies that showed that the public evaluates the trustworthiness of research organizations by assessing their competence in data handling and motivation for accessing the data [41]. A Citizens’ Jury on the use of free text for research carried out in 2018 [16] found a high degree of willingness to share EHR data for public benefit among public participants who were informed about the use of free-text data, although participants expressed caution owing to concerns around the lack of transparency in the use of data and increased privacy risks. Participants in the Citizens’ Jury suggested keeping patients informed about the use of their data and being transparent about ways to opt out of data sharing. These attitudes were mirrored in this study, as were views on risks related to deidentification of free text, which were in line with previous findings, including concerns around accuracy of removing patient identifiers.

Next Steps

The recommendations and advice resulting from the study are summarized in Textbox 1. The findings will be used to plan the next phase of developing the databank, including a pilot study to design the road map and communication plan and test the feasibility of donating to the databank. Next steps will involve identifying and reaching out to a broad range of stakeholders based on their diverse knowledge and skill sets to develop the vision for the databank and inform the road map and standards, including researchers, patients and public members, governance...
experts, providers of NHS, data controllers, charities, government, and industry. The road map and standards could be further informed by a national web-based survey that will be co-designed with stakeholders to explore in more detail the acceptability and design considerations highlighted in this study, including understanding whether stakeholders expect data to be stored in an identifiable or deidentifiable form. Planning the next steps will draw on recommendations across relevant themes that were highlighted in a position paper on developing data governance standards for the use of free-text data in health research, including the involvement of patients and public members at identifiable data stages and opt-in consent models for the reuse of free-text data [1].

Conclusions

Improved access to clinical free-text data will help support technological innovation for developing novel and valid NLP tools to support research for public benefit. One way to leverage access is through the creation of a consented databank to develop and train NLP tools outside the NHS via the lawful basis of informed consent. This study showed strong multistakeholder support for a databank for this purpose and an urgent need to move forward to develop something quickly. Stakeholders expressed commonality around many issues such as governance, communication, and sustainability, but there were also stakeholder-specific concerns such as clinician concern around increased workload and privacy and patient-and-public concern around inaccuracies in their personal EHRs and how their data will be used. These issues should be explored in more detail and targeted among individual stakeholder groups. Findings from this study will be used to inform the next steps for adopting a partnership approach to establish a national, funded databank of free text for use by the research community.

Acknowledgments

This work was funded by Healtx and Health Data Research UK. The funders have no role in developing the content of this manuscript. The authors would like to acknowledge and thank all the focus group participants for their expert knowledge and continued support of this work and Hopkins Van Mil for facilitating the focus groups. ADS is supported by NIHR (AI_AWARD01864 and COV-LT-0009), UKRI (Horizon Europe Guarantee for DataTools4Heart) and British Heart Foundation Accelerator Award (AA/18/6/24223).

Data Availability

Focus group discussions were recorded for the purposes of report writing. Recordings were destroyed after being transcribed, but transcriptions are stored securely for 24 months, after which they will be destroyed (May 2024).

Authors’ Contributions

All authors were involved in the conception and design of the study, critically reviewed the manuscript, and granted approval of the final version to be published. NKF led the participant recruitment and acquisition of data and wrote the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Topic guide: focus group questions and the distribution of questions between groups.

[DOC File, 88 KB - medinform_v11i1e45534_app1.doc ]

Multimedia Appendix 2

Summary of focus group findings and key areas of discussion.

[DOCX File, 30 KB - medinform_v11i1e45534_app2.docx ]

References

5. SNOMED CT (Systematized Nomenclature of Medicine -- Clinical Terms). Tech Target. URL: https://www.techtarget.com/searchhealthit/definition/SNOMED-CT [accessed 2022-12-05]


18. Medical information mart for intensive care. MIMIC. URL: https://mimic.mit.edu/ [accessed 2023-02-24]

19. Health Data Research UK homepage. HRD UK. URL: https://www.hdruk.ac.uk/ [accessed 2022-12-05]


28. Use my data homepage. Use My Data. URL: https://www.usemydata.org/ [accessed 2022-12-05]

29. The secure anonymised information linkage homepage. SAIL Databank. URL: https://saildatabank.com/ [accessed 2022-12-05]


31. Linguistic data consortium homepage. Linguistic Data Consortium. URL: https://www.ldc.upenn.edu/ [accessed 2022-12-05]

32. UK data service homepage. UK Data Service. URL: https://ukdataservice.ac.uk/ [accessed 2022-12-05]


40. Understanding Patient Data homepage. Understanding Patient Data. URL: https://understandingpatientdata.org.uk/weighing-up-risks [accessed 2022-12-05]


Abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>artificial intelligence</td>
</tr>
<tr>
<td>COREQ</td>
<td>Consolidated Criteria for Reporting Qualitative Research</td>
</tr>
<tr>
<td>EHR</td>
<td>electronic health record</td>
</tr>
<tr>
<td>GP</td>
<td>general practitioner</td>
</tr>
<tr>
<td>ICD-10</td>
<td>International Classification of Diseases, 10th revision</td>
</tr>
<tr>
<td>IG</td>
<td>information governance</td>
</tr>
<tr>
<td>MIMIC</td>
<td>Medical Information Mart for Intensive Care</td>
</tr>
<tr>
<td>NHS</td>
<td>National Health Service</td>
</tr>
<tr>
<td>NLP</td>
<td>natural language processing</td>
</tr>
<tr>
<td>REC</td>
<td>research ethics committee</td>
</tr>
<tr>
<td>SAIL</td>
<td>Secure Anonymised Information Linkage</td>
</tr>
<tr>
<td>SNOMED CT</td>
<td>Systematized Nomenclature of Medicine Clinical Terms</td>
</tr>
</tbody>
</table>

©Natalie K Fitzpatrick, Richard Dobson, Angus Roberts, Kerina Jones, Anoop D Shah, Goran Nenadic, Elizabeth Ford. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 03.05.2023. 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.
Acquisition of a Lexicon for Family History Information: Bidirectional Encoder Representations From Transformers–Assisted Sublanguage Analysis

Liwei Wang¹, MD, PhD; Huan He¹, PhD; Andrew Wen¹, MSc; Sungrim Moon¹, PhD; Sunyang Fu¹, PhD; Kevin J Peterson², PhD; Xuguang Ai³, MSc; Sijia Liu¹, PhD; Ramakanth Kavuluru⁴, PhD; Hongfang Liu¹, PhD

¹Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, United States
²Center for Digital Health, Mayo Clinic, Rochester, MN, United States
³Department of Computer Science, University of Kentucky, Lexington, KY, United States
⁴Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky, Lexington, KY, United States

Corresponding Author:
Hongfang Liu, PhD
Department of Artificial Intelligence and Informatics
Mayo Clinic
200 First Street SW
Rochester, MN, 55905
United States
Phone: 1 507 293 0057
Email: liu.hongfang@mayo.edu

Abstract

Background: A patient’s family history (FH) information significantly influences downstream clinical care. Despite this importance, there is no standardized method to capture FH information in electronic health records and a substantial portion of FH information is frequently embedded in clinical notes. This renders FH information difficult to use in downstream data analytics or clinical decision support applications. To address this issue, a natural language processing system capable of extracting and normalizing FH information can be used.

Objective: In this study, we aimed to construct an FH lexical resource for information extraction and normalization.

Methods: We exploited a transformer-based method to construct an FH lexical resource leveraging a corpus consisting of clinical notes generated as part of primary care. The usability of the lexicon was demonstrated through the development of a rule-based FH system that extracts FH entities and relations as specified in previous FH challenges. We also experimented with a deep learning–based FH system for FH information extraction. Previous FH challenge data sets were used for evaluation.

Results: The resulting lexicon contains 33,603 lexicon entries normalized to 6408 concept unique identifiers of the Unified Medical Language System and 15,126 codes of the Systematized Nomenclature of Medicine Clinical Terms, with an average number of 5.4 variants per concept. The performance evaluation demonstrated that the rule-based FH system achieved reasonable performance. The combination of the rule-based FH system with a state-of-the-art deep learning–based FH system can improve the recall of FH information evaluated using the BioCreative/N2C2 FH challenge data set, with the F1 score varied but comparable.

Conclusions: The resulting lexicon and rule-based FH system are freely available through the Open Health Natural Language Processing GitHub.

(JMIR Med Inform 2023;11:e48072) doi:10.2196/48072

KEYWORDS
electronic health record; natural language processing; family history; sublanguage analysis; rule-based system; deep learning

Introduction

Family history (FH) has long been regarded as a core element in caring for patients who have varied health concerns [1], with the capability to significantly enhance the delivery of precision medicine [2]. However, FH data are underused for actionable risk assessment [1]. One barrier to using FH information is provider preference in recording the collected FH information.
in an unstructured format (eg, clinical notes) [3] as opposed to within electronic health record (EHR) structured data [4]. As clinical text tends to be unstructured, the information contained within is computationally inaccessible relative to that contained in structured records. This lack of computational accessibilities poses a challenge in using FH information for downstream data analytics or clinical practice (eg, via clinical decision support). One approach to render information data computationally accessible is through the use of natural language processing (NLP), thus motivating our work to develop an NLP system that can extract and normalize FH information.

Despite the majority of clinical NLP measurement studies focusing on statistical approaches, rule-based NLP systems based on semantic lexicons and rule patterns are popular among observational studies [5] for cancer research and practice [6,7]. With the advantages of ensuring process transparency, implementability, and scientific rigor, semantic lexicons and rule patterns are interpretable and easily modifiable, conforming to the FAIR (Findable, Accessible, Interoperable, and Reusable) and RITE (Reproducible, Implementable, Transparent, and Explainable) principles [8,9] for scientific data management. In addition, semantic lexicons and rule patterns capture sublanguage characteristics of domains that can be portable and generalizable to other applications [10]. By sublanguage, we refer to domain-specific linguistic and lexical patterns that are more prominent in free text in specialized fields such as medicine.

One popular lexical resource for clinical NLP is the Unified Medical Language System (UMLS), a repository of biomedical vocabularies distributed by the US National Library of Medicine, integrating over 200 biomedical vocabularies. A source vocabulary contained within the UMLS is the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT), which is the recommended coding system for clinical problems. As not all terms in the UMLS or SNOMED-CT are part of the FH sublanguage, in this study, we exploited a corpus-driven method with pretrained language models to build an FH semantic lexicon with the normalization feature and reasonable size and coverage.

There have been previous efforts focused on creating semantic lexicons for clinical NLP. Johnson [11] automatically constructed a semantic lexicon based on the Specialist Lexicon of the UMLS, which can assist NLP analysis of a medical narrative with the semantic preference options of selecting semantic type. Luo et al [12] created a semantic lexicon using UMLS knowledge sources by leveraging a corpus from ClinicalTrials.gov. Liu et al [13] constructed a corpus-driven semantic lexicon based on the UMLS assisted by variants mined and usage information gathered from clinical text.

Regarding deep learning–based approaches, pretrained language models such as bidirectional encoder representations from transformers (BERT) [14] can learn the structure of language (ie, the basic semantic and syntax information) through unsupervised training on a large corpus of unlabeled text [15]. Given a new task, such pretrained models can be fine-tuned with a small number of annotated samples to perform well [16]. With respect to the FH information extraction task, we hypothesized that terms in a large-scale corpus having semantic types similar to the entities labeled in the data set used for fine tuning can be detected by fine-tuned BERT models in a named entity recognition task. Additionally, we hypothesized that corpus-driven methods would enable more term variants to be discovered from real-world EHR data, from which the lexicon results could further enhance and empower FH information extraction systems. Operating under these two hypotheses, we here present an FH lexicon derived through a combination of these two approaches. To demonstrate the usability of the FH lexicon, we further developed a rule-based FH system based on the lexicon that extracts FH entities and relations specified in previous FH challenges and evaluated its performance accordingly. In terms of system development, we consider that an FH system that prioritizes recall is highly desired for NLP-assisted curation in EHR-based studies.

**Methods**

**FH Concepts**

Before assembling an FH lexicon, we defined FH concepts as those belonging to the selected UMLS semantic types within the “DISO” (disorder) semantic group (Table 1), excluding T050 (experimental model of disease) [17]. Figure 1 shows our study design. We first fine-tuned BioClinicalBERT, UmlsBERT, and bert-base-uncased models, and selected the model with the best performance. We then used the selected model to extract potential disorder/finding related mentions in a large clinical corpus. Subsequently, potential FH mentions were automatically normalized, manually curated, and prepared into symbolic lexicon format compatible with the Open Health Natural Language Processing (OHNLP) Toolkit’s NLP engine MedTagger [18]. A coverage evaluation was conducted for the lexicon. To demonstrate the usability of the lexicon, a rule-based FH system was developed to extract FH information. In parallel, we also experimented with a deep learning–based FH system for FH information extraction. Previous FH challenge data sets were used for evaluation [19,20].
Figure 1. Study design. BERT: bidirectional encoder representations from transformers; ECH: employee and community health; FH: family history; I2B2: 2012 i2b2 natural language processing challenge data set (training data set); PURE: Princeton University Relation Extraction; SFCP: standardization framework for clinical problems.

<table>
<thead>
<tr>
<th>Semantic type code</th>
<th>Semantic type term</th>
</tr>
</thead>
<tbody>
<tr>
<td>T019</td>
<td>Congenital abnormality</td>
</tr>
<tr>
<td>T020</td>
<td>Acquired abnormality</td>
</tr>
<tr>
<td>T037</td>
<td>Injury or poisoning</td>
</tr>
<tr>
<td>T047</td>
<td>Disease or syndrome</td>
</tr>
<tr>
<td>T048</td>
<td>Mental or behavioral dysfunction</td>
</tr>
<tr>
<td>T049</td>
<td>Cell or molecular dysfunction</td>
</tr>
<tr>
<td>T190</td>
<td>Anatomical abnormality</td>
</tr>
<tr>
<td>T191</td>
<td>Neoplastic process</td>
</tr>
<tr>
<td>T033</td>
<td>Finding</td>
</tr>
<tr>
<td>T046</td>
<td>Pathologic function</td>
</tr>
<tr>
<td>T184</td>
<td>Sign or symptom</td>
</tr>
</tbody>
</table>

Resources

Overview

Here, we introduce the resources used for lexicon construction and all relevant evaluations. Specifically, the 2018 BioCreative FH challenge training set and the 2012 I2B2 training set were used as the supervised data sets to fine-tune the deep learning models for lexicon preparation. Various data sets were experimented with based on a hypothesis that more data could encompass more semantic contexts of potential FH mentions. The 2018 test set was used to evaluate lexicon coverage. The 2018 and 2019 FH challenge training sets were used for training of the rule-based FH system and fine-tuning the deep learning–based FH system, and the 2018 and 2019 FH challenge test sets were used for evaluating the performance of the FH systems in extracting FH entities (task 1) and relations (task 2). A large EHR corpus was used for collecting potential FH mentions. MedLex was used to further enrich term variants of FH concepts extracted by the selected model. The UMLS was used for semantic type selection, while UMLS and SNOMED-CT were used for a comparison of size with our corpus-driven dictionary. Table 2 summarizes specific applications of the resources in the lexicon construction and all relevant evaluations. Detailed descriptions for each resource are provided below.
## Table 2. Specific applications of resources in the lexicon construction and evaluations.

<table>
<thead>
<tr>
<th>Application</th>
<th>A&lt;sup&gt;a&lt;/sup&gt;</th>
<th>B&lt;sup&gt;b&lt;/sup&gt;</th>
<th>C&lt;sup&gt;c&lt;/sup&gt;</th>
<th>D&lt;sup&gt;d&lt;/sup&gt;</th>
<th>E&lt;sup&gt;e&lt;/sup&gt;</th>
<th>A+E</th>
<th>EHR&lt;sup&gt;f&lt;/sup&gt; corpus</th>
<th>MedLex</th>
<th>UMLS&lt;sup&gt;g&lt;/sup&gt;</th>
<th>SNOMED-CT&lt;sup&gt;h&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexicon construction</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Fine-tuning the BERT&lt;sup&gt;i&lt;/sup&gt; model</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Dictionary entry collection</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Dictionary concept enrichment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Semantic type selection</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>FH&lt;sup&gt;j&lt;/sup&gt; system development</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Development of rule-based FH system</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Fine-tuning deep learning–based FH system</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>Evaluation</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Dictionary coverage</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Challenge task 1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Challenge task 2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>A: Training set (BioCreative).

<sup>b</sup>B: Training set (N2C2).

<sup>c</sup>C: Testing set (BioCreative).

<sup>d</sup>D: Testing set (N2C2).

<sup>e</sup>E: Training set (I2B2/2010).

<sup>f</sup>EHR: electronic health record.

<sup>g</sup>UMLS: Unified Medical Language System.

<sup>h</sup>SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terms.

<sup>i</sup>BERT: bidirectional encoder representations from transformers.

<sup>j</sup>FH: family history.

### Synthetic FH Annotation Data Sets (A-D)

As the organizer of the BioCreative/OHNLP 2018 Family History Extraction Task [19] and 2019 NLP Clinical Challenge (N2C2)/OHNLP shared task [20], we curated deidentified annotation data sets based on synthetic clinical narratives. Data set A corresponds to data set C in the BioCreative Challenge and data set B corresponds to data set D in the N2C2 Challenge. FH was annotated as “observation” and defined as any health-related problem, including diseases, smoking, suicide, and drinking, while excluding auto accidents, surgeries, and medications [19]. Family members (FMs), observation, age, and living status were annotated as entities, and then all entities related to an FM category were linked into one chain. We further enhanced the resulting annotations in the data sets by normalizing observations to SNOMED-CT codes and correcting errors in previous annotations. The reannotated data sets are accessible based on the Data Use Agreement. Multimedia Appendix 1 shows the statistical comparison between original and enhanced annotations.

### I2B2 Data Set (D)

The 2012 I2B2 NLP challenge organizers provided a fully deidentified data set with annotations for temporal relations as well as those generated from previous challenges such as the 2010 challenge of clinical concept extraction (problems, tests, treatments) [21], where problems include symptoms, complaints, diseases, and diagnoses.

### EHR Raw Corpus

The raw corpus used in the study consists of 9,426,352 text segments extracted from the Family History section of clinical notes prior to 2013 of a primary care cohort (ie, the employee and community health cohort), which contains 83,000 patients at Mayo Clinic.

### Dictionary Resources

MedLex is a semantic lexicon built on a large corpus of clinical documents collected at Mayo Clinic and from the UMLS (2011AA version) [13]. MedLex contains term variants from real-world EHRs, serving as a practical dictionary resource for FH lexicon expansion. We used MedLex to further enrich term variants of FH concepts extracted by the selected model. The UMLS (2021AA) and SNOMEDCT_US concepts accessible through the UMLS were restricted to only English entries. The MRCONSO table of the UMLS, which includes over 200 source vocabularies, was used for coverage evaluation, and the MRSTY table of the UMLS was used for screening semantic types for each concept unique identifier (CUI) in the MRCONSO table.

### Ethical Approval

Use of the EHR raw corpus data was approved by the Mayo Clinic Institutional Review Board (17-003030) for Human Subject Research.
BERT-Based Corpus Analysis for Lexicon Construction

BERT Models for Extraction of Potential FH Mentions
BERT-base-uncased was pretrained on an unsupervised NLP data set using a masked language modeling approach [14]. Bio_ClinicalBERT was initialized from BioBERT and trained on all Medical Information Mart for Intensive Care notes [22]. UmlsBERT is a contextual embedding model that integrates domain knowledge during the pretraining process via a novel knowledge augmentation through the UMLS Metathesaurus [23]. UmlsBERT and Bio_ClinicalBERT are domains related to this study, while BERT-base-uncased could be used as a baseline comparison. We used configurations mostly consistent with the recommendations in the original release of the models. The maximum sequence length was set to 512, the batch size was set to 16, the total number of training epochs was set to 100, and the weight decay was set to 0.01. An early stopping method was used to determine the optimal number of epochs and to prevent overfitting. The train/test split was 80/20, where the “train” split was used for training and the “test” split was used for validation.

We fine-tuned these models on the BioCreative training data set, I2B2 training data set, as well as the combination of the BioCreative and I2B2 training data sets. We then selected the model with the best performance to extract any potential FH mentions in the raw corpus having coarse-grained semantic types similar to the entities labeled in the supervised data.

Normalization
The extracted FH mentions were automatically normalized through a standardization framework for clinical problems (SFCP) [24]. This framework converts free-text clinical problem descriptions into standardized forms based on the UMLS CUI corresponding to SNOMED-CT concepts and the Health Level 7 Fast Healthcare Interoperability Resources (FHIR)-based structured representations, including the codified problem and all relevant modifiers and context. The CUIs associated with the SNOMED-CT concept were used for coding. For example, for the mention “allergy-induced asthma,” the framework outputs “C0440102 | Various patch test substance” and “C0155877 | Allergic asthma.”

Manual Curation and Further Enrichment
We reviewed all normalized FH mentions and retained the main normalized problem concepts with semantic types corresponding to those previously selected. For example, for the mention “allergy-induced asthma,” the codified problem “C0155877 | Allergic asthma” was kept and “C0440102 | Various patch test substance” was removed from the final lexicon. In some cases, one BERT-extracted mention can be normalized to several individual concepts through the automatic standardization. We then enriched the FH lexicon by keeping all individual concepts and obtaining associated variants from MedLex.

In addition, we manually mapped high-frequency mentions that occurred across at least 20 patients that were not automatically standardized to corresponding CUIs.

Rule-Based FH System
To demonstrate the capability of the FH lexicon in extracting FH relations, we further implemented a rule-based FH system by integrating rules for FM identification with the resulting lexicon. Multimedia Appendix 2 shows three degrees of consanguinity we aggregated into FM identification rules. FH relations were then extracted based on co-occurrence within a clause of one sentence or across three adjacent sentences if coreference existed, as indicated by keywords such as “he,” “she,” “none of them,” “her,” or “his,” while excluding relations between FMs of spouse and FH. We implemented this as an OHNLP Toolkit module with code available on GitHub [25]. The implementation provides several output formats, including FHIR-based output, with FM and FH standardization conforming to FHIR standardization. The final output of the rule-based FH system includes entities and relations. The entity output includes file name (which links to document references with patients’ ID), sentence ID, chunk ID, entity type, concept, and certainty. The relation output includes file name, FM, side of family, text of observation, and certainty. An option is also available to output this information to CSV, instead of FHIR, format. To set SNOMED-CT condition codes as the standard, a separate mapping file is required due to SNOMED-CT licensing restrictions.

Deep Learning–Based FH System
As information extraction remains a challenging task, it is preferred to investigate what is the gain from deep learning–based models. Therefore, we further implemented a deep learning–based FH system as follows. Note that we experimented with fine-tuned models for two purposes in this study. The first was to fine-tune models for identifying and collecting potential FH mentions from clinical texts to build a dictionary, as described in the previous section. Here, the second purpose was to fine-tune models for information extraction to automatically identify FH entities and relations.

The Princeton University Relation Extraction (PURE) system is an approach where the entity model builds on span-level representations and the relation model builds on contextual representations specific to a given pair of spans. As this pipelined approach has been demonstrated to be extremely effective, we implemented PURE using scibert-scivocab-uncased as the base encoder and fine-tuned it based on the BioCreative and N2C2 training data [26].

Evaluation
Overview
We conducted two evaluation studies, including (1) a coverage evaluation of the lexicon and (2) a comparison study of a lexicon-based module with a deep learning–based module for FH information extraction.

Lexicon Coverage
To the best of our knowledge, our lexicon is the first to incorporate a large number of text variants and concepts of FH. Therefore, we compared the resulting lexicon with the UMLS and SNOMED-CT in terms of the number of concepts under each semantic type. We analyzed the lexicon coverage by
calculating the number of concepts under each semantic type. In addition, we calculated the number of concepts and variants covered by the corpus-driven lexicon run against the BioCreative testing data set relative to annotated gold standards. True positive (TP) rate based on a partial match, false negative (FN) rate, and recall (TP/[TP+FN]) at the concept level and variant level were calculated.

Performance of FH Information Extraction

We evaluated the utility of the lexicon in identifying mentions of FMs and their associated attributes (side of family) using the BioCreative testing set and the N2C2 testing sets (ie, task 1 of the challenges). Precision, recall, and F1-scores were calculated as the performance metrics of the lexicon-based module, the deep learning–based module, and both.

We also evaluated the rule-based FH system’s ability to identify relations between FMs, observations, and living status using the BioCreative testing set and N2C2 testing sets (ie, task 2 of the challenges). This task was different between the 2018 BioCreative and 2019 N2C2 challenges in that the latter added a certainty attribute (negated or nonnegated) into relation extraction. Three sets of precision, recall, and F1 values were separately calculated using varying setups: rule-based module only, deep learning–based module only, and a combination of both. Our evaluation scheme is the same as that applied in the 2018 BioCreative and 2019 N2C2 FH challenges. We performed a general error analysis to investigate error sources. In addition, to further investigate how much the lexicon contributes to system performance, we performed an ablation study and specific error analyses.

Results

Multimedia Appendix 3 shows the performance of the BERT-base-uncased, UmlsBERT, and Bio_clinicalBERT models fine-tuned on the BioCreative training data set, I2B2 training set, and on the combination of the BioCreative training set with the I2B2 training set. As the model fine-tuned on the combination of the two data sets outperformed other models, we selected the combined model to extract potential FH mentions from the corpus.

There were 72,518 unique entities identified by the Bio_clinicalBERT model fine-tuned on the combination data set, of which 47,250 (65.16%) were automatically normalized to 10,579 CUIs through the standardization framework for clinical problems. We manually normalized 148 entities occurring across more than 20 patients that were not automatically normalized to CUIs. For example, “typediabetes” with a frequency of 3693 was normalized to C0011854 (diabetes). Note that spellings such as “typediabetes” found in the EHRs are most likely typos due to physicians’ writing, backend EHR note processing, or tokenization challenges. Therefore, manual normalization is important. After semantic type screening, manual curation, and MedLex enrichment, the final FH lexicon contained 33,351 dictionary entries normalized to 6177 CUIs and 15,126 SNOMED-CT codes, with an average of 5.4 variants for each concept. Table 3 shows the comparison of sizes of various lexicons. The corpus-driven lexicon was more light-weighted, with more variants per concept. This implies that implementation with the lexicon for NLP tasks would be easier and more efficient with the corpus-driven lexicon than with the SNOMED-CT and UMLS lexicons.

Lexicon coverage evaluation on the BioCreative testing data set showed that there are 137 TP entities corresponding to 128 concepts and there are 16 FN entities, of which 6 entities had no corresponding concepts in the FH lexicon and 10 entities had 10 corresponding concepts in the FH lexicon. Concept-level recall was 95.8% and variant-level recall was 89.5%. For the N2C2 testing set, there were 507 TP entities corresponding to 214 concepts and 62 FN entities, of which 33 entities had no corresponding concepts in the FH lexicon and 29 entities had 26 corresponding concepts in the FH lexicon. Concept-level recall was 87.9% and variant-level recall was 89.1%. Table 4 shows the comparison of numbers of semantic types of CUIs in various lexicons. It can be observed that the corpus-driven lexicon contains less concepts under each semantic type compared with the UMLS and SNOMED-CT lexicons.

Table 5 shows the performance of FH systems for subtasks 1 and 2 of the BioCreative and N2C2 challenge data sets (original and reannotated) categorized by the rule-based FH system only, deep learning–based FH system only, and a combination of the two. For task 1, the highest F1 score was 0.8766 from the deep learning–based model on the original BioCreative data set and was 0.8061 on the original N2C2 data set. For task 2, the highest F1 score was 0.6206 from the deep learning–based module on the original BioCreative dataset and was 0.5940 from the combined results on the original N2C2 data set. The rule-based FH system based on the corpus-driven lexicon produced lower F1 scores in contrast with the deep learning–based FH system for both tasks, but higher or comparable recall for task 1 and higher recall for task 2. The combined results had the highest recall compared with the rule-based FH system or the deep learning–based FH system for task 1, ranging from 0.8669 to 0.9475. The combined results also showed the highest recall (ranging from 0.7109 to 0.8370) and varied F1 scores (ranging from 0.4288 to 0.6142) for task 2.
Table 3. Comparison of the size of lexicons.

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Concepts (CUI(^a))</th>
<th>Variants</th>
<th>Average number of variants per CUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus-driven lexicon</td>
<td>6177</td>
<td>33,351</td>
<td>5.40</td>
</tr>
<tr>
<td>SNOMED-CT(^b)</td>
<td>412,027</td>
<td>1,349,838</td>
<td>3.28</td>
</tr>
<tr>
<td>UMLS(^c)</td>
<td>4,440,279</td>
<td>9,569,507</td>
<td>2.16</td>
</tr>
</tbody>
</table>

\(^a\)CUI: concept unique identifier.

\(^b\)SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terms.

\(^c\)UMLS: Unified Medical Language System.

Table 4. Statistical summary for semantic types of concept unique identifiers (CUIs) in the lexicon.

<table>
<thead>
<tr>
<th>Semantic type code</th>
<th>Semantic type term</th>
<th>Number of CUIs</th>
<th>Corpus-driven lexicon</th>
<th>UMLS(^a)</th>
<th>SNOMED-CT(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T019</td>
<td>Congenital abnormality</td>
<td>61</td>
<td>11,237</td>
<td>7034</td>
<td></td>
</tr>
<tr>
<td>T020</td>
<td>Acquired abnormality</td>
<td>62</td>
<td>4326</td>
<td>2105</td>
<td></td>
</tr>
<tr>
<td>T037</td>
<td>Injury or poisoning</td>
<td>236</td>
<td>113,258</td>
<td>29,371</td>
<td></td>
</tr>
<tr>
<td>T047</td>
<td>Disease or syndrome</td>
<td>2645</td>
<td>113,780</td>
<td>45,946</td>
<td></td>
</tr>
<tr>
<td>T048</td>
<td>Mental or syndrome dysfunction</td>
<td>376</td>
<td>9152</td>
<td>3318</td>
<td></td>
</tr>
<tr>
<td>T049</td>
<td>Cell or molecular dysfunction</td>
<td>24</td>
<td>4175</td>
<td>583</td>
<td></td>
</tr>
<tr>
<td>T190</td>
<td>Anatomical abnormality</td>
<td>94</td>
<td>6943</td>
<td>1703</td>
<td></td>
</tr>
<tr>
<td>T191</td>
<td>Neoplastic process</td>
<td>866</td>
<td>43,532</td>
<td>11,054</td>
<td></td>
</tr>
<tr>
<td>T033</td>
<td>Finding</td>
<td>1018</td>
<td>309,971</td>
<td>44,547</td>
<td></td>
</tr>
<tr>
<td>T046</td>
<td>Pathologic function</td>
<td>403</td>
<td>27,601</td>
<td>9098</td>
<td></td>
</tr>
<tr>
<td>T184</td>
<td>Sign or symptom</td>
<td>389</td>
<td>14,167</td>
<td>2931</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)UMLS: Unified Medical Language System.

\(^b\)SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terms.

Table 5. Evaluation results for family history (FH) information extraction.

<table>
<thead>
<tr>
<th>Task and data set</th>
<th>Deep learning–based FH system</th>
<th>Rule-based FH system</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
</tr>
<tr>
<td>Task 1 Original</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2018 testing</td>
<td>0.8819</td>
<td>0.8729</td>
<td>0.8766</td>
</tr>
<tr>
<td>2019 testing</td>
<td>0.8271</td>
<td>0.7835</td>
<td>0.8061</td>
</tr>
<tr>
<td>Task 1 Reannotation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2018 testing</td>
<td>0.8830</td>
<td>0.8607</td>
<td>0.8709</td>
</tr>
<tr>
<td>2019 testing</td>
<td>0.7860</td>
<td>0.7747</td>
<td>0.7806</td>
</tr>
<tr>
<td>Task 2 Original</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2018 testing</td>
<td>0.7189</td>
<td>0.5464</td>
<td>0.6206</td>
</tr>
<tr>
<td>2019 testing</td>
<td>0.6841</td>
<td>0.5109</td>
<td>0.5851</td>
</tr>
<tr>
<td>Task 2 Reannotation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2018 testing</td>
<td>0.6777</td>
<td>0.5307</td>
<td>0.5962</td>
</tr>
<tr>
<td>2019 testing</td>
<td>0.3309</td>
<td>0.8168</td>
<td>0.4629</td>
</tr>
</tbody>
</table>

A general error analysis showed that errors could be divided into two major sources. First, varied definitions of FH between different subtasks represented a confounding factor. FH in the gold standards was defined as any health-related problem, including diseases, smoking, suicide, and drinking, excluding auto accident, surgery, and medications [19], whereas the definition in the FH lexicon is based on semantic types (Table 1). For example, mastectomy, a procedure, was annotated as an
observation (FH) in the challenge data sets, which is not in the scope of our lexicon. Second, the intrinsic difficulty of FH relation extraction and its textual representations presented several obstacles, particularly with symbolic systems. The rule-based FH system used simple heuristic rules, and therefore it is difficult to handle complex relationships, especially when the subtask 2 of challenges includes multiple layers of relationships.

In the ablation study, for task 2 based on the original BioCreative test set, the precision, recall, and F1 score for observation only (excluding living status) were 0.5419, 0.6265, and 0.5783, respectively, representing a slight improvement compared with the corresponding values of 0.5413, 0.6000, and 0.5673 for both observation and living status. For task 2 based on the original N2C2 test set, the precision, recall, and F1 score were 0.3984, 0.7033, and 0.5449, respectively, for observation only (excluding living status); 0.4275, 0.6658, and 0.5444, respectively, for both observation and living status; and 0.3730, 0.6675, and 0.5169, respectively, for both observation and certainty. Although living status and certainty alone had little impact on the performance, the combination of observation, living status, and certainty resulted in significantly lower performance of 0.4089, 0.5571, and 0.4716 for precision, recall, and F1, respectively.

**Discussion**

FH has its own sublanguage. However, some terms related to FH may not actually be used in practice or may be used very rarely, such as “ancestor,” “descendant,” or “genealogy.” For this reason, these terms may not appear in the lexicon. As FH is specifically related to blood relations (consanguinity), it relates to the patient themselves. Therefore, the FM of spouse should not be considered, and FH elements relating to a spouse (rather than the patient) should consequently not be extracted. The advantage of this operation is the consistency with definitions of FM, resulting in a list of FH of the relevant FMs. The disadvantage may be missing the spouse’s relatives and associated FH information. There are some social, behavioral, and environment factors shared in the same household, which also represent critical information. However, these are social determinants of health and not part of the blood relations.

Collecting lexicon entries can be defined as a named entity recognition task, which is an important task for identifying meaningful terms and multiword phrases in free text [27]. In this study, we fine-tuned several BERT-based models for the purpose of identifying potential FH mentions from an EHR corpus, leveraging various data sets for the purpose of providing more context for fine-tuning BERT-based models. Lexicon entries were collected from a large clinical EHR corpus, mitigating the problem of missing entities caused by limited amounts of data. The dictionary coverage evaluation showed that it covers a greater range of lexical variants and focuses primarily on clinical concepts typically reported as part of FH relative to a direct lexicon generated from the UMLS and SNOMED-CT. Our corpus-driven lexicon features FH definitions based on semantic types, concept normalization to UMLS and SNOMED-CT CUIs, and manual curation, with the potential to resolve semantic ambiguity and promote interoperability among various systems. The rule-based FH system also provides standard Health Level 7 FHIR output to foster interoperability.

FH relation extraction is more relevant for downstream analysis compared with entity extraction. In previous challenges, the F1 score was regarded as the most important metric for relation extraction evaluation. The highest F1 score obtained from challenges was 0.5708 in the BioCreative challenge [19] and was 0.681 in the N2C2 challenge [20]. However, it is not our aim to compete with previous studies in terms of F1 scores. As relation extraction is still a challenging task, an FH system that prioritizes recall is highly desired for NLP-assisted curation in EHR-based studies. Our evaluation results showed that the rule-based FH system on top of the corpus-driven lexicon produced higher recall than that obtained with the deep learning–based FH system. In addition, the combined results from both the rule-based module and the deep learning–based FH system resulted in the highest recall for relation extraction, ranging from 0.7109 to 0.8370, which were higher than the recall values obtained in any previous challenge results, ranging from 0.3732 to 0.6810.

Note that we did not observe higher performance when using the reannotated data as compared to the original data. There may be two underlying reasons for this. First, reannotated data have not been used for training the rule-based system. Second, reannotated data were obtained by a professional annotator with deep domain knowledge, which makes the information extraction task harder. In addition, we observed that performance on 2019 N2C2 FH challenge data was worse than that on the 2018 BioCreative FH challenge data. This is mainly because the 2019 N2C2 FH challenge added a certainty attribute (negated or nonnegated) into the relation extraction, which made the relation extraction task harder.

Our FH synthetic data sets used for training and testing were from real clinical sentences, for which the observations, FMs, and ethnicities are shuffled among the whole corpus using a heuristic deidentification process. The granularity of the synthetic FH data sets is the same as that of real FH data. In this study, we have not exhaustively compared all BERT-related models. Theoretically, large language models–empowered knowledge engineering is sufficient for lexicon entry collection from a clinical corpus. Our focus in this study was to provide a corpus-driven lexicon resource that leads to a rule-based FH baseline system for high-throughput analysis, while doing so in a manner that promotes interpretability and explainability for downstream applications.

We recommend that a comprehensive FH system include both a rule-based module and a deep learning–based module to obtain higher recall, which could facilitate manual curation. Although only the rule-based FH system can output normalized concepts, output from a deep learning–based FH system can be a rich source to enrich the lexicon. In the future, we will repeat the SFCP normalization for the output from the deep learning–based FH system to consistently improve the FH lexicon and the rule-based FH system. Meanwhile, we will continue to manually review the BERT-extracted entities without automatic

https://medinform.jmir.org/2023/1/e48072

**JMIR Medical Informatics**

Wang et al

p.270
normalization with frequency under 20, and look into other data sources such as social media so as to expand more concepts and/or term variants for the current lexicon. In addition, we will engage the user community to continuously refine the lexicon. We also plan to update the lexicon and rule-based FH system yearly, which will be distributed through the same open-source repository on GitHub.

There are three limitations to this study. First, although a large corpus with 83,000 patients was used for collection of potential FH variants, there is still a possibility that the FH information is not well represented. In addition, as the lexicon was developed using a largely monoinstitutional data resource, the lexicon may not be generalizable in other institutions. Second, during the entity normalization, we simply adopted an existing standardization framework, as it was not a priority of this study to focus on standardization method development. Third, we arbitrarily set a frequency cutoff of 20 for entities that were not automatically normalized to include in our manual review. However, we realize that some entities with low frequency also have the potential to contribute to the lexicon entries, such as “rectal ca” with a frequency of 18 and “high cholesterol” with a frequency of 12.

In summary, we constructed a corpus-driven FH lexicon to serve as a language resource for FH information extraction. Standardization of concepts in the FH lexicon and the rule-based FH system foster interoperability. The resulting lexicon and the rule-based FH system are freely available as part of the OHNLP Toolkit ecosystem. In the future, we will continue to expand more concepts and/or term variants of the current lexicon, and will explore the incorporation of the system for data curation efforts needed in various EHR-based studies and applications.

Acknowledgments
This work was made possible by National Institutes of Health (NIH) grants (1U01TR002062-01, U24CA194215-01A1, and R01LM013240).

Data Availability
The source code underlying this article is available on GitHub [25].

Authors' Contributions
LW designed the study, implemented models, performed data analysis and manual review, and wrote the manuscript. SM and KJP conducted automatic normalization. HH implemented the models and performed data analysis. AW implemented the baseline system and revised the manuscript. SF revised the manuscript. XA, SL, and RK implemented the models. HL conceptualized and designed the study, performed data analysis, and critically revised the manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Statistical comparison between original and enhanced annotations.

[DOCX File, 17 KB - medinform_v11i1e48072_app1.docx ]

Multimedia Appendix 2
Degree of consanguinity.

[DOCX File, 14 KB - medinform_v11i1e48072_app2.docx ]

Multimedia Appendix 3
Performance of various BERT models fine-tuned on different data sets.

[DOCX File, 15 KB - medinform_v11i1e48072_app3.docx ]

References
3. Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. AMIA Annu Symp Proc 2006;2006:925 [FREE Full text] [Medline: 17238544]


CUI: concept unique identifier
EHR: electronic health record
FAIR: Findable, Accessible, Interoperable, and Reusable
FH: family history
FHIR: Fast Healthcare Interoperability Resource
FM: family member
FN: false negative
N2C2: 2019 NLP Clinical Challenge
NLP: natural language processing
OHNLP: Open Health Natural Language Processing
PURE: Princeton University Relation Extraction
RITE: Reproducible, Implementable, Transparent, and Explainable
SFCP: standardization framework for clinical problems
SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terms
TP: true positive
UMLS: Unified Medical Language System
An Ontology-Based Approach to Improving Medication Appropriateness in Older Patients: Algorithm Development and Validation Study

Elena Calvo-Cidoncha¹, PharmD; Julián Verdinelli², PharmD; Javier González-Bueno³, PharmD, PhD; Alfonso López-Soto⁴, MD, PhD; Concepción Camacho Hernando¹, MBA, PharmD; Xavier Pastor-Duran⁵, MD, PhD; Carles Codina-Jané¹, PharmD, PhD; Raimundo Lozano-Rubi⁶, BSCS, MD, PhD

Corresponding Author:
Elena Calvo-Cidoncha, PharmD

Abstract

Background: Inappropriate medication in older patients with multimorbidity results in a greater risk of adverse drug events. Clinical decision support systems (CDSSs) are intended to improve medication appropriateness. One approach to improving CDSSs is to use ontologies instead of relational databases. Previously, we developed OntoPharma—an ontology-based CDSS for reducing medication prescribing errors.

Objective: The primary aim was to model a domain for improving medication appropriateness in older patients (chronic patient domain). The secondary aim was to implement the version of OntoPharma containing the chronic patient domain in a hospital setting.

Methods: A 4-step process was proposed. The first step was defining the domain scope. The chronic patient domain focused on improving medication appropriateness in older patients. A group of experts selected the following three use cases: medication regimen complexity, anticholinergic and sedative drug burden, and the presence of triggers for identifying possible adverse events. The second step was domain model representation. The implementation was conducted by medical informatics specialists and clinical pharmacists using Protégé-OWL (Stanford Center for Biomedical Informatics Research). The third step was OntoPharma-driven alert module adaptation. We reused the existing framework based on SPARQL to query ontologies. The fourth step was implementing the version of OntoPharma containing the chronic patient domain in a hospital setting. Alerts generated from July to September 2022 were analyzed.

Results: We proposed 6 new classes and 5 new properties, introducing the necessary changes in the ontologies previously created. An alert is shown if the Medication Regimen Complexity Index is ≥40, if the Drug Burden Index is ≥1, or if there is a trigger based on an abnormal laboratory value. A total of 364 alerts were generated for 107 patients; 154 (42.3%) alerts were accepted.

Conclusions: We proposed an ontology-based approach to provide support for improving medication appropriateness in older patients with multimorbidity in a scalable, sustainable, and reusable way. The chronic patient domain was built based on our previous research, reusing the existing framework. OntoPharma has been implemented in clinical practice and generates alerts, considering the following use cases: medication regimen complexity, anticholinergic and sedative drug burden, and the presence of triggers for identifying possible adverse events.

(JMIR Med Inform 2023;11:e45850) doi:10.2196/45850

KEYWORDS
biological ontologies; decision support systems; inappropriate prescribing; elderly; medication regimen complexity; anticholinergic drug burden; trigger tool; clinical; ontologies; pharmacy; medication; decision support; pharmaceutic; pharmacology; chronic condition; chronic disease; domain; adverse event; ontology-based; alert
**Introduction**

Medical advances have resulted in a rise of life expectancy. The prevalence of multimorbidity, which is defined as the coexistence of 2 or more chronic conditions, tends to be higher among older people [1]. As a result, the use of multiple medicines, which is commonly referred to as polypharmacy, has become a common phenomenon in this population [2,3].

Polypharmacy increases the risk of inappropriate medication [4,5], leading to a greater risk of adverse drug events (ADEs) [6]. ADEs are associated with hospital admissions, higher mortality rates, and increased health care expenditures [7-10]; therefore, improving medication appropriateness in older patients with multimorbidity is a priority [11].

One approach to improving medication appropriateness is to use clinical decision support systems (CDSSs) for assistance during the prescription process. CDSSs are intended to improve health care delivery by enhancing medical decisions with targeted clinical knowledge and patient information [12,13]. Relational databases are the predominant choice when it comes to designing a CDSS. However, due to the main challenges of CDSSs, such as the lack of interoperability or alert fatigue [14-16], there is increasing interest in using ontology-based CDSSs to overcome these challenges. An ontology is an explicit conceptualization of the entities of a domain [17,18]. Because ontologies add semantics to models, they enhance the reusability of data and are more efficient in dealing with changing requirements and maintenance requirements [19-21].

Previously, we used Protégé-OWL (Stanford Center for Biomedical Informatics Research) to develop OntoPharma—an ontology-based CDSS for reducing medication prescribing errors [22]. The domains addressed by OntoPharma include the identification and technical data of medicinal products, as well as data on drug appropriateness for ensuring the safe use of medicines. These domains were addressed in the following four use cases: maximum dosage alerts, a drug-drug interaction checker, renal failure adjustment, and a drug allergy checker. OntoPharma is currently implemented in a tertiary referral hospital.

Alerts generated by OntoPharma are, nowadays, commonly available. To leverage the ease of using ontologies to represent rich and complex knowledge, the modeling of drug knowledge that is absent in usual commercial databases is needed.

For this reason and on the basis of our previous research, the primary aim of this study was to model a domain for improving medication appropriateness in older patients with multimorbidity (hereinafter called the chronic patient domain). The secondary aim was to implement the version of OntoPharma containing the chronic patient domain in a hospital setting.

**Methods**

This study was conducted between 2020 and 2022 at a 710-bed tertiary hospital in Spain, which was equipped with computerized physician order entry (CPOE) and an electronic health record (EHR) system provided by SAP SE. The following 4-step development process was designed: defining the domain scope, representing the domain in the model, adapting the OntoPharma-driven alert module, and implementing the version of OntoPharma containing the chronic patient domain in a hospital setting.

**Ethics Approval**

This study was approved by the ethics committee of the Hospital Clinic of Barcelona (reference number HCB/2019/0735).

**Defining the Domain Scope**

**Overview of the Domain Scope**

The chronic patient domain focused on improving medication appropriateness in older patients with multimorbidity. Given the dimension of the domain, we decided to establish an expert advisory panel to limit the scope of the domain. The group of experts included geriatricians and clinical pharmacists, of whom all were members of the C3RG (Central Catalonia Chronicity Research Group) and had expertise in ensuring medication appropriateness in older patients with multimorbidity. Focus group sessions yielded consensus on the importance of the following three use cases: medication regimen complexity, anticholinergic and sedative drug burden, and the presence of triggers for identifying possible adverse events.

**Medication Regimen Complexity**

Complex medication regimens are challenging for patients, which may impact medication adherence and safety [23,24]. The Medication Regimen Complexity Index (MRCI), which was developed by George et al [25], is currently the most widely used scale for assessing medication regimen complexity. Medication complexity considers more factors than a simple medication count. The MRCI consists of 65 items, including weighted scores for types of prescribed dosage forms (section A), dosing frequency (section B), and additional administration instructions (section C). The sum of the scores of the three sections provides a total score, with higher scores indicating greater regimen complexity.

The MRCI has been translated and validated for other languages, including Spanish (Spanish MRCI [MRCI-E]) [26]. We used the MRCI-E as a source of information.

Section A provides weights for 32 dosage form and administration route combinations. For example, an oral tablet medication is given a weight of 1. More complex combinations result in higher weights.

Section B provides weights for 23 dosing frequencies (“scheduled” or “as needed”). The “once daily” frequency is used as the baseline (weight of 1), on which the other weightings are built.

Section C provides weights for 10 additional instructions that a patient may need to follow in adhering to a prescribed regimen. Additional administration instructions are related to taking medication at specific times, taking medication in relation to food, taking multiple units at one time, and needing to break or crush a tablet or needing to taper or increase a dose.
Anticholinergic and Sedative Drug Burden

Anticholinergic burden is defined as the cumulative effect of taking 1 or more drugs that are capable of causing anticholinergic adverse effects, and the load increases with the number of medications prescribed [27]. Anticholinergic toxicity is a common problem in older people. Anticholinergic effects are associated with peripheral manifestations (urinary retention, constipation, decreased secretions, etc) and central manifestations (delirium, cognitive disorders, and functional disorders) [27,28].

Several tools have been developed to estimate anticholinergic burden by giving a score to drugs according to their anticholinergic potential [29]. The Drug Burden Index (DBI) is the only scale that accounts for a patient’s dose [30]. In addition, the DBI considers not only anticholinergic effects but also sedative effects. The total DBI exposure is calculated as the sum of exposure to any DBI medication, according to the following formula:

\[
D = \sum_{i=1}^{n} d_i
\]

where “D” is the daily dose taken and “\(d_i\)” is the minimum effective daily dose for that drug.

Byrne et al [31] provided a master DBI list containing a final list of DBI medications and their minimum effective daily doses. The master DBI list included 156 entries. Each entry consisted of the following fields: drug description (ingredient), World Health Organization Anatomical Therapeutic Classification codes, anticholinergic and sedative effects, and minimum effective daily dose (expressed as mg) by route of administration (parenteral, sublingual, buccal, transdermal, rectal, and inhalation).

Triggers

A trigger is defined as a flag, occurrence, or prompt that alerts reviewers to initiate further in-depth investigations regarding a patient’s record to determine the presence or absence of an adverse event [32]. An example of a trigger is a potassium level of \(<2.9\) mEq/L in a patient with loop diuretics. Triggers are based on the assumption that any new condition may be due to the use of a drug. Multiple sets of triggers have been developed. Guzmán et al [33] identified the most appropriate triggers for detecting ADEs in older patients with multiple chronic conditions.

The trigger set developed by Guzmán et al [33] included a total of 51 entries. Each entry consisted of the following fields: high-alert medications for patients with chronic illnesses (therapeutic class or ingredient) and triggers for detecting potential ADEs (11 care module triggers, 9 antidote- and treatment-based triggers, 11 medication concentration–based triggers, 18 triggers based on abnormal laboratory values, and 1 emergency department trigger).

Domain Model Representation

Data sets were not organized in a predefined format. Prior to modeling the chronic patient domain through ontologies, we processed all of the information in a relational database to clean the data, detect redundancies, and detect relationships between different concepts.

To add this new domain to OntoPharma, we built on our previous research by using the existing framework, which was composed of 3 ontologies (Drugs, Decision support system [DSS], and Local pharmacy) [22]. The Drugs ontology was designed to represent the identification and technical data of medicinal products. The DSS ontology provides data on drug appropriateness. The Local pharmacy ontology was designed to represent local concepts from EHRs and CPOE in order to ensure interoperability.

The design, development, and maintenance of the chronic patient domain was driven by medical informatics specialists and clinical pharmacists. The information was represented in the Web Ontology Language (OWL) [34]. For encoding the OWL ontologies, we used the Protégé 3.5 editor tool [35]. The concepts of the chronic patient domain were organized hierarchically, following a top-down approach, as we did with all previous domains of OntoPharma. The development of the class hierarchy, the defining of properties, and the slotting of concepts were carried out at the same time. Finally, we defined individual instances of the classes represented.

OntoPharma-Driven Alert Module Adaptation

We reused the OntoPharma-driven alert module that was proposed in our previous research [22]. The integration between the CPOE system and the ontologies was performed through a REST API. A REST API call was published (in JSON format) each time a clinician added a new medication in the CPOE system, modified an existing one, or requested on-demand CDSS information. The request contained patient-specific clinical data. SPARQL (Apache Jena Fuseki server) was used to query ontologies [36]. After applying the queries, a returning REST API, with the results, was published.

It was necessary to update the content of the REST API published each time OntoPharma was triggered. We specifically had to add more laboratory parameters (to date, the only one considered was glomerular filtration rate). New local concepts were manually mapped with existing concepts in the ontologies. In addition, we created new SPARQL queries to ensure the safe use of medicines in older patients.

Alerts were shown in the CPOE system in cases of high medication regimen complexity, in cases of high anticholinergic and sedative drug burden, or in cases where triggers for detecting ADEs in older patients were present. In addition, patients were required to be older than 65 years.

In accordance with the recommendations of end users, the user interface proposed in the previous paper [22] was slightly modified to ensure usability and minimal interference with the clinician’s workflow.

Formal testing was performed to demonstrate that the new version of the ontology-driven alert module met functional requirements. In addition, clinical pharmacists performed manual testing in a control environment (the SAP quality assurance server) to evaluate whether the alert module functioned properly when generating the prescribing alerts.
Implementation of the Version of OntoPharma Containing the Chronic Patient Domain in a Hospital Setting

In July 2022, the version of OntoPharma containing the chronic patient domain was implemented at one ward of the internal medicine unit, which had capacity for 20 admissions. Informatics staff and clinical pharmacists were responsible for the diffusion and for providing support.

A retrospective analysis of the alerts generated was performed. We included patients who were admitted to the internal medicine ward from July to September 2022. The following patient data were collected: gender, age, duration of hospital stay, and number of medications during hospital stay. We further examined the alerts, including the number of alerts, the types of alerts, clinical relevance, and the acceptance rates.

Quantitative variables were expressed as means and SDs for variables with a normal distribution or as medians and IQRs for variables with a skewed distribution. Qualitative variables were expressed as percentages. Data analysis was carried out by using SPSS 20.0 (IBM Corp).

Results

Knowledge Representation Using Ontologies

Overview of Ontologies

For modeling the chronic patient domain, we proposed new classes and properties, introducing the necessary changes in the ontologies previously created (Drugs, DSS, and Local pharmacy). The three ontologies are interconnected. The import schema of ontologies is shown in Figure 1.

Figures 2-4 provide diagrams showing the relationships between classes for defining the chronic patient domain in the Drugs ontology, DSS ontology, and Local pharmacy ontology, respectively.

Multimedia Appendix 1 contains a list of the medication knowledge concepts and their definitions, which were used to define the chronic patient domain. Multimedia Appendix 2 contains a list of properties and their facets, which were also used to define the chronic patient domain.
Figure 1. Import schema of the ontologies used in OntoPharma. For modeling drug-related knowledge, 3 ontologies have been developed (Drugs, DSS, and Local pharmacy). Each ontology has been divided into 2 parts. The first part provides concepts and classes (also known as T-Box), and the second provides the instances of these concepts (also known as A-Box). The three ontologies are interconnected, as shown in Figure 1. DSS: Decision support system.
Figure 2. Diagram showing the relationships between classes in the Drugs ontology for defining the chronic patient domain. The Drugs ontology was designed to represent the identification and technical data of medicinal products.
Figure 3. Diagram showing the relationships between classes in the DSS ontology for defining the chronic patient domain. The DSS ontology provides data on drug appropriateness. Concepts that were specifically created to define the chronic patient domain are highlighted in yellow. DBI: Drug Burden Index; DSS: Decision support system; MRCI: Medication Regimen Complexity Index.
Figure 4. Diagram showing the relationships between classes in the Local pharmacy ontology for defining the chronic patient domain. The Local pharmacy ontology was designed to represent local concepts from electronic health records and computerized physician order entry. Each local concept is mapped to the corresponding OntoPharma concept.

Medication Regimen Complexity

The MRCI quantifies drug regimen complexity based on dosage form, dosage frequency, and additional instructions.

To represent the weighted scores for types of prescribed dosage forms (section A), we first created the concept "MRCI A form" (DSS ontology), which comprises 30 subclasses for identifying the possible dosage form and route of administration combinations. To provide the weight for each dosage form and route of administration combination, we introduced the property "mrci A weight.”

To represent the weighted scores for dosage frequency (section B), we introduced the following two attached properties within the class “Local frequency” (Local pharmacy ontology): “mrci B,” which provides weights for “scheduled” dosing frequencies, and “mrci B PRN,” which provides weights for “as needed” dosing frequencies.

We identified 231 distinct frequency combinations. Frequency weights were assigned, considering that frequency data also contained indicators that qualify for component C scoring, such as indicators to take medication less often than once per day (e.g., once every 48 hours) or indicators to take medication at specific times (before a meal, at bedtime, etc).
To represent the weighted scores for additional administration instructions (section C) related to taking medication with or without food, we introduced 1 attached property (“mrci C”) within the class “Virtual medicinal product (VMP)” (Drugs ontology). A virtual medicinal product is an abstract representation of an active medicinal ingredient associated with strength information and a route of administration (eg, “omeprazole 20mg capsule”). We assigned a total of 6257 weights. Figure 5 provides a class diagram to model medication regimen complexity, which is explained with an example.

**Figure 5.** Class diagram to model medication regimen complexity. Circles represent the classes needed to quantify drug regimen complexity. Squares represent the attached properties (object or data) within each class. An example is given in brackets. MRCI: Medication Regimen Complexity Index; VMP: virtual medicinal product.

---

**Anticholinergic and Sedative Drug Burden**

The OWL concept that was used to enter the data on anticholinergic and sedative drug burden was “DBI” (DSS ontology), in reference to the scale used for its calculation. Because the DBI is a dose-related measure of anticholinergic and sedative drug exposure, we created the “DBI” concept as a subclass of the “Dose appropriateness” concept. To provide enough information to calculate the DBI, we introduced the property “medd,” which describes the minimum effective daily dose of each drug.

The “DBI” class contains 164 individuals. Each individual contains the following knowledge: ingredient (eg, alprazolam),
route of administration (eg, oral), age range (eg, 65-999 years), minimum effective daily dose (eg, 0.5), unit (eg, mg), base unit (eg, every 24 hours), and alert-related data. **Figure 6** provides a class diagram to model anticholinergic and sedative drug burden, which is explained with an example.

**Triggers**

The OWL concept that was used to enter the triggers for detecting ADEs in older patients with multiple chronic conditions was “Trigger tool” ([DSS ontology](https://medinform.jmir.org/2023/1/e45850)). We created the “Trigger tool” concept as a subclass of the “Appropriateness lab test” concept because we only included triggers based on abnormal laboratory values. Introducing new properties was not required.

The “Trigger tool” class contains 821 individuals. Each individual contains the following knowledge: ingredient (eg, furosemide), route of administration (eg, parenteral), age range (eg, 65-999 years), lab test (eg, serum glucose), lab test unit (eg, mg/dL), low value (eg, 0), high value (eg, 110), and alert-related data. **Figure 7** provides a class diagram to model triggers for detecting ADEs in patients with multimorbidity, which are explained with an example.
**Alerts**

In addition to the abovementioned actions, we made changes related to the “Alert” class (DSS ontology). The “Alert” class included the information that was displayed when appropriateness criteria were not met. The following fields related to the “Alert” class remained unchanged: alert description, alert recommendation, alert source, alert date (the date when the alert was last updated), and related information (supporting documentation). In addition to the existing instances of the “Alert level” class (“not recommended,” “contraindicated,” or “unallowed prescription”), we created a new one called “risk minimization.” We also introduced a new class—the “Drug intervention” subclass of the “Alert” class (DSS ontology). We defined the following two types of drug intervention: “global drug” intervention for when a complete treatment revision is required (eg, high MRCI) and “specific drug” intervention for when a partial treatment revision is required (eg, high DBI).

**Knowledge Not Represented Using Ontologies**

With regard to medication regimen complexity, we did not represent complexity based on the additional instructions related to taking multiple units at one time or needing to break or crush a tablet.
The triggers used to identify ADEs can be abnormal laboratory values, the use of certain medications or antidotes, or changes in clinical status that may indicate a possible medication-related harm. We only represented triggers based on abnormal laboratory values.

**OntoPharma-Driven Alert Module Adaptation**

The OntoPharma-driven alert module works the same as it did in our previous study [22]. Once the patient-specific clinical data are sent from the CPOE system and EHR to the ontologies, local concepts are matched to their equivalent OntoPharma concepts. With regard to the chronic patient domain, we defined the following decision rules.

The MRCI is obtained by summing the scores of the three sections.

The section A score is estimated by considering the dosage form and route of administration combination. Because administering the same dosage form more than once is easier than administering different dosage forms, each dosage form and route of administration combination is counted only once within a regimen. For example, if a patient’s regimen consists of taking 3 tablets orally, their component A score is 1, not 3.

Section B and C scores are estimated by considering the dosage frequency and the virtual medicinal product prescribed, respectively. The cutoff point selected for triggering an alert (MRCI≥40) was determined by the expert advisory panel, in accordance with the literature.

The total DBI is calculated with the equation $\delta = \frac{D}{\delta}$, where “$D$” is the daily dose taken by the individual patient and “$\delta$” is the minimum effective daily dose for that drug. The daily dose taken for each DBI medication is estimated by considering the dose, dose unit, and frequency. We have defined conversion factors for cases where the drug dose unit prescribed is different from the unit dose defined in the ontologies. The minimum effective daily dose is represented in the DSS ontology for each ingredient and route of administration combination. The cutoff point selected for triggering an alert (DBI≥1) was determined by the advisory panel, in accordance with the literature.

To evaluate the presence of triggers for identifying ADEs, we consider the ingredient regardless of the dosage. If a patient has several laboratory values, we consider the most recent values. An alert is triggered when a value is outside of the defined range [33].

Medications prescribed “as needed” were not considered in previous cases.

With regard to the interface, alerts are shown in different colors (red, orange, and yellow) according to their clinical relevance (contraindicated, moderate relevance, and low relevance). We added a new label (blue) to identify alerts aimed at risk minimization. To date, the advisory text contains the generic drug name, a short description of the possible concern, and a recommendation for improving medication appropriateness. The generic drug name is still displayed if the alert requires “specific drug” intervention. In cases where the alert requires “global drug” intervention, the text “Review total treatment” is displayed. We also included a hyperlink to relevant literature.

Alerts related to the chronic patient domain were defined, such as soft-stop alerts, so that the clinician can decide whether to ignore or accept the alert. To avoid alert fatigue, if an alert is ignored once, it will not be displayed again.

The interface that displays the alerts also includes a link to a user guide and an activity registry that serves as a traceability system.

Despite the addition of new use cases, the results show that the response time for generating decision support remains short (within milliseconds), with minimal impact on the user’s workflow.

**Implementation of the Version of OntoPharma Containing the Chronic Patient Domain in a Hospital Setting**

A total of 107 patients were included. The median age was 86 (IQR 80-90) years, and the majority of patients were women (n=63, 58.9%). The median length of hospital stay was 8 (IQR 5-13) days. Patients had a median of 15 (IQR 11-19) medications.

Of the 107 patients, 96 (89.7%) received at least one alert. OntoPharma generated 364 alerts (mean 3.9, SD 5.3 alerts per patient). Of these, 296 (81.3%) alerts were considered of low (contraindicated, moderate relevance, and low relevance). We included a new label (blue) to identify alerts aimed at risk minimization. Further, 154 (42.3%) alerts were accepted.

Details of the types of alerts and the acceptance rates are included in Table 1. The most frequent alerts were alerts due to high anticholinergic and sedative drug burden (231/364, 63.5%), followed by alerts due to high medication regimen complexity (68/364, 18.7%) and alerts due to the presence of triggers (65/364, 17.8%).

### Table. Description of the types of alerts generated by OntoPharma and the acceptance rates.

<table>
<thead>
<tr>
<th>Type of alert</th>
<th>Frequency (N=364), n (%)</th>
<th>Acceptance rate, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medication regimen complexity</td>
<td>68 (18.7)</td>
<td>40 (58.8)</td>
</tr>
<tr>
<td>Anticholinergic and sedative drug burden</td>
<td>231 (63.5)</td>
<td>84 (36.4)</td>
</tr>
<tr>
<td>Triggers</td>
<td>65 (17.8)</td>
<td>30 (46.2)</td>
</tr>
</tbody>
</table>

*Percentages were calculated by using the numbers in the “Frequency” column as denominators.
Principal Results

This paper presents a modeling approach, which was formalized in ontological terms, for defining the chronic patient domain that provides support for improving medication appropriateness in older patients with multimorbidity. The chronic patient domain was built on OntoPharma—an ontology-based CDSS for reducing medication prescribing errors that has already been implemented in a tertiary referral hospital [22].

There are already ontology-based CDSSs that address medication management in patients with chronic conditions [37]. However, to the best of our knowledge, this is the first ontology-based approach that models medication regimen complexity, anticholinergic and sedative drug burden, and triggers for identifying possible adverse events. Farrish and Grando [38] built an ontology to assist with the management of polypharmacy prescriptions for patients with multiple chronic conditions to reduce the overall treatment complexity. Recently, Román-Villarán et al [39] developed an ontology-based CDSS for patients with complex chronic conditions. However, the knowledge sources used were different from ours, including clinical practice guidelines, the LESS-CHRON (List of Evidence-Based Deprescribing for Chronic Patients) criteria, and the STOPP/START (Screening Tool of Older Persons’ Prescriptions and Screening Tool to Alert to Right Treatment) criteria, among others. These ontology approaches for patients with chronic conditions have been validated with patient data from databases. However, it is important to note that they are not implemented in a real environment, unlike our ontology approach. OntoPharma provides rapid and real-time support to improve medication appropriateness in older patients with multimorbidity.

Using ontologies instead of relational databases, which are the predominant choice in current commercial CDSSs, has distinct advantages [40,41]. First, the semantic approach and the use of OWL enable a convenient infrastructure for reuse. Hence, we reused the existing OntoPharma framework, without having to start from scratch. In addition, ontologies are more flexible and efficient in dealing with changes; thus, it was possible to add a new domain to OntoPharma without major complications. We were able to model a complex domain, creating only 6 new classes and 5 new properties. This was possible because the three ontologies (Drugs, DSS, and Local pharmacy) are interconnected (Figure 1), and classes are linked between them through object properties.

To ensure flexibility, scalability, and sustainability, we operated on the most appropriate level of abstraction. To define anticholinergic drug burden and triggers, we considered the ingredient. However, to define weighted scores for additional administration instructions (MRCI section C), we considered the class “Virtual medicinal product (VMP).”

To integrate structured clinical data with clinical knowledge, we reused the mappings previously established in the ontology Local pharmacy. It was only necessary to add some new mappings related to laboratory parameters.

End users participated throughout the development of the chronic patient domain in order to ensure usability and gain user acceptance [42,43]. As a result, we introduced some changes in the user interface, such as new clinical relevance levels and a hyperlink to relevant literature. Some proposals for improvement, such as showing the laboratory values next to the alert, have not been implemented yet. Usability may also be influenced by the response time for generating decision support. However, response time has not been modified, showing that OntoPharma is scalable.

Limitations

In terms of evaluation, we have not identified a database-based system for direct comparison with OntoPharma. van der Sijs et al [44] conducted a systematic review, concluding that drug safety alerts are overridden by clinicians in 49% to 96% of cases. Our acceptance rate (154/364, 42.3%) was expected to be better, considering that an expert advisory panel selected the most useful information to improve medication appropriateness in older patients. This may be partly explained by the following limitations. First, appropriateness criteria were evaluated if patients were older than 65 years. Considering that the older population is heterogeneous, we should also have considered frailty—a known factor indicative of vulnerability to medication-related problems [45]. Second, the alerts with a lower acceptance rate (84/231, 36.4%) were related to the DBI. Interventions for reducing the DBI commonly involve progressive medication deprescribing, which is difficult to realize in a tertiary hospital and would be easier in intermediate care [46]. On the other hand, poor adherence is one of the major consequences of high MRCI scores [23]. In hospitals, the administration of medications is primarily the nurses’ responsibility; therefore, clinicians may have not given sufficient importance to MRCI alerts. Acceptance rates might improve in outpatient care. Our research focused primarily on clinician decision-making. The variables analyzed allowed us to identify the scale of potentially inappropriate medications and the usefulness of OntoPharma. However, evaluating OntoPharma’s influence on health outcomes is a challenge that we should take up in future.

As mentioned in our previous paper on OntoPharma [22], one limitation of this study is maintaining the evidence and keeping it relevant and up to date [47]. To create individual instances, we extracted the information from papers. Since this was not done via automatic extraction, it was a time-consuming process. With regard to the maintenance, we must assign a complexity weight if there is a new dose form or dosing frequency; these data are not updated frequently. In addition, we must check if new medications have additional administration instructions, are capable of causing anticholinergic adverse effects, or are included in the set of triggers for detecting ADEs in older patients.

Of note, although mapping in this study did not take excessive time, we are aware that manual mapping is a resource-intensive and ongoing process.

We have not represented all of the knowledge from the sources of information. OntoPharma relies on structured data; therefore, we have prioritized representing data that are structured in text
format within the EHR. As a result, medication regimen complexity may be underestimated because special instructions are underrepresented. In addition, there are triggers that are different from abnormal laboratory values that are not represented in ontologies. Even though there exist large amounts of health care data, the main challenge to improving results of CDSSs is converting free-text data into structured fields computationally [48].

In future implementations, we will continue to represent complex drug knowledge that is absent in commercial databases. We are currently modeling knowledge for supporting the neonatal population and populations at risk for hepatitis B virus reactivation. To capture clinicians’ reasoning processes, we must place a high priority on increasing structured patient data within EHRs.

Other areas for future work are mentioned in our previous OntoPharma paper [22], such as developing a more complex CDSS that can be applied across the entire treatment process and is not only restricted to the medication prescription process. Finally, we must continue working on customized alerts to avoid alert fatigue [49].

Despite the limitations, we believe that our methods have been successful in modeling knowledge related to the chronic patient domain and that the proposed version of OntoPharma is an enhancement of the previous one. Although optimizing care in older patients is a context-dependent, complex process, we believe that developing an ontology to support the chronic patient domain constitutes a major step toward improving medication appropriateness in a generalizable and reusable way.

Conclusions
Polypharmacy in the older population poses challenges to the delivery of medical care because of the increased difficulties in guaranteeing appropriate prescription. We proposed an ontology-based approach to provide support for improving medication appropriateness in older patients with multimorbidity in a scalable, sustainable, and reusable way. OntoPharma has been implemented in clinical practice and generates alerts, considering the following use cases: medication regimen complexity, anticholinergic and sedative drug burden, and the presence of triggers for identifying possible adverse events.

Acknowledgments
We thank Central Catalonia Chronicity Research Group (C3RG)-Line of research “Person-Centred Prescription” for providing advice in the conceptualization of the use cases. We especially thank Joan Espaulella-Panicot, Daniel Sevilla-Sánchez, and Núria Molist-Brunet.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Medication knowledge concepts that are represented in OntoPharma to define the chronic patient domain.

[DOCX File, 22 KB - medinform_v11i1e45850_app1.docx ]

Multimedia Appendix 2
Properties and their facets, which are represented in OntoPharma to define the chronic patient domain.

[DOCX File, 20 KB - medinform_v11i1e45850_app2.docx ]

References


34. Deborah L, van Harmelen F. W3C. OWL Web Ontology Language overview. 2004. URL: www.w3.org/TR/owl-features/ [accessed 2023-01-17]


Abbreviations

ADE: adverse drug event
C3RG: Central Catalonia Chronicity Research Group
CDSS: clinical decision support system
CPOE: computerized physician order entry
Agreement Between Experts and an Untrained Crowd for Identifying Dermoscopic Features Using a Gamified App: Reader Feasibility Study

Jonathan Kentley\textsuperscript{1,2*}, MSc, MBBS; Jochen Weber\textsuperscript{2*}; Konstantinos Liopyris\textsuperscript{3}, MD, PhD; Ralph P Braun\textsuperscript{4}, MD; Ashfaq A Marghoob\textsuperscript{2}, MD; Elizabeth A Quigley\textsuperscript{2}, MD; Kelly Nelson\textsuperscript{5}, MD; Kira Prentice\textsuperscript{6}, BA; Erik Duhaime\textsuperscript{6}, MPhil, PhD; Allan C Halpern\textsuperscript{2}, MD; Veronica Rotemberg\textsuperscript{2}, MD, PhD

\textsuperscript{1}Department of Dermatology, Chelsea and Westminster Hospital, London, United Kingdom
\textsuperscript{2}Dermatology Section, Memorial Sloan Kettering Cancer Center, New York, NY, United States
\textsuperscript{3}Department of Dermatology, Andreas Syggros Hospital of Cutaneous and Venereal Diseases, University of Athens, Athens, Greece
\textsuperscript{4}Department of Dermatology, University Hospital Zurich, Zurich, Switzerland
\textsuperscript{5}Department of Dermatology, The University of Texas MD Anderson Cancer Center, Houston, TX, United States
\textsuperscript{6}Centaur Labs, Boston, MA, United States

*these authors contributed equally

Corresponding Author:
Veronica Rotemberg, MD, PhD
Dermatology Section
Memorial Sloan Kettering Cancer Center
530 E 74th Street
New York, NY, 10021
United States
Phone: 1 8336854126
Email: rotembev@mskcc.org

Abstract

Background: Dermoscopy is commonly used for the evaluation of pigmented lesions, but agreement between experts for identification of dermoscopic structures is known to be relatively poor. Expert labeling of medical data is a bottleneck in the development of machine learning (ML) tools, and crowdsourcing has been demonstrated as a cost- and time-efficient method for the annotation of medical images.

Objective: The aim of this study is to demonstrate that crowdsourcing can be used to label basic dermoscopic structures from images of pigmented lesions with similar reliability to a group of experts.

Methods: First, we obtained labels of 248 images of melanocytic lesions with 31 dermoscopic “subfeatures” labeled by 20 dermoscopy experts. These were then collapsed into 6 dermoscopic “superfeatures” based on structural similarity, due to low interrater reliability (IRR): dots, globules, lines, network structures, regression structures, and vessels. These images were then used as the gold standard for the crowd study. The commercial platform DiagnosUs was used to obtain annotations from a nonexpert crowd for the presence or absence of the 6 superfeatures in each of the 248 images. We replicated this methodology with a group of 7 dermatologists to allow direct comparison with the nonexpert crowd. The Cohen $\kappa$ value was used to measure agreement across raters.

Results: In total, we obtained 139,731 ratings of the 6 dermoscopic superfeatures from the crowd. There was relatively lower agreement for the identification of dots and globules (the median $\kappa$ values were 0.526 and 0.395, respectively), whereas network structures and vessels showed the highest agreement (the median $\kappa$ values were 0.581 and 0.798, respectively). This pattern was also seen among the expert raters, who had median $\kappa$ values of 0.483 and 0.517 for dots and globules, respectively, and 0.758 and 0.790 for network structures and vessels. The median $\kappa$ values between nonexperts and thresholded average–expert readers were 0.709 for dots, 0.719 for globules, 0.714 for lines, 0.838 for network structures, 0.818 for regression structures, and 0.728 for vessels.

Conclusions: This study confirmed that IRR for different dermoscopic features varied among a group of experts; a similar pattern was observed in a nonexpert crowd. There was good or excellent agreement for each of the 6 superfeatures between the
crowd and the experts, highlighting the similar reliability of the crowd for labeling dermoscopic images. This confirms the feasibility and dependability of using crowdsourcing as a scalable solution to annotate large sets of dermoscopic images, with several potential clinical and educational applications, including the development of novel, explainable ML tools.

(JMIR Med Inform 2023;11:e38412) doi:10.2196/38412

KEYWORDS
dermatologist; diagnosis; diagnostic; labeling; classification; deep learning; dermoscopy; dermatoscopy; skin; pigmentation; microscopy; dermoscopic; artificial intelligence; machine learning; crowdsourcing; crowdsourced; melanoma; cancer; lesion; medical image; imaging; development; feasibility

Introduction

The use of dermoscopy, a low-cost, noninvasive diagnostic technique based on a hand-held device with a light source and magnifying lens, is routine practice for the evaluation of pigmented skin lesions and has been shown to increase sensitivity for early melanoma detection [1,2]. Dermoscopy allows examination of morphological features below the stratum corneum that would not be visible by visual inspection alone [3]. Diagnosis of melanoma using dermoscopy relies on assessment of lesion morphology and identification of dermoscopic features. A number of diagnostic criteria and algorithms have been developed for this purpose, including pattern analysis [4], the ABCD (asymmetry, border, color, diameter) rule [5], the Menzies method [6], the 7-point checklist [7], and the CASH (color, architecture, symmetry, homogeneity) score [8].

As use of dermoscopy has expanded, so too has dermoscopic vocabulary, resulting in a vast number of published feature definitions and 2 competing terminologies: metaphoric and descriptive. In recent years, efforts have been made to harmonize nomenclature, and the 2016 International Dermoscopy Society terminology consensus proposed 31 specific “subfeatures” of melanocytic lesions, falling into 9 “superfeatures” based on structural similarities (Textbox 1) [9].

However, interrater reliability (IRR) for identifying melanoma-specific dermoscopic structures has been shown to be poor [10]. Our research group recently performed the EASY (Expert Agreement on the Presence and Spatial Location of Melanocytic Features in Dermoscopy) study, which found that agreement was highly variable when 20 dermoscopy experts were asked to identify the 31 dermoscopic subfeatures in an image set specifically curated for this purpose. IRR across 248 images was poor to moderate for all 7 features. We demonstrated that when individual subfeatures were collapsed into 9 superfeatures, increased agreement was observed, ranging from a pairwise Fleiss κ of 0.14 for the detection of dots to 1.0 for the detection of a pigment network structure.

Machine learning (ML) methods have recently been investigated in the field of dermatology, and the majority of developed algorithms are diagnostic binary classifiers [11,12]. A number of studies have evaluated the performance of algorithms developed to detect specific dermoscopic features, including pigment network structures, vessels, and blue-white veil; however, many algorithms were trained and tested on relatively small data sets and have achieved only moderate accuracy [13-21].

Due to the vast dimensionality of medical images, classifier algorithms are typically of an uninterpretable “black box” nature, a term that describes the phenomenon whereby functions that connect input pixel data to output labels cannot be understood by the human brain. There has been a push by medical regulators and the artificial intelligence community to develop explainable algorithms; however, it has been acknowledged that this may come at the cost of decreased accuracy [22]. Incorporating detection of dermoscopic features into melanoma classifier algorithms may allow for better explainability and therefore greater acceptance into clinical practice by clinicians and regulatory bodies [23,24].

The International Skin Imaging Collaboration (ISIC) archive provides an open-access data set comprising almost 70,000 publicly available dermoscopic images at the time of writing, including 5598 melanomas and 27,878 nevi. As well as hosting the regular ISIC Grand Challenge to promote the development of ML for melanoma detection, the archive has been extensively utilized to train independent ML algorithms and acts as a comprehensive educational resource for dermatologists via the Dermoscopedia platform [25,26]. Most public images in the archive have labels serving as a diagnostic ground truth for supervised learning. However, accurate feature annotations are thus far lacking. As part of the 2018 ISIC Challenge, 2595 images were annotated for 5 dermoscopic patterns (pigment network structures, negative network structures, streaks, milia-like cysts, and dots/globules) [27]. However, the ground truth labels were provided by only 1 clinician and the performance of the 23 submitted algorithms was acknowledged to be exceptionally low, likely as a result of this [27].

As medical data sets continue to rapidly expand and computing power increases, it is widely recognized that one of the major limiting factors for the development of robust and generalizable ML in dermatology is the need for large, comprehensively labeled data sets [28,29]. Obtaining annotations of medical images by medical experts is both time-consuming and expensive, creating a bottleneck in the development pipeline and making it challenging to obtain annotations at scale [30].

Crowdsourcing provides a potential solution to these problems. Crowdsourcing involves the recruitment of groups of individuals of varying levels of knowledge, heterogeneity, and number who voluntarily complete an online task, often with financial incentives [31,32]. Monetary compensation is typically less than US $0.10 per annotation, and tasks can be distributed to a...
large number of workers in parallel, aggregating the crowd’s knowledge to complete the task in a cost- and time-effective manner [33,34]. One study reported that it took 6 months to obtain expert labels comprising 340 sentences from radiology reports written by 2 radiologists, whereas the authors obtained crowdsourced annotations of 717 sentences in under 2 days at a cost of less than $600. A classification algorithm trained using these crowdsourced annotations outperformed an algorithm trained using the expert-labeled data as a result of the increased volume of available training examples [32].

Given the heterogeneity of biomedical data, the utility of crowdsourcing may decrease with the complexity of the task. For example, the 14 million images contained in the ImageNet archive were easily annotated by the untrained public, whereas the ability to classify and segment radiological images may require many years of specialist training [28,30,35]. Nevertheless, crowdsourcing has proven effective in a wide range of applications for biomedical imaging, most commonly histopathology or retinal imaging [34].

Feng et al [36] reported that a crowd of South Korean students were able to reach similar diagnostic accuracy as experts for diagnosing malaria-infected red blood cells after only 3 hours of training, allowing the authors to build a gold standard library of malaria-infection labels for erythrocytes. The authors used a game-based tool that made the task easy to complete by including points and a leaderboard on the platform. This method of so-called gamification is frequently used by crowdsourcing platforms and has been shown to increase the engagement of the crowd and improve the quality of the crowdsourced work [37]. Bittel et al [38] used a hybrid crowd-ML approach to create the largest publicly available data set of annotated endoscopic images. Heim et al [28] found that a crowd was able to segment abdominal organs in computed tomography (CT) images with comparable quality to a radiologist, but at a rate up to 350 times faster.

There are few studies published to date evaluating crowdsourcing in the field of dermatology, and to the best of the authors’ knowledge, there are no published studies on the utility of crowdsourcing for the annotation of features present in dermoscopic images [39,40].

The aim of this study is to demonstrate that crowdsourcing can be employed to label dermoscopic subfeatures of melanocytic lesions with equivalent reliability to a small group of dermatologists. This will allow for efficient annotation of a large repository of dermoscopic images to aid the development of novel ML algorithms [32]. Incorporating detection of dermoscopic features into diagnostic algorithms will result in explainable outputs and may therefore improve the acceptability of these outputs to the medical community.

Table 1. List of superfeatures (in bold) and corresponding subfeatures seen in melanocytic lesions [9].

<table>
<thead>
<tr>
<th>Dots</th>
<th>Irregular, regular</th>
</tr>
</thead>
<tbody>
<tr>
<td>Globules</td>
<td>Cobblestone pattern, irregular, regular, rim of brown globules</td>
</tr>
<tr>
<td>Lines</td>
<td>Branched streaks, pseudopods, radial streaming, starburst</td>
</tr>
<tr>
<td>Network structures</td>
<td>Atypical pigment network, broadened pigment network, delicate pigment network, negative pigment network, typical pigment network</td>
</tr>
<tr>
<td>Regression structures</td>
<td>Peppering/granularity, scarlike depigmentation</td>
</tr>
<tr>
<td>Shiny white structures</td>
<td></td>
</tr>
<tr>
<td>Patterns</td>
<td>Angulated lines, polygons, zigzags</td>
</tr>
<tr>
<td>Structureless areas</td>
<td>Irregular blotches, regular blotches, blue-whitish veil, milky red areas, structureless brown areas, and homogenous (not otherwise specified)</td>
</tr>
<tr>
<td>Vessels</td>
<td>Comma, corkscrew, dotted vessel, linear irregular vessel, polymorphous vessel, milky red globules</td>
</tr>
</tbody>
</table>

Methods

Ethics Approval

This study was conducted as part of the umbrella ISIC research protocol and was approved by the Memorial Sloan Kettering Cancer Center Institutional Review Board (16-974). All images were deidentified and do not contain any protected health information as per the terms of use agreement for the ISIC archive.

Materials

This study was performed in 3 separate experiments, each using the same set of 248 lesion images used in the EASY study. Briefly summarized, clinical experts contributed 964 lesion images showing 1 of 31 preselected subfeatures, as described
by Kittler et al. Clinicians were asked to submit images of “excellent quality showing the exemplar feature in focus.” Three experts chose 248 of these images, roughly balancing benign and malignant lesions and ensuring image quality. Each of the 31 features was the exemplar in 8 of the lesion images submitted. However, each image could, and typically did, show multiple features.

Subfeatures and Superfeatures

As described earlier, low to moderate IRR was observed for the majority of subfeatures. Hence, we used only the superfeature terms for our scalability investigation. While each of the subfeatures had 8 exemplar images, collapsing the labels into superfeatures created some imbalance. The full list of subfeatures is shown in Textbox 1. The 9 superfeatures (dots, globules, lines, network structures, patterns, regression structures, shiny white structures, structureless areas, and vessels; shown in Multimedia Appendix 1, Table S1) were presented to participants during the tutorial on the DiagnosUs smartphone app, adapted from Marghoob and Braun.

Agreement Measure

To measure agreement across raters, we employed the Cohen $\kappa$ [42], which has a value of 0 for completely random choices, increasing toward a maximum value of 1.0 with improved IRR. Measures of agreement are interpreted as poor (0-0.4), fair to good (0.4-0.75), and excellent (0.75-1.0) [43]. This measure was primarily chosen to accommodate the nature of the 3 separate studies (see below), allowing for partial data between pairs of raters using the binary choice of “feature present” or “feature absent.” Throughout this paper, we use the term “median $\kappa$” to refer to the median of $\kappa$ values across the set of pairwise comparisons as a measure of central tendency, given the nonnormal distribution of $\kappa$ values.

Initial Expert Annotations (Study 1)

For the first study, we used a custom programmed annotation platform built for the ISIC archive. We asked a total of 20 dermatoscopy experts to each annotate 62 images (2 per exemplar feature) in 4 substudies of nonoverlapping image sets. Experts for study 1 were clinicians with $\geq 5$ years of experience. We recruited 7 experts to use the DiagnosUs platform and annotate the same 248 images from studies 1 and 2 for the presence of the 6 superfeatures. For each of the features, we selected the first 5 dermatologists who completed annotation of the image set.

Gold Standard for the Crowd Study

After collapsing the subfeatures into the 9 abovementioned superfeatures, we found that 3 had very poor agreement and too few exemplars to allow reliable evaluation by the crowd: patterns, shiny white structures, and structureless areas. For the remaining 6 superfeatures (dots, globules, lines, network structures, regression structures, and vessels), images in which at least 3 of 5 experts in study 1 had selected $\geq 1$ of the subfeatures within the same superfeature as present were used as the gold standard for “superfeature present.” Images in which none of the 5 experts had identified any of the subfeatures within the same superfeature as present were used as the gold standard for “superfeature absent.”

Nonexpert Crowd Annotations (Study 2)

To collect nonexpert image annotations, we used the commercially available platform DiagnosUs (Centaur Labs) [44] through a collaboration agreement. Users can sign up to the app and participate in competitions, which increases engagement and improves accuracy [37]. Users are recruited via a referral system or advertisements on social media. To ensure that only users somewhat skilled at a task computed average detection values, gold standard images were used for both training and validation. This left the remaining images, for which either 1 or 2 expert raters annotated a subfeature within the same superfeature as being present, as true test images. If a user did not reach at least 83% correctness for the validation items, that user’s choices were not used in the subsequent analysis. Each of the 6 superfeatures was presented as a separate task. In addition to the binary choice of presence or absence of a superfeature, we also collected reaction times to assess decision difficulty [45].

Expert Crowd Annotations (Study 3)

As study 1 allowed experts to select from the 31 subfeatures, we replicated the methodology of study 2 to allow direct comparison with the nonexpert crowd. Experts in study 2 were dermatologists with $\geq 5$ years of experience. We recruited 7 experts to use the DiagnosUs platform and annotate the same 248 images from studies 1 and 2 for the presence of the 6 superfeatures. For each of the features, we selected the first 5 dermatologists who completed annotation of the image set.

Reaction Times

For each of the tasks in studies 2 and 3, we computed the per-item averaged logged reaction times as the log of (1 + reaction time) to approximate a normal distribution of measurement errors. These averaged logged reaction times were then regressed against the average responses and a quadratic term, allowing for an inverted-U–shaped response function, which peaked roughly at the (across-readers) point of indecision.

Results

Initial Expert Annotations (Study 1)

In study 1, we found that dots showed poor agreement (median $\kappa=0.298$), whereas vessels showed excellent agreement (median $\kappa=0.768$). All other superfeatures showed fair to good agreement (Table 1). The resulting distributions of pairwise Cohen $\kappa$ values are shown in Figure 1A. The number of resulting gold-standard images for each of the 6 superfeatures was as follows (0 readers/at least 3 readers, respectively): dots (93/61), globules (57/92), lines (129/60), network structures (63/140), regression structures (113/59), and vessels (152/66).
Table 1. Median Cohen κ values for pairwise readers. For study 2, pairs of readers were considered only if both readers saw at least 62 of the same images.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Study 1 (experts), median κ</th>
<th>Study 2 (nonexpert crowd), median κ</th>
<th>Study 3 (expert crowd), median κ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dots</td>
<td>0.2977</td>
<td>0.5264</td>
<td>0.4829</td>
</tr>
<tr>
<td>Globules</td>
<td>0.4075</td>
<td>0.3945</td>
<td>0.5166</td>
</tr>
<tr>
<td>Lines</td>
<td>0.5205</td>
<td>0.3983</td>
<td>0.4433</td>
</tr>
<tr>
<td>Network structures</td>
<td>0.6175</td>
<td>0.5810</td>
<td>0.7575</td>
</tr>
<tr>
<td>Regression structures</td>
<td>0.4643</td>
<td>0.5066</td>
<td>0.4730</td>
</tr>
<tr>
<td>Vessels</td>
<td>0.7683</td>
<td>0.7977</td>
<td>0.7903</td>
</tr>
</tbody>
</table>

Figure 1. Pair-wise Cohen κ values for study 1 (A), study 2 (B), and study 3 (C).

Nonexpert Crowd Annotations (Study 2)

Providing demographic data pertaining to the users’ jobs and their reasons for using the DiagnosUs platform was optional; these data were collected from 190 users. Of these, 23 (12.1%) were physicians (2 dermatologists, 21 other specialties), 72 (37.9%) were medical students, 11 (5.8%) were nurse practitioners, 8 (4.2%) were physician assistants, and 76 (40%) were “other” or “other healthcare student.” The most common reason for using DiagnosUs was “improve my skills” (134/190, 70.5%), followed by “earn money” (37/190, 19.5%) and “compete with others” (19/190, 10%).

The number of users that engaged with each of the features varied for dots (92 users), globules (111 users), lines (82 users), network structures (97 users), regression structures (79 users), and vessels (95 users). Equally, the median number of ratings made per user per task varied for dots (160 images rated per user), globules (131 images), lines (177 images), network structures (91 images), regression structures (124 images), and vessels (104 images). The total number of crowd base ratings obtained in this study was 139,731, including 25,466 total ratings for dots, 40,853 for globules, 21,074 for lines, 17,114 for network structures, 17,020 for regression structures, and 18,204 for vessels.

The pattern we found in study 1 was largely replicated by the nonexperts. To ensure that there was sufficient and comparable overlap for images between pairs of readers, only pairs in which both readers saw at least 62 of the same images were evaluated. Dots and globules showed relatively lower agreement (with median κ values of 0.526 and 0.395, respectively), whereas network structures and vessels showed the highest agreement (with median κ values of 0.581 and 0.798, respectively). To allow a direct comparison between studies 1 and 2, we have compiled the 6 superfeatures into a panel figure (Figure 1A and 1B).

Expert Crowd Annotations (Study 3)

Again, the patterns found in studies 1 and 2 were replicated, such that dots and globules showed relatively lower agreement (median κ values were 0.483 and 0.517, respectively), whereas network structures and vessels showed the highest agreement (median κ values were 0.758 and 0.790, respectively; Figure 1C).

We computed κ values for each nonexpert reader from study 2 compared to a single simulated expert by thresholding the responses in study 3 from 3 of 5 experts into a binary variable. The median κ values between nonexperts and the thresholded average expert reader were as follows: for dots, 0.709; for
globules, 0.719; for lines, 0.714; for network structures, 0.838; for regression structures, 0.818; and for vessels, 0.728.

**Reaction Times**

Irrespective of task, the reaction time varied by user (the median IQR for reaction time across users was 2.5 seconds to 4.3 seconds) and across images (the median IQR for the difference in reaction time per user was –0.93 seconds to +1.5 seconds), suggesting that the variability within users was somewhat greater than the variability across users.

For both the nonexperts and experts, the quadratic term accounting for the inverted-U–shaped response in averaged logged reaction times reached statistical significance across all tasks. Among the nonexperts, the $t$ values (calculated with a 2-tailed $t$ test) ranged from $t_{244}=-14.3$ (for dots) to $t_{244}=-20.09$ (for vessels). Among the experts, probably due to higher noise, the $t$ values ranged from $t_{244}=-7.63$ (for regression structures) to $t_{244}=-10.62$ (for vessels). All $t$ values were highly significant ($P<.001$). In all tasks and for both sets of readers, the linear term had a negative sign and was also significant (at lower levels), meaning that in all cases readers were faster to respond when a feature was present compared to when it was absent (Figure 2).

**Discussion**

**Principal Findings**

The main findings of this study confirmed the variable, and sometimes low, IRR between experts for identifying dermoscopic superfeatures on images of melanocytic lesions. The patterns of repeatability were mirrored in all 3 studies, highlighting that some features are more challenging to identify regardless of experience level. We found that the IRR between the untrained crowd and expert crowd was good to excellent for all superfeatures, suggesting that crowdsourced labels can be reliably used for future research. Reaction times were slower for lesions that would be considered more challenging in both cohorts, and therefore may be used as a proxy for decision difficulty.

**Initial Expert Annotations (Study 1)**

In Study 1, the lowest level of agreement was observed for dots and globules, and the highest agreement was observed for network structures and vessels. This is in keeping with the findings of previous studies evaluating IRR for the identification of dermoscopic patterns among a group of experienced dermatosopists [10,46]. It has been suggested that poor agreement on criteria such as structureless areas, streaks, and dots or globules may be the result of lack of standardization in dermoscopy education [46,47]. Furthermore, the definition of dermoscopic structures may evolve over time. Whereas vascular structures and pigment network structures are easily recognizable, and their definitions have been consistent in the literature to date, dots and globules may be more difficult to categorize. Tiny, numerous gray dots may be categorized as regression structures, and red dots may be defined as vascular structures [48-50]. Globules are defined as measuring >0.1 mm, which may be challenging to identify in dermoscopic images without a unit of measurement as a reference point. Going forward, it may be more feasible to consider dots and globules as a single criterion to eliminate the
challenges encountered when attempting to differentiate them based on size.

**Nonexpert Crowd Annotations (Study 2)**

A similar pattern of results was seen in study 2, suggesting that the gridlike pattern of a pigment network structure and the distinctive red color of vascular structures may be more repeatably identified by an untrained crowd. In keeping with the results of study 1, dots and globules were identified with poor repeatability. Again, this may be as a result of the ambiguity in distinguishing between the two on the basis of their diameter.

Prior studies have shown that dermoscopy by novice clinicians is no more accurate than visual inspection alone, and so an untrained crowd would not be expected to identify complex dermoscopic patterns, particularly when agreement between a group of world experts is known to be low, such as in our EASY study. To obtain reliable crowdsourced labels for complex medical images, an easier set of images may be used or participants may receive extended training; the study must also be designed to accommodate a large number of redundant labels [28]. In a study evaluating crowdsourcing as a method of identifying colonic polyps in CT colonoscopy images, McKenna et al [51] found that the crowd performance deteriorated with increasing difficulty, as well as with increasing reaction time. By collapsing the 31 subfeatures into 6 superfeatures, we created a more achievable task for a crowd with no prior experience of dermoscopy.

**Expert Crowd Annotations (Study 3)**

The results from study 3 showed that agreement between experts was higher for dots, globules, and network structures when compared to study 1, in which annotations for subfeatures were aggregated into superfeature categories. It is known that there is a greater potential for disagreement with an increased number of categories and that the Cohen \( \kappa \) is typically observed to be lower in this circumstance [52]. Thus, if experts have been asked to choose from 6 superfeatures rather than 31 subfeatures, there would have been less potential for disagreement.

When comparing the median \( \kappa \) across all 3 studies, we found that repeatability for identifying all 6 superfeatures was similar across the experts and nonexperts. When comparing the median nonexpert annotations in study 2 to the thresholded expert annotations in study 3 for the same task, we saw that agreement was excellent for network structures and regression structures and good for the 4 remaining superfeatures. This suggests that the crowd was able to both repeatably and reliably identify dermoscopic superfeatures. Interestingly, agreement for vessels was higher within groups than between groups; thus, crowd annotations, although repeatable, were less accurate than expert annotations, suggesting that the crowd may be less reliable when annotating vessels. Vessels had the highest number of subfeatures (6) with distinct morphologies, several of which were not presented to the crowd during training on the DiagnosUs platform. Redesigning the tutorial may result in better accuracy for crowd annotations of vessels.

**Reaction Times**

For both experts and nonexperts, there were 2 common patterns of response time (ie, the time it took a participant to feel confident enough to log a response varied as a function of estimated difficulty). For images for which the crowd showed low agreement (the average response was approximately 0.5 seconds), the response times were significantly slower than for images for which the crowd showed high agreement. For gold standard images (those for which \( \geq 3 \) of 5 experts in study 1 agreed on the presence or absence of a feature) reaction times were faster than those for images of lesions upon which only 1 or 2 experts agreed, highlighting the challenging nature of these images. Furthermore, images where the feature was present had faster reaction times than those where the feature was absent, regardless of level of agreement. Overall, experts took longer to respond to images than nonexperts, suggesting that they exerted more effort to ensure a correct response. In addition, there was no financial reward for experts in this study; thus, they were less motivated to annotate as many lesions as possible within a designated timeframe.

**Limitations**

One of the fundamental limitations of this study and future implications that can be drawn from it is the potentially low dependability of crowdsourced annotations. Although we found high repeatability and reliability of labels in study 3, this was for a relatively small set of images that had been carefully curated to have high-quality examples of a limited number of superfeatures.

There are a number of proposed methods to improve the quality of crowdsourced data. Crowd performance has been shown to improve with increased time spent training for the task, and participants that complete more readings have been observed to perform better [36,53]. Therefore, we may be able to improve performance of the crowd by providing additional training, as well as by increasing participant engagement, such as with greater financial rewards. This may, however, come at the cost of increased time and a smaller number of participants. Although crowdsourced annotations may be marginally less accurate than those provided by experts, the increased number of available labels for training ML algorithms has been shown to make them more robust to noisy data [54].

In this study, we validated the participants’ performance against gold standard images to ensure the quality of labels, and poorly performing participants were not included. In the absence of an expert-labeled image, DiagnosUs allows a ground truth to emerge with an unlabeled competition design in which images that show internal consistency across raters become the initial gold standard. Filtering of individuals may also be achieved by evaluating participants based on previously performed tasks or providing a pretask test [34]. Aggregating results via majority voting is another commonly used method of preprocessing to improve annotation quality. Annotations may also be evaluated by using them to train a ML model and using the model’s performance as a proxy for crowd performance [34].

It is essential that some level of quality assurance take place for crowdsourced annotations in the absence of expert labeling for
comparison, as would be the case in future studies. Although agreement is traditionally considered an indicator of data reliability, it has been suggested that participants’ competence and confidence should be taken into account [55]. This can be achieved by filtering participants with poor accuracy on gold standard images, aggregating annotations, and using reaction time as a proxy for decision confidence. Images that give rise to long reaction times and a low level of agreement may then be transferred to an expert for annotation.

Many of the lesions in the archive are complex and have multiple dermoscopic patterns, which we observed created challenges for the experts to reliably identify, let alone the untrained crowd. Obtaining annotations for only 6 superfeatures may limit the diagnostic value of an ML tool. Crowdsourced labeling of the ISIC archive may be limited by its size; at the time this study was conducted, approximately 10,000 superfeature annotations were collected per day. However, engagement with the DiagnosUs platform continues to grow exponentially, and it currently receives in excess of 1 million crowd opinions daily across multiple tasks. Therefore, it may be entirely achievable to annotate the ISIC archive with crowdsourced labels within a timeframe of weeks to months.

Although the images in this study were subject to a manual quality assurance process, they were not standardized. For example, some images contained a unit of measure, which may have introduced bias when differentiating between dots and globules, as mentioned earlier in the discussion. Insufficient demographic data were collected by the DiagnosUs platform to allow meaningful subanalyses; however, disparities in experience level between users were highlighted. Importantly, 2 physicians specializing in dermatology participated in the crowd, and it therefore cannot be truly considered untrained. Due to the nature of the platform, it appeals to medical professionals as a learning tool with the aim of driving innovation in medical artificial intelligence, and the platform provides meaningful labels at scale regardless of the background of its users.

Future Work
Given the sheer size of the ISIC archive, it would be infeasible to obtain annotations by expert dermatologists for all images. We have shown the feasibility of obtaining crowdsourced annotations; this method can be used in several ways. First, it will allow hierarchical organization of the archive, allowing users to filter lesions based on dermoscopic patterns. Second, it may act as a teaching tool, allowing novice dermatologists to learn patterns and corresponding diagnoses. And third, these annotated data may be used to develop novel ML tools. Even if only a small proportion of images are labeled by the crowd, a pattern classification or segmentation algorithm could be used to annotate additional images in the archive though a weakly supervised technique [56]. A hybrid crowd-algorithm approach has been successfully developed by several groups for the purpose of segmenting large databases of medical images [28,38,54,57].

The issues regarding “black box” algorithms have been raised as a barrier to implementation of these tools in clinical practice. Given the complexity of medical imaging data, a fully explainable algorithm would be unlikely to have adequate performance; however, use of interpretable outputs may go some way to assuage hesitancy in uptake. A classification tool that is also able to detect dermoscopic patterns that have influenced its decision would allow dermatologists to make more informed decisions when evaluating the output of the algorithm [22]. Furthermore, a multidimensional algorithm that is trained on both diagnoses and dermoscopic features may have increased accuracy when compared to those trained on diagnoses alone.

The next steps in exploring the applications of crowdsourced data are to expand labeling to a larger sample of images with a robust quality assurance process and incorporate the labels into a pattern-detection algorithm to be evaluated in a study of readers. Should this algorithm display acceptable performance measures, it may be deployed to label further images and be incorporated into a classification algorithm to improve its explainability.

Conflicts of Interest
JK has provided services for Skin Analytics, Ltd and IQVIA, Inc. ACH has provided services for Canfield Scientific, Lloyd Charitable Trust, and SciBase; has ownership and equity interests in HCW Health, LLC; and has a fiduciary role, intellectual property rights, and ownership and equity interests in SKIP Derm, LLC. VR has provided services for Inhabit Brands, Ltd. AAM has received royalties from UpToDate. KN is an employee of Centaur Labs. KP is an employee of Centaur Labs. ED is the CEO of Centaur Labs.

Multimedia Appendix 1
Select dermoscopic features as presented to participants during the tutorial on the DiagnosUs smartphone app.

References


19. Kentley et al. JMIR MEDICAL INFORMATICS


Abbreviations

CT: computed tomography
EASY: Expert Agreement on the Presence and Spatial Location of Melanocytic Features in Dermoscopy
IRR: interrater reliability
ISIC: International Skin Imaging Collaboration
ML: machine learning

©Jonathan Kentley, Jochen Weber, Konstantinos Liopyris, Ralph P Braun, Ashfaq A Marghoob, Elizabeth A Quigley, Kelly Nelson, Kira Prentice, Erik Duhaime, Allan C Halpern, Veronica Rotemberg. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 18.01.2023. 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.
Toward Individualized Prediction of Binge-Eating Episodes Based on Ecological Momentary Assessment Data: Item Development and Pilot Study in Patients With Bulimia Nervosa and Binge-Eating Disorder

Ann-Kathrin Arend1; Tim Kaiser2; Björn Pannicke1; Julia Reichenberger1; Silke Naab3; Ulrich Voderholzer3,4,5; Jens Blechert1

1Department of Psychology, Centre for Cognitive Neuroscience, University of Salzburg, Salzburg, Austria
2Department of Clinical Psychology, University of Greifswald, Greifswald, Germany
3Schoen Clinic Roseneck, Prien am Chiemsee, Germany
4Department of Psychiatry and Psychotherapy, University Hospital, Ludwig Maximilian University of Munich, Munich, Germany
5Department of Psychiatry and Psychotherapy, University Hospital of Freiburg, Freiburg, Germany

Corresponding Author:
Ann-Kathrin Arend
Department of Psychology
Centre for Cognitive Neuroscience
University of Salzburg
Hellbrunnerstraße 34
Salzburg, 5020
Austria
Phone: 43 66280445102
Fax: 43 66280445126
Email: ann-kathrin.arend@plus.ac.at

Abstract

Background: Prevention of binge eating through just-in-time mobile interventions requires the prediction of respective high-risk times, for example, through preceding affective states or associated contexts. However, these factors and states are highly idiosyncratic; thus, prediction models based on averages across individuals often fail.

Objective: We developed an idiographic, within-individual binge-eating prediction approach based on ecological momentary assessment (EMA) data.

Methods: We first derived a novel EMA-item set that covers a broad set of potential idiosyncratic binge-eating antecedents from literature and an eating disorder focus group (n=11). The final EMA-item set (6 prompts per day for 14 days) was assessed in female patients with bulimia nervosa or binge-eating disorder. We used a correlation-based machine learning approach (Best Items Scale that is Cross-validated, Unit-weighted, Informative, and Transparent) to select parsimonious, idiographic item subsets and predict binge-eating occurrence from EMA data (32 items assessing antecedent contextual and affective states and 12 time-derived predictors).

Results: On average 67.3 (SD 13.4; range 43-84) EMA observations were analyzed within participants (n=13). The derived item subsets predicted binge-eating episodes with high accuracy on average (mean area under the curve 0.80, SD 0.15; mean 95% CI 0.63-0.95; mean specificity 0.87, SD 0.08; mean sensitivity 0.79, SD 0.19; mean maximum reliability of rD 0.40, SD 0.13; and mean rCV 0.13, SD 0.31). Across patients, highly heterogeneous predictor sets of varying sizes (mean 7.31, SD 1.49; range 5-9 predictors) were chosen for the respective best prediction models.

Conclusions: Predicting binge-eating episodes from psychological and contextual states seems feasible and accurate, but the predictor sets are highly idiosyncratic. This has practical implications for mobile health and just-in-time adaptive interventions. Furthermore, current theories around binge eating need to account for this high between-person variability and broaden the scope of potential antecedent factors. Ultimately, a radical shift from purely nomothetic models to idiographic prediction models and theories is required.
Recently, the methodologies of just-in-time adaptive interventions for high-risk states. As such states can fluctuate quickly, they need to be assessed as soon as possible, before the binge-eating pressure builds up. Antecedents are promising as they attempt to stop the process to resist, interventions that target binge eating based on its impulsive, automatic, and difficult-to-resist nature. However, although particular efforts have been directed at theories claiming that emotional eating underlies binge eating [5,6], nomothetic binge-eating models often fail to translate to idiographic–cognitive behavioral therapy for eating disorders (EDs) is effective for only about 65% of individuals with an ED [2] and has high relapse rates (26.8% across EDs) [3].

Nomothetic Binge-Eating Models

To predict binge eating, researchers typically rely on nomothetic theories—theories that are based on the average characteristics of multiple individuals in groups. Some nomothetic findings hold that individuals with BN and BED overeat in response to negative emotions, whereas healthy controls do not [4]. However, although particular efforts have been directed at predicting high-risk states for binge eating based on a variety of measures (eg, negative emotions or irregular eating patterns) [5,6], nomothetic binge-eating models often fail to translate to an idiographic–individual–level [7-10]. To illustrate, nomothetic theories claiming that emotional eating underlies binge eating [11] imply that emotion regulation interventions provide causative help [12,13]. However, this reasoning might not be applicable to patients who are prone to binge eating when impulsive, after extensive fasting periods, or experiencing dissociative states [6,13,14]. Correspondingly, various nomothetic theories of binge eating have proliferated. They differ substantially in the assumed causal mechanisms, which include, but are not limited to, emotional eating, impulsivity, restrained eating, food addiction, ego depletion, associative learning, and emotion-regulation or coping with emotions [4,15-21].

Idiographic Binge-Eating Models and Interventions

As binge eating can be highly impulsive, automatic, and difficult to resist, interventions that target binge eating based on its antecedents are promising as they attempt to stop the process as soon as possible, before the binge-eating pressure builds up. As such states can fluctuate quickly, they need to be assessed and evaluated with a high timely resolution to inform about the appropriate timing for interventions for high-risk states. Recently, the methodologies of just-in-time adaptive interventions (JITAI)s [22] and high-frequency ecological momentary assessment (EMA) have merged into a methodological framework that can be applied to binge-eating prediction and prevention. JITAI have been shown to enhance cognitive behavioral therapy in BED and BN [23] and have been successfully implemented in other domains of eating behavior (eg, in weight loss) [24].

In the “OnTrack” weight-loss intervention, Forman et al [24] sampled emotions and stress, next to eating history and context conditions such as watching television or alcohol consumption. By investigating a wide range of antecedents for dietary lapses, they go well beyond what emotional-eating theory suggests as predictors (eg, negative emotions). Similarly, with their “Think Slim” app, Spanakis et al [25] showed that a sample of participants with normal weight and overweight can be clusters into multiple groups according to the different momentary states in which they tend to eat unhealthily. Therefore, applying a single nomothetic binge-eating theory might be insufficient to identify a broad spectrum of individually varying antecedents and would yield inaccurate predictions of binge eating in most individuals [26]. Instead, to cover all relevant antecedents for many patients, a broad set of EMA items is required. Notably, items that serve this purpose in weight disorders (eg, “OnTrack” JITAI-enhanced weight-loss intervention by Forman et al [24]) may not cover all antecedent states that arise in patients with clinical binge eating. Furthermore, despite being often disregarded within nomothetic frameworks, protective factors (eg, positive emotions or healthy coping [27,28]) have the potential to improve prediction accuracies in idiographic machine learning frameworks because of their negative associations with binge-eating likelihood. However, to balance participant burden with broad sampling, a baseline phase with the full item set could be followed by a phase with a reduced EMA-item set, based on a prediction model that identifies the idiographic subset of items that best predict binge eating for a given individual.

Aims and Hypothesis

This study examined the feasibility of the first part of this approach, that is, whether subsets of items could be found with good prediction accuracies for binge eating. Furthermore, 2 studies were conducted to establish a conceptual and empirical foundation for JITAI on binge eating. In study 1, we collected a comprehensive set of binge-eating antecedents in the form of EMA items. We combined a literature review with qualitative and quantitative interviews (focus group with 11 inpatients) following Soyster and Fisher [29]. In study 2, an algorithm was used to select idiographic subsets of binge-eating predictors based on Elleman et al [30], Kaiser et al [10], and Soyster et al [31]. We hypothesized that these idiographic binge-eating antecedents would predict binge eating with high...
accuracy. This selection and prediction were tested in 13 patients with BN or BED.

**Methods**

**Ethics Approval**

All participants signed an informed consent form (stating which data were stored, where and for how long, who the investigator was, and the purpose of the study) approved by the ethics committee of the University of Salzburg (EK-GZ: 37/2018).

**Study 1—Development of the EMA-Item Set**

**Literature Research**

A PhD-level researcher systematically searched Google Scholar, PsycINFO, and PubMed databases for articles with the word “binge” in their title and the terms “ecological momentary” or “experience sampling” to find risk state descriptors with relevance to binge eating in the literature. The search resulted in 509 articles that were deduplicated and scanned for relevance. Only empirical articles reporting the results of EMA studies on binge eating were retained. A total of 262 articles were subsequently analyzed (see Multimedia Appendix 1, Figure S1 for an attrition diagram).

**Text Analysis Using Word Embedding**

Abstracts of all articles in the literature were retrieved. The R package text2vec [32] was used to perform global vector word-embedding analysis on these abstracts. Word embedding is an “unsupervised” learning algorithm that maps words to a vector space based on their similarity. It is unsupervised as no labeling of training data is needed because training is performed on aggregated global word-word cooccurrence statistics from a corpus. A matrix is calculated where each element $X_{ij}$ represents how often word $i$ appears in the context of word $j$ (ie, in the same sentence). Thus, words can be represented numerically and their similarities can be compared [32].

The following parameters were set for training the word vectors (vector dimensions=100, window size=15, and minimum word count to be included in the model=5). The English stop words were removed. Single words (eg, “sadness”), as well as combinations of 2 words (eg, “negative affect”), were allowed in the model. The cosine similarity between word vectors was used to quantify the similarity between word embeddings. This metric computes the angle between 2 vectors to quantify the similarity in the vector space they inhabit. The interpretation of cosine similarity resembles that of the correlation coefficients. Perfectly similar word vectors have a cosine similarity of 1, whereas perfectly dissimilar vectors have a similarity of –1. We calculated the cosine similarity of all retained words with the words “binge” or “binges” retaining only words that had at least a cosine similarity of +.10 or −.10 (resembling a small effect according to the criteria of Cohen [33] for the interpretation of correlation coefficients). In this way, we intended to find words that were conceptually similar to “binge eating” while covering a wide range of binge-eating antecedents.

**Integration Into a Preliminary Item List**

In the next step, 2 authors independently rated whether a given retained word was quantifiable with a psychometric item (ie, the words “dissociation” or “dissociative” were rated as quantifiable with the item “I feel detached from myself.” [0=not at all to very much=100]) and in terms of usefulness for an EMA survey. Items were only retained if they were rated as quantifiable and useful by both authors. Overlapping constructs were organized into categories to reduce redundancy. Finally, a preliminary list of 47 items was compiled from the empirical and theoretical constructs and complemented by constructs derived from previous EMA studies (Multimedia Appendix 2, Table S1).

**Patient Focus Group**

A focus group of inpatients (11 female adolescents and young adults in treatment for regular binge-eating episodes at the Schoen Clinic Roseneck, Germany) complemented this literature-based approach. It was conducted to tap into antecedents that nomothetic EMA research might have overlooked so far. After an individual written brainstorming session on “triggers and circumstances associated with binge eating,” the inpatients rated the preliminary list of EMA items on relevance to their binge-eating episodes (“happens before/during/after binge eating:...”; 1=[almost] never, 3=might or might not, 5=[almost] always). A moderated discussion of the brainstormed and provided items concluded the sessions.

Next, 2 researchers analyzed the rating data and integrated patient-generated items. This led to the following changes: several constructs missing in the preliminary item list were identified and items were added to cover these gaps (eg, eating based on internal opposed to external motivation: “Did you eat on your own accord?”; (not) following a regular meal structure: “How much did you follow a regular meal structure today?”; and restricting specific foods: “Are you restricting on certain foods right now?”).

The focus group participants further rated 27 of the provided items as positively associated with their binge-eating episodes (mean >3.5), 11 items as negatively associated (mean <2.5), and 9 items as unrelated to their binge-eating episodes (mean 2.5-3.5; Multimedia Appendix 2, Table S1). Some items were scored as unrelated (eg, “Right now I feel: tired” and “I engaged in increased levels of sport.”), and items with large SDs (SD >1.00; eg, “Right now I feel: relieved;” “Right now I am shopping for groceries.” and “I acted upon my plans regarding my eating behavior.”) were disregarded, merged (eg, “I am in company.” with “I am on my own.”), or exchanged (eg, “I feel strained due to...work / university / school; close social network; wider social network; everyday stressors” with “Do you feel like you can handle all upcoming tasks and problems?”). As the patients expressed concerns over the redundancy of emotional states, 4 more items were disregarded (“Right now I feel: calm/ashamed/guilty/frustrated”). Finally, 4 items regarding eating behaviors such as “resistance to food craving” or “restriction” were rephrased to map more accurately on constructs introduced by the focus group (see Multimedia Appendix 3, Figure S1 for all item iterations).
Feedback of Clinicians

Finally, clinicians with experience in ED treatment (n=4) provided feedback on the gaps in the included constructs. This feedback was integrated by adding concepts such as accessibility to tasty food, day structure (ie, regular sleep and eating patterns), self-regulation intentions, and eating alone. This feedback further led us to include the autoregressive effect of binge-eating episodes on subsequent binge-eating risk in our models [34].

First Pilot

The EMA items were then piloted by 2 authors and 1 female patient with BN (consistent with the Diagnostic Statistical Manual-5 [DSM-5] [1]) to evaluate content, coverage, wording, and participant burden. Piloting revealed that some items needed further changes to map more accurately on the intended constructs: 1 item about adaptive coping strategies was added (“How much did you try to distract yourself from a possible urge to overeat by healthy strategies [eg, relaxation, social activity, mindfulness, etc]?”) to complement the items on dysfunctional coping and distraction strategies, which were merged into one item (“How much did you try to distract yourself from a possible urge to overeat by unhealthy strategies [eg, alcohol, cigarettes, drugs, self-harm, etc]?”). Two items were rephrased, and 1 item assessing food craving was split up and rephrased to differentiate food craving, overeating, and objective binge-eating episodes (food craving: “How strong is your craving for certain foods right now?”; overeating: “How strong is your urge to overeat right now?”; and binge-eating episodes: “How high would you rate your risk for a binge-eating episode right now?”).

The highly compliant participant with BN (all 84 EMA signals answered) reported that the participant burden was too high. Thus, 6 more items were disregarded to shorten the extensive list of items assessing different forms of self-licensing [35,36] and restrictions. Finally, the authors integrated the information gathered in the previous steps (ie, literature review, feedback of the focus group, feedback from clinicians, and feedback of the pilot patient) to make final iterations to the EMA-item set (see Multimedia Appendix 3, Figure S1 for all item iterations, and Multimedia Appendix 4, Tables S1 and S2 for the final EMA-item set).

Study 2—Idiographic Predictor Selection and Prediction of Binge Eating From EMA Data

Participants

Female patients with current BN (n=12) or BED (n=1) were recruited via mail from the waiting list for inpatient treatment of the Schoen Clinic Rosenbeck, Germany (n=10), and from web-based forums on eating disorders and psychology (n=3; see Multimedia Appendix 5, Figure S1 for a CONSORT [Consolidated Standards of Reporting Trials] flowchart). This study was advertised as a pilot study for a smartphone-based binge-eating intervention. The data were collected between April 2020 and April 2021.

Procedure

All participants completed the following study protocol. First, the BN and BED research diagnoses according to DSM-5 [1] were determined via telephone using the Eating Disorder Examination interview [37] and the Structured Clinical Interview for DSM-IV [38]. Both interviews were adapted to the diagnostic criteria of the DSM-5 (eg, 1 binge-eating episode per week for 3 months instead of 2 binges per week for 3 months).

The participants were then introduced to the EMA items and logged into the customized smartphone app SmartEater. SmartEater was used during the subsequent EMA phase, in which signal-based EMA questionnaires were inquired up to 84 times per participant (6 signal-contingent prompts per day, in intervals of 2.5 hours for 2 weeks; questionnaires expired 1 hour after the initial prompt). In addition, an event-contingent EMA questionnaire on overeating, loss of control, and binge-eating episodes was accessible. Participants were instructed to fill in this event-contingent questionnaire whenever they felt like they overate or felt a sense of loss of control over food intake or both. The event-contingent questionnaire included questions to differentiate between subjective and objective binge eating and objective overeating (Multimedia Appendix 4, Table S2). EMA items assessing emotions were presented in a randomized order. However, the other items were presented in a fixed order to prevent carryover effects. The participants were able to review and change their answers through a “back” button. Answering all items (except branched items) was mandatory for submission of the questionnaires.

After the EMA phase of 2 weeks, a JITAI phase of 2 weeks started, in which the participants received short intervention suggestions from the app to prevent binge-eating episodes at ideographically predicted high-risk times. Every study stage was accompanied by web-based questionnaires that assessed current eating behavior pathology, demographic data, perceived acceptability, feasibility, and so on. Data from the intervention phase were not covered in the present article. For reimbursement, the participants received €30 (US $32.80) and personalized feedback on their EMA data and psychometric web-based questionnaires.

Data Preparation and Measures

To avoid the violation of the assumption of equally spaced time series [39], empty rows were inserted in the data set after every last signal for a given day. This prevented the prediction algorithm from regressing on data from the previous day.

Binge-Eating Episodes—Criterion

Objective binge-eating episodes, characterized by (1) “feelings of loss of control over eating behavior” and (2) “consumption of objectively large, inappropriate amounts of food” [1,37,40], were identified from eating episodes reported over the signal-based (1 item: “Was your meal a main meal, snack, or binge?”) and event-based EMA questionnaires (2 items: “Would other people rate the amount of food as excessive under similar circumstances?” and “Did you feel like you are losing control of your eating behavior?”). The signal-based and the 2 event-based items were recoded into a binary variable indicating the occurrence of an objective binge-eating episode (binge-eating episode reported=1, no binge-eating episode reported=0). As the algorithm was supposed to predict future
binge-eating episodes, this variable was shifted backward in time by one signal (approximately 2.5 hour).

**EMA Predictors**

An unshifted version of the binge-eating variable was included as a possible predictor of the autoregressive effects of binge eating. Furthermore, additional EMA items (n=31) were used to model possible binge-eating antecedents. Thus, only items that were assessed with every signal-based questionnaire were included (aside from the binge-eating classifier), as each variable needed to have a sufficient percentage of data points within a person (see Multimedia Appendix 4, Table S1 for the wording of each item).

**Time Predictors**

Time variables, especially in the form of circles and distinct times of day, have been shown to be highly predictive in everyday life [41]. EMA studies have even found peak times for certain binge-eating antecedents (ie, food cravings or hunger; [42]) and binge eating itself [43-45]. Thus, as temporal data are passively collected in the EMA setting via timestamps, without additional participant input, we decided to include different temporal predictors that could detect a single high-risk time per day (24-hour oscillation) or several times per day (sub-24 hour oscillation).

Variables representing 8-, 12- and 24-hour sinusoidal and cosinusoidal cycles were computed based on the cumulative sum of time differences between assessments (eg, 10:30 AM-8 AM, 1 PM-10:30 AM=2.5, 5, 7.5...). For example, a 24-hour sinusoid cycle was calculated using the following formula: \( \sin(2\pi \times \Delta_t) \), where \( \Delta_t \) is the difference between assessment points in hours (here: 2.5). Finally, dummy-coded variables representing the time of day were calculated for each signal (morning, late morning, early afternoon, afternoon, evening, and late evening). This allows for identifying a daytime when binge eating is particularly likely for a given participant (eg, when returning from work) that could not be well captured by the cyclical predictors.

**Application of the Best Items Scale that is Cross-validated, Unit-weighted, Informative, and Transparent Algorithm to EMA Data (for Idiographic Predictor Selection and Prediction of Binge Eating)**

The machine learning algorithm Best Item Scales that are Cross-validated, Unit-weighted, Informative and Transparent (BISCUIT) [30] of the bestScales function from the R package psych [46] was applied separately to the EMA data of each patient to select the best idiographic predictors of binge-eating episodes. This method was chosen because of its (1) robustness to missing data; (2) use of unit-weighted scoring of predictors, which was found to be more generalizable, especially in the context of prediction; and (3) tendency to select more parsimonious predictor sets compared with other approaches such as Elastic Net regression [30,47,48]. BISCUIT is a simple algorithm that correlates a set of predictors (here all EMA and time variables) with a criterion (here, the binary time-shifted binge-eating variable at t+1) and retains the predictors with the highest correlation to form a unit-weighted scale [10,30,46].

This scale was then used to estimate the out-of-sample predictive performance using 10-fold cross-validation. The average correlation of the scale with the criterion across 10 cross-validation splits was then computed, and the set of items with the highest cross-validated correlation was retained [30,46]. The output of BISCUIT is the selection of items showing maximum predictive validity, as the cutoff values that lead to the highest combination of sensitivity and specificity are retained [30,46].

Thus, multiple Rs (pairwise Pearson correlations) of all predictors with time-shifted binge-eating episodes (at t+1) were calculated for each participant separately to select the idiographic predictor sets. Furthermore, the area under the curve (AUC) with a bootstrapped 95% CI, specificity, sensitivity, and within- and out-of-sample reliability were calculated as prediction accuracy measures of the idiographic predictor sets and their prediction of binge eating in the next 2.5 hours (t+1).

**Results**

**Study 1—EMA-Item Set**

The final signal-contingent EMA questionnaire included 36 EMA items (momentary emotions, stress, exhaustion, and context; eg, being alone, social interactions, dissociations, eating behavior, resistance to food craving, distraction, and coping), which were designed to be assessed 6 times per day. In addition, an optional event-contingent EMA questionnaire on overeating, loss-of-control eating, and binge eating was self-initialized and included 20 items. See Multimedia Appendix 4, Tables S1 and S2 for all interval- and event-contingent items and their wording. A flowchart of all iterations applied to the EMA-item set can be found in Multimedia Appendix 3, Figure S1).

**Study 2—Idiographic Predictor Selection and Prediction of Binge Eating From EMA Data (by Application of the BISCUIT Algorithm)**

**Selection of Idiographic Predictor Subsets**

The patients (n=13) answered on an average 67.3 out of 84 EMA prompts (SD 13.4; range 43-84; see Multimedia Appendix 6, Table S1 for EMA compliance and occurrences of binge-eating episodes per patient). Across participants, the algorithm selected highly heterogeneous predictor sets of varying sizes (mean 7.31, SD 1.49; range 5-9 predictors) for the prediction of binge-eating episodes.

Figure 1 shows the idiographic predictor selection that showed maximum predictive validity for each participant. Thus, the predictors (at t) with the highest multiple Rs (pairwise Pearson correlations) with time-shifted binge-eating episodes (at t+1) were selected. All listed items were selected as idiographic predictors of binge eating, independent of their significance. However, we additionally calculated the significance of the correlations for the context. The exact P values, codes, and data can be found in the corresponding project in the Open Science Framework [49]. Note that the results might vary slightly, as the R function set.seed does not apply to the cross tables.
Prediction of Binge Eating by Idiographic Predictor Subsets

The selection of idiographic predictor sets resulted in good average prediction accuracy (mean AUC 0.80, SD 0.15; mean 95% CI 0.63-0.95; mean specificity 0.87, SD 0.08; mean sensitivity 0.79, SD 0.19; mean maximum reliability of $r_D$ 0.40, SD 0.13; mean $r_{CV}$ 0.13, SD 0.31). The mean AUC of 0.80 indicates that there is on average an 80% chance that the idiographic models predict binge and non-binge episodes accurately. The mean specificity of 0.87 indicates that the idiographic models mistakenly classified 13 of 100 episodes as binge-eating episodes. The mean sensitivity of 0.79 indicates that the idiographic models mistakenly classified 21 out of 100 binge-eating episodes as non-binge episodes. Table 1 shows the prediction accuracies of the idiographic predictor subsets for binge-eating episodes per participant. R code and data are available from the Open Science Framework [49].
**Table 1. Model fit indices for prediction of binge eating in the next 2.5 hours from idiographic predictors, selected by the Best Items Scale that is Cross-validated, Unit-weighted, Informative, and Transparent (BISCUIT) algorithm, separately for each participant.**

<table>
<thead>
<tr>
<th>Participants</th>
<th>01</th>
<th>02</th>
<th>03</th>
<th>04</th>
<th>05</th>
<th>06</th>
<th>07</th>
<th>08</th>
<th>09</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model fit indices</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUC&lt;sup&gt;a&lt;/sup&gt; (95% CI)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.92 (0.75-1.00)</td>
<td>0.97 (0.92-1.00)</td>
<td>0.84 (0.70-0.98)</td>
<td>0.51 (0.23-0.80)</td>
<td>0.73 (0.45-1.00)</td>
<td>0.93 (0.83-1.00)</td>
<td>0.85 (0.75-0.95)</td>
<td>0.63 (0.41-0.85)</td>
<td>0.75 (0.58-0.93)</td>
<td>0.72 (0.53-0.92)</td>
<td>0.60 (0.29-0.90)</td>
<td>0.98 (0.94-1.00)</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>0.84</td>
<td>0.89</td>
<td>0.74</td>
<td>1.00</td>
<td>0.86</td>
<td>0.85</td>
<td>0.90</td>
<td>0.81</td>
<td>1.00</td>
<td>0.74</td>
<td>0.87</td>
<td>0.91</td>
<td>0.96</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>1.00</td>
<td>1.00</td>
<td>.86</td>
<td>.45</td>
<td>.67</td>
<td>1.00</td>
<td>1.00</td>
<td>.80</td>
<td>.56</td>
<td>.73</td>
<td>.69</td>
<td>.56</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Derivation step (within-sample performance)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( r_{DP}^{c,d} ) (SD)</td>
<td>0.48 (0.04)</td>
<td>0.42 (0.05)</td>
<td>0.53 (0.04)</td>
<td>0.33 (0.05)</td>
<td>0.34 (0.03)</td>
<td>0.47 (0.06)</td>
<td>0.53 (0.06)</td>
<td>0.51 (0.10)</td>
<td>0.10 (0.22)</td>
<td>0.46 (0.06)</td>
<td>0.50 (0.06)</td>
<td>0.18 (0.10)</td>
<td>0.32 (0.23)</td>
</tr>
<tr>
<td><strong>Validation step (out of sample performance)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( r_{CV}^{d,e} ) (SD)</td>
<td>0.41 (0.34)</td>
<td>0.10 (0.56)</td>
<td>-0.02 (0.72)</td>
<td>0.36 (0.38)</td>
<td>0.38 (0.62)</td>
<td>-0.36 (0.28)</td>
<td>0.29 (0.30)</td>
<td>0.54 (0.42)</td>
<td>-0.11 (0.46)</td>
<td>0.34 (0.27)</td>
<td>0.27 (0.64)</td>
<td>0.08 (0.78)</td>
<td>-0.56 (0.57)</td>
</tr>
</tbody>
</table>

<sup>a</sup>AUC: area under the curve.

<sup>b</sup>CI: bootstrapped 95% CI of the AUC.

<sup>c</sup>\( r_{DP} \): multiple \( R \) of the unit-weighted scale in the derivation step.

<sup>d</sup>: pairwise Pearson correlation of the item with time-shifted binge-eating episodes (at \( t+1 \)).

<sup>e</sup>: \( r_{CV} \): average cross-validated multiple \( R \) of the derived scale.

**Discussion**

**Principal Findings**

**Study 1—EMA-Item Set**

This study used a mixed methods approach to develop a conceptual and statistical basis for an idiographic JITAI for binge eating. The EMA-item development in study 1 followed a replicable procedure similar to Soyster and Fisher [29] while considering nomothetic theories on binge-eating antecedents (ie, emotional eating) and underwent several qualitative (literature research and focus group brainstorming and discussion) and quantitative (focus group ratings) iterations and piloting.

This resulted in a broad EMA-item set (Multimedia Appendix 4), including several constructs underrepresented in the nomothetic literature (eg, “I feel detached from myself”, and “specific” restrictions “did you restrict yourself [eg, by eating less, avoiding certain foods]?” [12,50,51]). This approach also helped us shorten the extensive lists of emotional states (eg, “right now I feel...calm/relieved/ashamed/guilty/frustrated.”) because within-person ratings for similar emotions were often identical, and concerns about redundancy were expressed during the moderated discussion. Furthermore, we did not only incorporate risk factors into the EMA-item set but also protective factors that could potentially decrease the likelihood of binge eating (ie, healthy coping strategies to keep oneself from binge eating or positive emotions [27,28]). The role of protective factors is often overlooked in nomothetic binge-eating theories but is crucial to idiographic binge-eating prediction and intervention models.

**Study 2—Prediction Based on Idiographic Predictor Subsets**

Regularly completing extensive EMA-item sets (such as the present one with 36 interval-contingent and 20 event-contingent items) becomes increasingly burdensome over prolonged study periods. Thus, we applied a machine learning algorithm to the EMA data of patients with BN and BED to select parsimonious idiographic subsets of EMA items. This data-driven selection optimizes the predictive power within participants and decreases potential researcher bias.

The idiographic item subsets predicted binge-eating episodes with a high average accuracy (mean AUC 0.80) across 13 patients. Notably, the sensitivity approached 100% (successful prediction of every reported binge) in several patients, without forfeiting much specificity (predicting no binge when none occurred). This is noteworthy as outcome frequency was not extremely high (mean 10.4, SD 7.4; range 2-28 binge-eating episodes; see also Multimedia Appendix 6, Table S1 “Number of binge-eating episodes and total data points per participant”).

**Secondary Findings**

Regarding the composition of the selected item sets, a high selection rate of items with high proximity to the binge-eating construct was evident (ie, hunger, food craving, urge to overeat, subjective binge-eating risk, and preceding binge-eating episodes). This suggests that some patients may accurately predict upcoming binge-eating episodes. This reveals a relatively high level of insight into the temporal evolution of the symptoms in some patients. Surprisingly, hunger and food craving were negatively correlated with binge eating in 3 patients. One could speculate that the negative correlations between hunger and binge eating in patients 06 and 11 point to disinhibition, that is,
because of the temporary abandonment of rigid diet rules after eating in the absence of hunger [52,53].

In addition to items with conceptual similarity to binge eating, emotional items were selected in 9 patients. This supports the relevance of emotional eating in binge-eating predictions [4,11,19]. However, the selected emotion sets were highly heterogeneous across the 9 patients. In fact, no single emotion item (or specific set of emotion items) was consistently selected across all patients. This speaks against a singular and generalizable emotional eating theory of binge eating. Similarly, because no other nonemotion–related predictor was consistently selected across all patients, our pilot data provide no evidence for other generalizable nomothetic theories of binge eating. Thus, several nomothetic theories are needed to explain present heterogeneity, which may in turn explain the multitude of competing nomothetic binge-eating theories. Clearly, nomothetic theories must model individual differences more explicitly to account for these findings. These findings also support the use of a broad EMA-item set that covers a large range of possible binge-eating antecedents in the context of idiographic prediction [4-6,14-21].

Interestingly, time-derived predictors were selected only in 7 patients. In 6 of these patients, discrete time predictors were chosen that were consistent with the literature on the timing of binge-eating peaks (ie, afternoon to late evening) [43-45]. Time cycles were only selected in 3 patients. This is surprising given the observation of cyclic symptoms (eg, in depression [41]). However, time-based predictors may be more powerful if EMA items with conceptual similarity to binge eating are omitted. In the case of binge eating, time cycles could represent a rising and falling urge to overeat (eg, due to prolonged restriction between meals) [41,54]. Discrete time variables could represent the time of the day where a patient usually binges (eg, due to contexts such as being alone at home every afternoon) [41]. Assessing such time-derived variables does not require user input and thus does not contribute to the participant burden. This makes them valuable for the predictions in the JITAI framework.

Limitations and Strengths

Compared with the high average within-sample performance (mean \( r_D \) 0.40), the average out-of-sample performance (mean \( r_{CV} \) 0.13) was lower, suggesting limited out-of-sample generalizability. This might be because of 10-fold cross-validation, which does not account for the serial correlation and potential nonstationarity of time-series data [55]. Future studies could resort to alternative time-series–specific techniques (ie, roll forward cross-validation and out-of-sample evaluation) that ensure that training data always precede test data. However, X-fold cross-validation has been shown to outperform these techniques [55]. Furthermore, the number of observations was limited (max 84 per participant), leading to relatively small splits in the 10-fold cross-validation. Thus, there was a high possibility of randomly drawn training sets that were unrepresentative of the data set. The results from the validation step might also vary slightly, as the R function `set.seed` does not apply to the cross tables.

Another general drawback of the BISCUIT algorithm is that nonlinear trends and interaction effects among predictors are not considered. In addition, when applied under optimal conditions (ie, big data sets and no missing data), gold standard machine learning approaches, such as random forests [56] and XGBoost [57] in combination with super learners [58], calibrate better to the data. However, for typical EMA data sets, the conditions are rarely optimal for these algorithms. Missing data and a limited number of observations are typical features of high-burden EMA sampling schemes. However, BISCUIT was created to handle these problematic properties. BISCUIT outperformed random forest and elastic net approaches in other studies with smaller idiographic data sets and more missing data (Beck et al [59]; mean 57.4, SD 16.3; range 40-109 EMA observations; present data: mean 67.3, SD 13.4; range 43-84 EMA observations).

Finally, the idiographic approach used in this study precludes mechanistic and theoretical inferences about binge eating. Generally, machine learning algorithms are silent about the underlying mechanisms; instead, they tailor models as close as possible to the given data and conditions. Thus, the present results are highly specific, for example, to the used “prediction interval” of 2.5 hours between predictors and outcome. This could be problematic as it has been shown that emotions and eating can influence each other at different time intervals [60].

Implications and Future Directions

In addition to emphasizing the importance of a broad predictor set, the results have a direct implication for the JITAI and EMA methodology: participant burden in longer EMA sampling periods precludes the use of large EMA-item sets. Thus, such EMA studies might prune their large EMA-item sets after a “calibration period” by applying the described predictor-selection approach. Therefore, the participant burden is reduced, whereas accurate idiographic binge-eating predictions are retained. Such predictions can then be used to trigger JITAI, as done by Forman et al [24,61] in a JITAI on dietary lapses.

Future studies may transfer the present work to a range of disordered and maladaptive eating behaviors (eg, purging behaviors or food restrictions) to develop low-threshold JITAI. EMA-item–based prediction should be compared with predictions generated from passive data sources (ie, smartphone sensors, use data, and wearable data) that do not inflict much user burden [10,62-64]. In the long term, acceptance, dropout rates, and effectiveness of JITAI protocols on binge eating need to be tested in microrandomized trials [65] and classic randomized controlled trials against nonadaptive, non–real-time interventions before the ultimate recommendation as the gold standard.

Finally, feeding back personal binge-eating predictors can serve as a psychoeducational intervention and raise awareness of personal risk and protective factors. Such personal binge-eating predictors can also inform conventional face-to-face psychotherapy. Patients with a clear dominance of emotion-related predictors might profit from emotion-focused interventions [66] more than patients with a dominance of...
impulsive or craving-related predictors, who might profit more from impulse control intervention [67].

Acknowledgments
This project received funding from the European Research Council under the European Union’s Horizon 2020 Research and Innovation Program (ERCStG-2014 639445 NewEat) and the Austrian Science Fund (KLI 762). Publishing open access was supported by the Paris Lodron University of Salzburg and Austrian Science Fund. The funding bodies were not involved in the study design, assessment and analysis of data, interpretation of results, or writing of the manuscript.

Authors’ Contributions
AKA, TK, BP, JR, and JB conceptualized the studies; AKA conducted the studies; AKA, TK, and BP analyzed the data; AKA, TK, and JB wrote the paper; and BP, JR, SN, and UV contributed to the interpretation of the studies and critically revised the work for important intellectual content. All authors have read and approved the final manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Attrition diagram.
[DOCX File, 70 KB - medinform_v11i1e41513_app1.docx]

Multimedia Appendix 2
First item list.
[DOCX File, 63 KB - medinform_v11i1e41513_app2.docx]

Multimedia Appendix 3
Iterations in the creation of the item set.
[DOCX File, 82 KB - medinform_v11i1e41513_app3.docx]

Multimedia Appendix 4
Final ecologic momentary assessment items.
[DOCX File, 72 KB - medinform_v11i1e41513_app4.docx]

Multimedia Appendix 5
CONSORT (Consolidated Standards of Reporting Trials) flow diagram.
[DOCX File, 83 KB - medinform_v11i1e41513_app5.docx]

Multimedia Appendix 6
Number of binge-eating episodes and ecologic momentary assessment observations per patient.
[DOCX File, 58 KB - medinform_v11i1e41513_app6.docx]

References


38. Wittchen HU, Zaudig M, Fydrich T. SKID. Strukturiertes Klinisches Interview f


Abbreviations

AUC: area under the curve
BED: binge-eating disorder
BISCUIT: Best Item Scales that are Cross-validated, Unit-weighted, Informative and Transparent
BN: bulimia nervosa
CONSORT: Consolidated Standards of Reporting Trials
DSM-5: Diagnostic Statistical Manual-5
ED: eating disorder
EMA: ecologic momentary assessment
JITAI: just-in-time adaptive intervention
Toward Individualized Prediction of Binge-Eating Episodes Based on Ecological Momentary Assessment Data: Item Development and Pilot Study in Patients With Bulimia Nervosa and Binge-Eating Disorder


JMIR Med Inform 2023;11:e41513

URL: https://medinform.jmir.org/2023/1/e41513
doi:10.2196/41513
PMID:36821359
Deployment of Real-time Natural Language Processing and Deep Learning Clinical Decision Support in the Electronic Health Record: Pipeline Implementation for an Opioid Misuse Screener in Hospitalized Adults

Majid Afshar¹, MSc, MD; Sabrina Adelaine¹, PhD; Felice Resnik¹, PhD; Marlon P Mundt¹, PhD; John Long¹, MS; Margaret Leaf¹, MS; Theodore Ampian¹, MS; Graham J Wills¹, PhD; Benjamin Schnapp¹, MS; Michael Chao¹, MS; Randy Brown¹, MD, PhD; Cara Joyce², PhD; Brihat Sharma¹, MS; Dmitriy Dligach², PhD; Elizabeth S Burnside¹, MPH, MD, MS; Jane Mahoney¹, MD; Matthew M Churpek¹, MD, PhD; Brian W Patterson¹, MPH, MD; Frank Liao¹, PhD

¹University of Wisconsin - Madison, Madison, WI, United States
²Loyola University Chicago, Chicago, IL, United States

Corresponding Author:
Majid Afshar, MSc, MD
University of Wisconsin - Madison
1685 Highland Avenue
5158 Medical Foundation Centennial Building
Madison, WI, 53705
United States
Phone: 1 3125459462
Email: majid.afshar@wisc.edu

Abstract

Background: The clinical narrative in electronic health records (EHRs) carries valuable information for predictive analytics; however, its free-text form is difficult to mine and analyze for clinical decision support (CDS). Large-scale clinical natural language processing (NLP) pipelines have focused on data warehouse applications for retrospective research efforts. There remains a paucity of evidence for implementing NLP pipelines at the bedside for health care delivery.

Objective: We aimed to detail a hospital-wide, operational pipeline to implement a real-time NLP-driven CDS tool and describe a protocol for an implementation framework with a user-centered design of the CDS tool.

Methods: The pipeline integrated a previously trained open-source convolutional neural network model for screening opioid misuse that leveraged EHR notes mapped to standardized medical vocabularies in the Unified Medical Language System. A sample of 100 adult encounters were reviewed by a physician informaticist for silent testing of the deep learning algorithm before deployment. An end user interview survey was developed to examine the user acceptability of a best practice alert (BPA) to provide the screening results with recommendations. The planned implementation also included a human-centered design with user feedback on the BPA, an implementation framework with cost-effectiveness, and a noninferiority patient outcome analysis plan.

Results: The pipeline was a reproducible workflow with a shared pseudocode for a cloud service to ingest, process, and store clinical notes as Health Level 7 messages from a major EHR vendor in an elastic cloud computing environment. Feature engineering of the notes used an open-source NLP engine, and the features were fed into the deep learning algorithm, with the results returned as a BPA in the EHR. On-site silent testing of the deep learning algorithm demonstrated a sensitivity of 93% (95% CI 66%-99%) and specificity of 92% (95% CI 84%-96%), similar to published validation studies. Before deployment, approvals were received across hospital committees for inpatient operations. Five interviews were conducted; they informed the development of an educational flyer and further modified the BPA to exclude certain patients and allow the refusal of recommendations. The longest delay in pipeline development was because of cybersecurity approvals, especially because of the exchange of protected health information between the Microsoft (Microsoft Corp) and Epic (Epic Systems Corp) cloud vendors. In silent testing, the resultant pipeline provided a BPA to the bedside within minutes of a provider entering a note in the EHR.
Conclusions: The components of the real-time NLP pipeline were detailed with open-source tools and pseudocode for other health systems to benchmark. The deployment of medical artificial intelligence systems in routine clinical care presents an important yet unfulfilled opportunity, and our protocol aimed to close the gap in the implementation of artificial intelligence–driven CDS.

Trial Registration: ClinicalTrials.gov NCT05745480; https://www.clinicaltrials.gov/ct2/show/NCT05745480

(JMIR Med Inform 2023;11:e44977) doi:10.2196/44977

KEYWORDS
clinical decision support; natural language processing; medical informatics; opioid related disorder; opioid use; electronic health record; clinical note; cloud service; artificial intelligence; AI

Introduction

Background
As of 2017, >95% of the hospitals in the United States adopted an electronic health record (EHR), and >80% are collecting electronic clinical notes [1]. Clinical decision support (CDS) and intelligent data-driven alerts are part of federal incentive programs for Meaningful Use [2,3]. With the increasing capacity of EHR data and financial incentives to improve quality care, hospitals are increasingly well equipped to leverage computational resources to improve case identification and care throughput [4].

The unstructured narrative of EHRs provides a rich source of information on patients’ conditions that may serve as CDS tools. Detailed medical information is routinely recorded in providers’ intake notes. However, this information is neither organized nor prioritized during routine care for augmented intelligence at the bedside. Moreover, clinical notes’ free-text format hinders efforts to perform analytics and leverage the large domain of data. The computational methods of natural language processing (NLP) can derive meaning from clinical notes, from which machine learning algorithms can screen for conditions such as opioid misuse.

In 2020, overdose deaths from opioid misuse soared to an all-time high, with a record 93,000 deaths nationwide during the pandemic year [5]. Substance misuse ranks second among the principal diagnoses for unplanned 7-day hospital readmission rates [6]. Screening for patients at risk for opioid use disorders is not part of the admission routine at many hospitals, and many hospitalized patients in need are never offered opioid treatment. The high prevalence rate of substance use disorders in hospitalized adults exceeds the rates in the general population or outpatient setting and reveals the magnitude of this lost opportunity [7]. We previously trained a convolutional neural network (CNN) that outperformed a rule-based approach and other machine learning methods for screening opioid misuse in hospitalized patients. The CNN substance misuse classifier had >80% sensitivity and specificity and demonstrated that clinical notes captured during hospitalization may be used to screen for opioid misuse [8].

There remains a paucity of evidence on the implementation of clinical NLP models in an interoperable and standardized CDS system for health operations and patient care [9]. The interactions among an artificial intelligence (AI) system, its users, its implementation, and the environment influence the AI intervention’s overall potential effectiveness. Few health systems have been able to accommodate the complexities of an NLP deep learning model integrated into an existing operational ecosystem and EHR [10]. Much of the literature on NLP-driven CDS has described retrospective studies [11,12] outside the clinical workflow or simulated clinical environments [13,14]. Others have used NLP for information extraction efforts aimed at quality improvement without direct integration into the clinical workflow and operations [15,16]. Few provide a real-time NLP system but do not share an implementation framework or pipeline details to ensure fidelity and reproducibility [17]. Although the field of AI-driven CDS is growing, sharing knowledge in development and operations for health care delivery is lacking in best practices for processes and technologies in application planning, development, delivery, and operations.

This Study
This protocol describes a cloud service designed to ingest, process, and store clinical notes as standardized and interoperable messages from a major EHR vendor in an elastic cloud computing environment. We subsequently demonstrate the use of multiple open-source tools, including an open-source NLP engine for processing EHR notes and feeding them into a deep learning algorithm for screening for opioid misuse. Our resultant NLP and deep learning pipeline can process clinical notes and provide decision support to the bedside within minutes of a provider entering a note into the EHR.

To our knowledge, this is the first protocol for a bedside implementation of an NLP-driven CDS tool. We expect that our protocol will serve as a guide for other health systems to leverage open-source tools across interoperable data standards and ontologies. We provide an implementation framework and a cost-effectiveness analysis of a tool developed for the automated screening of hospitalized adults for opioid misuse. We aimed to describe a hospital-wide protocol and computing architecture for implementing a real-time NLP-driven CDS tool.

Methods

Hospital Setting and Study Period
The NLP CDS tool was implemented at the University of Wisconsin (UW) Hospital across the surgical and medical hospital inpatient wards. The EHR system used at the UW Health is Epic (Epic Systems Corp). The tool was designed for hospitalized adults (aged ≥18 years) and was assessed using a
pre-post quasi-experimental study design over 30 months (24 months of usual care and 6 months for the implementation of automated screening). The study was a quality improvement initiative by the health system to provide an automated hospital-wide screening system for opioid misuse and was registered on ClinicalTrials.gov (NCT05745480).

**Preintervention Period: Usual Care With Ad Hoc Addiction Consultations**

The UW Hospital launched an Addiction Medicine inpatient consult service in 1991 to address the high prevalence of substance use disorders among hospitalized adults. A screening, brief intervention, and referral to treatment program [18] was instituted for alcohol misuse. Screening, intervention flow sheets, and consult order sets were built into EHR-driven workflows for inpatient nurses and social workers for alcohol screening using the Alcohol Use Disorders Identification Test–Concise [19], a best practice alert (BPA) for patients at risk of alcohol use disorder, and order sets for withdrawal treatment. For other drugs, a single screening item queries “marijuana or other recreational drug use,” but no formal screening process was in place specifically targeting opioid misuse. For patients at risk of an opioid use disorder, the practice was ad hoc consultations at the discretion of the primary provider.

**Postintervention Period: Computing Architecture and Real-time Implementation**

**Overview**

The technical architecture that enabled the real-time NLP CDS tool incorporated industry-leading and emerging technological capabilities. Figure 1 details the overall NLP CDS infrastructure that exported the notes from the EHR, organized them, and fed them into an NLP pipeline; input the processed text features into the opioid screener deep learning model; and delivered the resultant scores back to the bedside EHR as a BPA. The final architecture was a real-time NLP CDS tool, and the 6 components of the architecture are further detailed in the subsequent sections.

**Component 1: Transferring Clinical Notes From the EHR to Cloud Computing**

Health Level 7 (HL7) refers to the standards for transferring health care data between data sources. Cloverleaf (Infor Cloverleaf Integration Suite) was the UW’s vendor solution that served as an application programming interface (API) gateway for accessing the clinical narratives in the EHR using HL7 version 2. To initialize the data feed, a UW Health interface analyst created a new entry in the Cloverleaf vendor software detailing the desired record information, which included clinical note text and identifiers. The analyst then “activated” the data feed, which began a continuous Transmission Control Protocol message–generating process. The Transmission Control Protocol messages were communicated using the HL7 application protocol to the Azure Virtual Machine (version 2022; Microsoft Corp) host at a port designated by the data science engineering team. This port was reserved by a NET program (the “TCP listener”), which wrote the message to the cloud file system and replied to the Cloverleaf server with acknowledgment messages.
These services were “always on,” but each had a trigger operated independently and communicated through data stores. The code triggered a custom Python script to extract the text and fed it the Azure cloud environment. In 15-minute intervals, Databricks on a virtual machine to place them at a known location within EHR and used the Azure File Transfer Protocol server running as clinical notes were entered into the EHR for an individual "heroin misuse," C3266350. For generalizability, we used the "heroin misuse" from the text was assigned C0600241 as its medical prescription normalized ontologies. For instance, Systemized Nomenclature of Medicine–Clinical Terms and mapped to a concept unique identifier (CUI) using the UMLS an anatomical sites, and procedures). Each named entity was relevant clinical contexts (eg, drugs, diseases, symptoms, System (UMLS), which is a repository of groups of words with engines used in the clinical domain [21]. cTAKES provided for other use cases. cTAKES is one of the most ubiquitous NLP engines used in the clinical domain [21]. cTAKES provided named entities from the free text that were mapped from the EHR and used the Apache OpenNLP NLP toolkit [20]. This configuration contained several engines for sentence detection, tokenization, part-of-speech tagging, concept detection, and normalization to extract information from the clinical narrative in the EHR. We did not use the negation module because it was not used in the current use case; however, this can be turned on for other use cases. cTAKES is one of the most ubiquitous NLP engines used in the clinical domain [21]. cTAKES provided named entities from the free text that were mapped from the National Library of Medicine’s Unified Medical Language System (UMLS), which is a repository of groups of words with relevant clinical contexts (eg, drugs, diseases, symptoms, anatomical sites, and procedures). Each named entity was mapped to a concept unique identifier (CUI) using the UMLS Systemized Nomenclature of Medicine–Clinical Terms and medical prescription normalized ontologies. For instance, “heroin misuse” from the text was assigned C0600241 as its CUI, which was different from the CUI assigned to “history of heroin misuse,” C3266350. For generalizability, we used the default cTAKES pipeline [22]. As clinical notes were entered into the EHR for an individual patient, Cloverleaf relayed the notes via HL7 from the Epic EHR and used the Azure File Transfer Protocol server running on a virtual machine to place them at a known location within the Azure cloud environment. In 15-minute intervals, Databricks triggered a custom Python script to extract the text and fed it into the cTAKES pipeline to map and extract the CUIs. The CUIs were stored in the Azure Data Lake with appended data, including patient ID, encounter ID, and note time stamp, and were ready to be fed into any machine learning model. The code executed for the pipeline consisted of several services that operated independently and communicated through data stores. These services were “always on,” but each had a trigger condition that initiated the code execution. The pseudocode for these services is provided in Multimedia Appendix 1.

Component 4: Text Feed From the NLP Pipeline Into the Deep Learning Model

We previously published a substance misuse screening algorithm using CUIs fed into a CNN called the Substance Misuse Algorithm for Referral to Treatment Using Artificial Intelligence (SMART-AI) [8]. SMART-AI was trained on the first 24 hours of all clinical notes entered into the EHR, starting from the patient’s arrival time. This approach provided sufficient time not only for robust training but also for the addiction consult service to intervene before hospital discharge. For ease of implementation, the model was not trained on any specific note type and followed a time stamp approach for all notes filed within 24 hours of arrival at the hospital. SMART-AI is a supervised model with target labels that were derived from the manual screening data of over 50,000 patients who self-reported on the validated Drug Abuse Screening Test [23] and answered follow-up questions about opioid use. SMART-AI is publicly available to run the trained model [24], and more details about the model architecture and development can be found in the original development and validation publication [8]. The model’s development and validation followed guideline recommendations [25]. Temporal validation of the classifier (trained on data between 2017 and 2019 and tested on data from 2020) at an outside hospital demonstrated an area under the precision-recall curve of 0.87 (95% CI 0.84-0.91) for screening for opioid misuse. Similar results were derived in an external validation at a second independent health system [8]. Multiple cutoff points were examined for the optimal threshold selection for the BPA, including the point on the area under the receiver operating curve that minimized the difference between sensitivity and specificity. During validation on the full cohort of hospitalized patients, the optimal cutoff point for screening for opioid misuse was 0.05. At that cutoff, the sensitivity was 0.87 (95% CI 0.84-0.90), specificity was 0.99 (95% CI 0.99-0.99), negative predictive value was 0.99 (95% CI 0.99-0.99), and positive predictive value was 0.76 (95% CI 0.72-0.88). The number needed to evaluate was 1.4, which translates to 26 alerts per 1000 hospitalized patients [8]. This was deemed an acceptable workload for consultation requests in live production for the UW Addiction Medicine clinicians. Additional silent testing was performed at the UW Health to examine sensitivity and specificity with 95% CI in our practice setting. All notes from the first 24 hours of arrival at the UW Hospital were combined into a single document per patient encounter and converted into sequences of vector representations (eg, embeddings). The CUI embeddings defined the input layer to the SMART-AI model at the encounter level. The model provided prediction probabilities for opioid misuse and stored them in a Databricks table with the predefined cutoff point for screen positives.
Component 5: Real-time Delivery of the Prediction Results

The Nebula Cloud Platform was Epic’s Software as a service platform for integrating new technology and specifically supported clinical prediction modeling. Nebula capabilities included the deployment of machine learning models, including a library of Epic-curated models for health care and custom algorithms. Our solution leveraged the latter to facilitate triggers from Epic to call out to the Databricks environment and provided the predictions for BPAs.

In the case of SMART-AI, we designed a BPA (Figure 2) to trigger once a clinician opened a patient chart in the EHR. Epic called its Nebula component to determine whether a BPA should be generated. Nebula made an HTTP call to Databricks to request the score. The RESTful HTTP API provided the SMART-AI model score that was serviced using MLFlow. The parameters included UMLS dictionaries, model results, patient identifiers, and other attributes necessary for individual-level predictions. The score was returned to Nebula, which was used to trigger a BPA if SMART-AI met the cutoff score for opioid misuse. For screen positives, the alert recommended the clinician to consult with the UW’s Addiction Medicine consult service. The following were internal targets to meet the real-time needs of the end user at the bedside: (1) a throughput of 1000 notes per minute (<60 ms each); (2) three-nines (99.9%) availability—equivalent of <9 hours of downtime annually; and (3) an established error rate threshold.

Figure 2. In an iterative design with feedback from end users, a final BPA was implemented for bedside care. The BPA triggers upon opening a chart for a patient that meets the cutpoint predicted probability for opioid misuse from the NLP and deep learning model (SMART-AI). BPA: best practice alert; NLP: natural language processing; SMART-AI: Substance Misuse Algorithm for Referral to Treatment Using Artificial Intelligence.

Component 6: Cybersecurity

Two principles of security were applied: (1) defense in depth and (2) zero trust. The zero-trust architecture was outlined in the National Institute of Standards and Technology Special Publication 800 to 207 [26]. To secure access between Azure Databricks MLFlow and Epic’s Nebula, we used an authentication token and IP range restriction (Databricks admin utility). The authentication token was issued via Databricks standard authentication. As a security best practice, we used the Databricks service principal and its Databricks access token to provide automated tool and system access to Databricks resources.

Implementation Framework

The Consolidated Framework for Implementation Research–Expert Recommendations for Implementing Change matching tool [29], we will tailor relevant implementation strategies to enhance provider uptake and use of the tool. In addition, during the pilot phase, we will interview providers on the hospital units beyond the pilot units to identify and explore their determinants for the use of the BPA. After a pilot implementation period of 3 months, we will optimize provider training, enhance educational materials, and institute quality monitoring preparatory to a hospital-wide rollout.

Patient Outcome Analysis and Power Calculation

The SMART-AI study intervention sample consisted of all hospitalized patients who were screened positive for opioid misuse through the NLP CDS tool. The primary effectiveness measure was the percentage of hospitalized patients in the NLP CDS intervention sample who screened positive for opioid misuse and received an intervention by the inpatient addiction consult service. A control sample was derived by retrospectively applying the NLP CDS tool to all inpatient EHR records from the 2 years before this study’s initiation in March 2023. Hospitalized patients who were screened positive retrospectively through the NLP CDS tool will form the usual care control group.

The primary outcome was the percentage of inpatients who were screened positive (or would have screened positive)
through the NLP CDS tool and who received an addiction consult with any of the following interventions: (1) receipt of opioid use intervention or motivational interviewing (MI), (2) receipt of medication-assisted treatment (MAT), or (3) referral to substance use disorder treatment. The primary outcome will be reported as a percentage in the preintervention and postintervention periods and will be measured through substance use screening and treatment service engagement for hospitalized patients screened for opioid misuse. The secondary outcomes included the 30-day unplanned hospital readmission rate. The criteria for unplanned hospital readmissions were adopted from the Centers for Medicare and Medicaid Services [30].

Hypothesis testing for intervention effects will be conducted using independent tests of the difference in the proportion of patients receiving MI, MAT, or referral to substance use disorder treatment. The null hypothesis was that the proportion of patients who screened positive and received any of the aforementioned interventions was lower (inferior) in the postintervention period than in the preintervention period, that is, $H_0: p_1 - p_2 \geq M$, where $M$ denotes the noninferiority (eg, equivalence) margin. $p_1$ denotes the preperiod proportion and $p_2$ denotes the postperiod proportion. The alternative 1-tailed test for noninferiority, that is, $H_1$: $p_1 - p_2 < M$, will be tested using the $Z$ statistic. The noninferiority design was adopted to demonstrate that comprehensive screening may be as effective as manual screening but less costly via automated solutions. Our use case was an example of an AI system intended to improve efficiency and throughput within a reasonable timeframe for hospital operations. In these cases, statistically superior performance on outcomes may not be expected or required for prospective implementation, and interventions may be desirable if they are both substantially equivalent (noninferior) on clinical outcomes and cost-effective, given the high cost of building IT infrastructure, hiring vendors, and obtaining licensing and software support.

In hospital-wide screening, we expected a prevalence of 3% of adult inpatients with opioid misuse based on prior findings of hospital-wide analyses. A total sample size of 12,500 patients, with 10,000 in the preintervention 2-year period and 2500 in the postintervention 6-month period, had 85% power to detect a difference of +0.75% in the postintervention period (3.75%) compared with the preintervention period (3%), with a noninferiority difference of $-0.5\%$ using a 1-sided $Z$ test with a significance level of 0.025.

**Cost-effectiveness Analysis**

**Overview**

Cost-effectiveness analysis will estimate the incremental costs of the SMART-AI intervention for the 6 months after the implementation compared with the 6 months before the implementation (ie, the added costs of the SMART-AI tool in reference to usual care) relative to the incremental effectiveness for the primary and secondary outcomes. The health economic evaluation would determine incremental intervention costs by examining the following: (1) the opportunity start-up costs of implementing the SMART-AI tool, (2) the incremental medical costs resulting from usual care for hospitalized patients with opioid misuse versus SMART-AI automated screening–supported care costs, and (3) the ongoing costs of administering and maintaining the SMART-AI tool.

The start-up costs of establishing SMART-AI substance use screening care would include the costs associated with developing and implementing the NLP CDS tool: (1) the cost of supporting the NLP and machine learning components and building the BPA in the EHR and (2) the cost of training the health professionals on tool use. The incremental costs between usual care and SMART-AI automated screening care were determined by calculating medical care costs before and after the implementation of SMART-AI. Medical costs associated with the hospitalization stay and all subsequent medical costs for the 30 days following hospital admission for the pre– and post–SMART-AI intervention periods were derived from hospital billing records and presented from the single-payer (a health system) perspective.

The following 3-pronged approach will be applied to identify the administration and maintenance costs associated with SMART-AI screening workflow changes introduced by the NLP CDS tool: (1) conducting in-depth interviews with hospital administrators, (2) performing activity-based observations of health care personnel who use SMART-AI, and (3) querying the clinician messaging system in the EHR. Average hospital compensation rates were used for valuing health care personnel time costs. Research-related costs were excluded.

**Analytical Approach to Cost-effectiveness Analysis**

The cost-effectiveness analysis was reported in terms of the incremental cost-effectiveness ratio (ICER) per additional patient who received substance use treatment. For this study, the ICER was calculated as the difference between preimplementation and postimplementation intervention costs divided by the difference between preimplementation and postimplementation intervention effectiveness as measured by the rates of patient engagement with substance use treatment services (ie, primary outcome) and 30-day hospital readmission (ie, secondary outcome).

The usual care control group and SMART-AI intervention group were characterized by the pathway probabilities of receiving substance use treatment and meeting the primary outcome. The pathway probabilities of patients’ engagement with inpatient substance use consult, brief intervention or MI, MAT, and referral to substance use treatment for both study groups would result in 8 treatment combinations, which are displayed in Figure 3.
The differential costs pre– and post–SMART-AI intervention were determined as the difference in the weighted sum of the individual pathway costs, using the pathway probabilities as weights for the intervention and control groups. Effectiveness was determined as the difference in the rates of hospitalized patients engaging with substance use disorder treatment before and after the implementation of SMART-AI for the intervention and control groups. The ICER was calculated as follows:

Sensitivity analyses will introduce uncertainty in substance use treatment receipt rates and costs for the intervention and control groups. The Monte Carlo–based simulation estimation used the rates of substance use treatment service uptake observed in the intervention and control groups as a reference to simulate a cohort of postimplementation hospitalized patients and a cohort of usual care hospitalized patients. The ICER per additional individual who received an inpatient substance use consult, brief intervention, MI, MAT, or referral to substance use treatment was calculated by drawing a random sample with replacement from the observed distributions for health care costs ($\mu_{COSTi}$) and substance use treatment services ($\mu_{TRTi}$) for the intervention and control groups. This process was repeated (n=1000) to produce bootstrap estimates of the 95% CI for the ICER per additional individual who received an inpatient substance use consult, brief intervention, MI, MAT, or referral to substance use treatment. These probabilistic sensitivity analyses estimated the elasticity of the differential cost per patient relative to the differential substance use treatment service rates for the intervention and control groups.

**Ethics Approval**

This clinical study was reviewed by the UW Institutional Review Board (ID 2022-0384). The study was part of a larger quality improvement initiative at the UW Health and met the exemption status for human participant research according to the UW Institutional Review Board. The study was secondary research with the collection of existing EHR data that met category 4 exemption. The study met the requirements for a waiver of consent, and all study results were anonymized or deidentified. No compensation was provided in the human participant research.

**Results**

**Preimplementation Testing and Approvals**

Early-stage investigations were performed to assess the AI system’s predictive performance in a retrospective setting and evaluate the human factors surrounding the BPA before initiating the quasi-experimental clinical study. During the silent testing of SMART-AI at the UW Health, a random sample of 100 adult patient encounters (with an oversampling of patients with the International Classification of Diseases codes for substance use) in 2021 were extracted and reviewed by an inpatient physician and a clinical informatics expert. SMART-AI performed similarly to previously published reports for screening for opioid misuse, with a sensitivity of 93% (95% CI 66%-99%) and specificity of 92% (95% CI 84%-96%).

Before the deployment of SMART-AI, approvals were received across hospital committees for inpatient operations, EHR super users, CDS, and nursing documentation. The proposal protocol
was also reviewed by the Center for Clinical Knowledge Management to confirm that there were no competing interests or roles with existing protocols for screening for substance use conditions in the health system. In addition, SMART-AI was reviewed by the UW’s Clinical AI and Predictive Analytics Committee. A model review form providing details on the clinical problem, model value proposition, model description, proposed workflow integration, internal validation, and monitoring strategy (including fairness and equity) was reviewed and approved by a multidisciplinary committee of clinicians, informaticians, bioethicists, executive leadership, and data scientists. The planned workflow from introduction to implementation is shown in Figure 4.

**Figure 4.** Flow diagram for the process to bedside implementation and evaluation. AI: artificial intelligence; SMART-AI: Substance Misuse Algorithm for Referral to Treatment Using Artificial Intelligence.
Implementation Framework

An end-user interview guide and survey were developed to examine the user acceptability of the BPA. Open-ended questions were asked about the barriers to and facilitators of the use of the BPA. A total of 5 interviews were conducted (with 3 nurse practitioners, 1 family medicine resident, and 1 surgical attendant), and the responses led the production team to create an educational flyer, modify the BPA with more details and options for consultation refusal, and modify when and where the BPA would trigger. Figure 2 shows the final production version of the BPA for deployment. Dissemination efforts included Grand Round presentations to the Addiction Medicine Division, Department of Family Medicine, and notification via the hospital’s weekly electronic newsletter.

The longest delay in operational workflow and architecture was for receiving cybersecurity approvals, especially for the exchange of protected health information between the Microsoft and Epic cloud vendors. An additional 6 months of delay occurred for achieving acceptable security monitors and checks. The go-live of SMART-AI in the EHR was scheduled for January 2023.

Discussion

Principal Findings

We offer one of the first protocols that detailed the components of a real-time NLP-driven CDS system for health care delivery at the bedside. We further detailed an implementation framework with human-centered design principles and a planned iterative process to evaluate the cost-effectiveness and health outcomes of screening for opioid misuse. We shared the components and pseudocode with open-source technologies involved in the implementation of an end-to-end NLP pipeline that processed the notes entered by the provider and returned a BPA within minutes for patients at risk of an opioid use disorder. Interviews and user-centered design as well as educational efforts for improving adherence led to changes in the BPA. Finally, we shared an experimental design with a rapid PDSA cycle and cost-effectiveness setup with a noninferiority design to evaluate the screening system for continued implementation or deimplementation.

The digital era in medicine continues to grow exponentially in terms of both the quantity of unstructured data collected in the EHR and the number of prediction models developed for detection and diagnostic, prognostic, and therapeutic guidance. In parallel, the clinical NLP field has grown in its capabilities with the advent of transformer architectures and more affordable and efficient cognitive computing of big data [31]. However, a major bottleneck remains in the successful implementation of NLP and deep learning models in clinical practice. Much of the progress in NLP has focused on information retrieval and extraction [32]; however, the application of these methods at scale with a combination of software developers and operations remains challenging at health care institutions. The role of NLP in BPAs has been limited to date, and prior BPAs have used existing technologies embedded into the EHR [33]. Similar to prior motivations for BPAs delivered to bedside clinicians [34], our intention was to support and enhance decision-making at the bedside with a recommendation for an Addiction Medicine consult in patients who may otherwise not receive it or have it delayed, similar to another NLP-driven BPA [17]. However, given the lack of capacity of many EHR vendors to incorporate custom NLP models, we offer an interoperable pipeline to integrate external AI tools with existing EHRs.

Applied clinical NLP has predominantly remained a rule-based approach, but statistical machine learning models are now the leading method in the research literature [21]. Few vendors who provide NLP services rely entirely on machine learning, and a gap remains in effectively applying NLP models to EHRs that go beyond disease detection, which is limited to explicit keyword mentions [35]. Several barriers exist with neural language models, including the need to remove protected health information so that the trained models may be shared and the computational requirements to run complex deep learning models in a production environment [9]. We offer solutions for both barriers using a feature engineering approach to map free text to coded vocabulary and describe a large computing infrastructure with a connection between a data science cloud platform and the EHR to support direct data feeds into any machine learning model. The NLP CDS pipeline accomplishes efficiency in data standardization and scalability [36] for successful implementation and is extensible to other NLP engines. The benefit of augmented intelligence remains unknown and its identification using our health care outcomes and cost-effectiveness analysis is the next step in a clinical study.

Our implementation framework is largely guided by a team of implementation scientists supported by the UW’s Clinical and Translational Science Award. We leveraged our Clinical and Translational Science Award’s Dissemination and Implementation Launchpad to help bridge the gap between evidence-based research and practice [37]. The Dissemination and Implementation Launchpad serves to accelerate the pace of disseminating research findings and increase the adoption and implementation of effective interventions, leading to sustainable practice and policy changes. It uses strategies from implementation science, design thinking, and human-centered engineering for the better integration of AI technologies into health systems. As part of the preimplementation phase, we assessed contextual factors that may impact implementation by engaging both adopters, who are the decision-makers, and end users, who are the main implementers, of the tool [38]. We conducted qualitative interviews with end users to evaluate the need for the tool and BPA design. We involved adopters early in the process to inform the intervention or implementation process through consultations during the design, feasibility testing, and implementation phases. An iterative process ensued to address the constraints and contextual factors that affect the adoption and implementation of the tool in our health system.

During the preimplementation phase, the project team clarified roles with the project management, with the readiness of the clinical workflow approved through hospital committee meetings and individual interviews with end users. Our health system is an early adopter of AI governance with a review process similar to that of other health systems [39]. The Clinical AI and Predictive Analytics Committee follows the Minimum
Information About Clinical AI Modeling checklist [40]. The offline validation of our model incorporated principles from multiple reporting guidelines on prediction models, bias, fairness, and validation [41]. Clinical evaluation after the go-live of SMART-AI will follow the reporting guideline for the early-stage clinical evaluation of decision support systems driven by AI (Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by AI) [42].

The build of an enterprise-wide AI infrastructure for data-driven CDS is an important feature of a data-driven learning health system. At the UW, learning health system activities dating back to 2013 established an evidence-based framework with a series of organizational-level quality improvement interventions [43]. In 2020, the UW Health reaffirmed its strategic plan for embedding discovery and innovation as well as diversity, equity, and inclusion in clinical care. Successful implementation included coaching staff and administrative leaders for working in PDSA with lean management to get the problem, analysis, corrective actions, and action plan down on a single sheet of large (A3) paper, also known as “A3” thinking [44]. A rapid PDSA cycle is important in the advent of AI-driven interventions that require rigorous evaluation for implementation or deimplementation. Furthermore, the pipeline developed for the opioid screener use case is applicable to other CDS tools that use machine learning and NLP. We designed our architecture to ingest different modalities of data and provide a computing environment that is flexible to different data modalities and machine learning algorithms.

Several limitations exist in the deployment and sustainability of our NLP-driven CDS tool. First, calibration drift is a real concern with changes in medical practice, evidence, and demographic shifts over time that may affect model performance [45]. During implementation, reviews by the Clinical AI and Predictive Analytics Committee will include quarterly evaluations of the sustained effectiveness of the tool, an audit of the fairness of the tool across parity groups, and examination for alert fatigue. Others have shown benefits in recalibration approaches and domain adaptation with additional training data to update the models over time [46]. Furthermore, the start-up costs of the pipeline may be cost-prohibitive for small health systems. Our proposed cost-effectiveness analysis will provide a perspective on both the start-up costs of implementing the NLP tool and the ongoing incremental costs. The start-up costs are more of a burden to a small health system than the incremental costs, but we expect that our results will be informative in terms of both these costs.

**Conclusions**

The deployment of medical AI systems in routine clinical care presents an important yet unfulfilled opportunity [47], and our protocol aims to close the gap in the implementation of AI-driven CDS. Our protocol implementation for an enterprise-wide production environment of an AI opioid misuse screener provides a model for other health systems to use to bring NLP models into practice for CDS. We highlight opportunities to leverage the expertise of our applied data science team to use the open-source tools for feature engineering and model development inside a larger infrastructure with vendor support for hardware and software dependencies. Given the sensitive nature of health care data, the biggest challenges are ensuring high standards for cybersecurity and meeting the privacy requirements for protecting patient data.

**Acknowledgments**

The authors acknowledge support from the University of Wisconsin Institute for Clinical and Translational Research, which, in turn, is supported by the Clinical and Translational Science Award program through the National Institutes of Health National Center for Advancing Translational Sciences grant (2UL1TR002373). The research was also supported by the National Institute on Drug Abuse of the National Institutes of Health (R01DA051464; CJ, DD, MA, and RB), National Library of Medicine Temporal Histories of Your Medical Event (THYME) project (R01LM010090; DD), and National Institute of Diabetes and Digestive and Kidney Diseases R01DK126933). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the other funding sources listed in this section.

**Data Availability**

The raw electronic health record data are not available because of regulatory and legal restrictions imposed by the University of Wisconsin-Madison Institutional Review Board. The original data were derived from the institution’s electronic health record and contain patients’ protected health information. Data are available from the University of Wisconsin Health Systems for researchers who meet the criteria for access to confidential data and have a data use agreement with the health system. Only the final trained model that is fully deidentified with a vocabulary of mapped concept unique identifiers is open source and available [38]. Our deidentification approach has been previously described [39].

**Authors’ Contributions**

MA, FL, BWP, SA, FR, CJ, MPM, DD, and MMC led the conception and design of the study and supervised the study. MA, FL, MPM, TA, GJW, and MC had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. CJ, B Sharma, and DD could not access the original data directly because of limitations in the data use agreement but take responsibility for the accuracy of the data analysis and did have access to all the data presented in the manuscript. MA, SA, MPM, ML, TA, GJW, B Schnapp, MC, CJ, B Sharma, ESB, FR, JM, DD, BWP, MMC, and FL performed
the analysis or interpretation of data. Administrative, technical, and material support were provided by MA, BWP, SA, JL, TA, GJW, B Schnapp, and FL. All authors reviewed the manuscript and provided edits and revisions. All authors take responsibility for the integrity of the work as a whole, from inception to the finished article, and all authors approved the final version submitted. MA was responsible for the decision to submit the manuscript.

Conflicts of Interest

Research conducted by RB is supported by grants to the University of Wisconsin by the Heffter Research Institute, Usona Institute, Revive Therapeutics, and the Etheridge Foundation. MMC is a named inventor on a patent for a risk stratification algorithm for hospitalized patients (US patent #11,410,777).

Multimedia Appendix 1

Pseudocode for custom Python scripts.

References


Abbreviations

AI: artificial intelligence
API: application programming interface
BPA: best practice alert
CDS: clinical decision support
CNN: convolutional neural network
cTAKES: Clinical Text Analysis and Knowledge Extraction System
CUI: concept unique identifier
EHR: electronic health record
HL7: Health Level 7
ICER: incremental cost-effectiveness ratio
MAT: medication-assisted treatment
MI: motivational interviewing
MLFlow: machine learning model life cycle management
NLP: natural language processing
PDSA: Plan-Do-Study-Act
SMART-AI: Substance Misuse Algorithm for Referral to Treatment Using Artificial Intelligence
UMLS: Unified Medical Language System
UW: University of Wisconsin
A Comprehensive and Improved Definition for Hospital-Acquired Pressure Injury Classification Based on Electronic Health Records: Comparative Study

Mani Sotoodeh\textsuperscript{1}, BSc, MSc, PhD; Wenhui Zhang\textsuperscript{2}, BSN, MSc, PhD; Roy L Simpson\textsuperscript{2}, DNP, RN, DPNAP; Vicki Stover Hertzberg\textsuperscript{2}, PhD, PStat; Joyce C Ho\textsuperscript{3}, BSc, MSc, PhD

\textsuperscript{1}Public Health Research Institute of University of Montreal, University of Montreal, Montreal, QC, Canada
\textsuperscript{2}School of Nursing, Emory University, Atlanta, GA, United States
\textsuperscript{3}Department of Computer Science, Emory University, Atlanta, GA, United States

Abstract

Background: Patients develop pressure injuries (PIs) in the hospital owing to low mobility, exposure to localized pressure, circulatory conditions, and other predisposing factors. Over 2.5 million Americans develop PIs annually. The Center for Medicare and Medicaid considers hospital-acquired PIs (HAPIs) as the most frequent preventable event, and they are the second most common claim in lawsuits. With the growing use of electronic health records (EHRs) in hospitals, an opportunity exists to build machine learning models to identify and predict HAPI rather than relying on occasional manual assessments by human experts. However, accurate computational models rely on high-quality HAPI data labels. Unfortunately, the different data sources within EHRs can provide conflicting information on HAPI occurrence in the same patient. Furthermore, the existing definitions of HAPI disagree with each other, even within the same patient population. The inconsistent criteria make it impossible to benchmark machine learning methods to predict HAPI.

Objective: The objective of this project was threefold. We aimed to identify discrepancies in HAPI sources within EHRs, to develop a comprehensive definition for HAPI classification using data from all EHR sources, and to illustrate the importance of an improved HAPI definition.

Methods: We assessed the congruence among HAPI occurrences documented in clinical notes, diagnosis codes, procedure codes, and chart events from the Medical Information Mart for Intensive Care III database. We analyzed the criteria used for the 3 existing HAPI definitions and their adherence to the regulatory guidelines. We proposed the Emory HAPI (EHAPI), which is an improved and more comprehensive HAPI definition. We then evaluated the importance of the labels in training a HAPI classification model using tree-based and sequential neural network classifiers.

Results: We illustrate the complexity of defining HAPI, with <13\% of hospital stays having at least 3 PI indications documented across 4 data sources. Although chart events were the most common indicator, it was the only PI documentation for >49\% of the stays. We demonstrate a lack of congruence across existing HAPI definitions and EHAPI, with only 219 stays having a consensus positive label. Our analysis highlights the importance of our improved HAPI definition, with classifiers trained using our labels outperforming others on a small manually labeled set from nurse annotators and a consensus set in which all definitions agreed on the label.

Conclusions: Standardized HAPI definitions are important for accurately assessing HAPI nursing quality metric and determining HAPI incidence for preventive measures. We demonstrate the complexity of defining an occurrence of HAPI, given the conflicting and incomplete EHR data. Our EHAPI definition has favorable properties, making it a suitable candidate for HAPI classification tasks.
pressure ulcer; decubitus ulcer; electronic medical records; bedsore; nursing; data mining; electronic health record; EHR; nursing assessment; pressure ulcer care; pressure ulcer prevention; EHR data; EHR systems; nursing quality

Introduction

Background and Significance

Hospital-Acquired Pressure Injury, a Key Nursing Metric

Localized damage to the skin or underlying tissues characterizes pressure injury (PI). PI is typically found over a bony prominence or under a medical device and can be caused by lying down or sitting in one place for too long without much movement [1,2]. Hospital-acquired PI (HAPI) is classified according to the PI stage and the time of its development or progression. HAPI is associated with extended hospital stays, high readmission rates, reduced quality of life, and mortality [3]. HAPI is the most frequent preventable adverse event in hospitals according to the Center for Medicare and Medicaid (CMS) and the second most common claim in wrongful death lawsuits [3]. CMS and the Agency for Healthcare Research and Quality (AHRQ) consider HAPI a “never event,” that is, events with profound financial penalties to providers on reimbursement [4]. More than 2000 US hospitals are part of the National Database of Nursing Quality Indicators program to measure nursing quality metrics, that is, events that are directly associated with the quality of nursing care. The National Database of Nursing Quality Indicators requires participating facilities to perform a quarterly survey of patients to estimate the incidence of HAPI [2]. Thus, accurate information on the incidence of HAPI in a health care unit is critical for assessing nursing quality and planning by hospital administrators.

Electronic Health Records and HAPI Identification, Opportunities, and Challenges

Electronic health records (EHRs) provide extensive information on existing and new PIs, including diagnosis codes; characteristics in structured charts, such as stage, depth, and location of PIs; and PI keywords in semistructured or unstructured clinical notes. Automatic detection of HAPI in EHRs using computational models facilitates clinical decision-making and patient care [5]. Predictive models for HAPI depend on the quality, dependability, and consistency of the data set. However, the complexity and subjectivity of PI documentation, detection, and staging impact the reliability of PI documentation. PI documentation reliability also depends on the competency and continuity of nursing staff and their roles as well as changes in data entry or the EHR system.

Despite advances in prevention and treatment, HAPI persists and is difficult to identify from EHRs. Data sources provide contradictory information on PIs. Furthermore, the predictive model accuracy relies heavily on the definition of HAPI and accurate labels. Previous studies using EHR data have used inconsistent definitions of the HAPI. Some describe medical conditions that indicate HAPI [2,6,7]; some identify HAPI in all records associated with a hospital stay [7-9]; and others use prior laboratory data to predict HAPI [10].

Inconsistent HAPI labels adversely impact the model performance in HAPI classification and complicate comparison of multiple models. Thus, the correct identification of HAPI labels from EHR data is essential for HAPI studies. Assessing the performance of machine learning models for HAPI tasks using fixed benchmark data requires accessing appropriate clinical data from EHR databases, unifying multiple data sources, and using them consistently with regulatory guidelines. Here, we propose a HAPI definition that meets these requirements.

Toward a Unified HAPI Definition and More Accurate HAPI Classification

We illustrate the challenges in detecting HAPI in EHRs using the Medical Information Mart for Intensive Care III (MIMIC-III) [11] as a case study. MIMIC-III is one of the most widely used open benchmark data sets, built over CareVue and Metavision EHR systems that encompass approximately 59,000 hospital stays. The patient data include demographics, vital signs, laboratory results, physiological measurements, diagnoses, and clinical and nursing notes.

This study highlights the gaps between existing HAPI definitions for MIMIC-III and the guidelines set by CMS and other regulatory bodies. We propose the Emory HAPI (EHAPI) definition, which better adheres to the regulatory guidelines. We then demonstrate the impact of our improved definition in training a more accurate HAPI classification model. The classification performance was evaluated using a manually labeled set from nurse annotators as a proxy for the HAPI ground truth.

Our main contributions are as follows:

1. An improved HAPI definition that leverages diverse data sources and accounts for their reliability while adhering more closely to clinical guidelines
2. Illustrating the impact of the noncomprehensive HAPI definition on training a HAPI prediction model

To achieve these objectives, we pursue the following steps: (1) describe challenges in finding evidence for HAPI within different sources in the MIMIC-III data set, (2) use nursing expertise with clinical information to prioritize and combine conflicting data sources, (3) establish core parameters for a practically reasonable HAPI definition, and (4) determine the impact of the definition on the performance of tree-based and neural network–based HAPI classifiers.
Methods

Overview of Data Sources for PI in Hospital Stays
There are 4 major sources of EHR data that may contain information on PI: patient chart events, diagnosis codes, notes, and procedures performed.

Chart Events
Chart events constitute the largest portion of structured clinical data and include many medical services, including laboratory tests, vital signs, nurses’ assessments, and general indicators such as patient mental status. Chart events are time-stamped and provide information on the time and order of events during the hospital stay.

Diagnosis Codes
For billing purposes, each hospital stay contains a limited set of diagnosis codes. These codes usually include the most important diagnoses during the hospital stay; however, financial concerns and imperfect mapping of clinical findings to predetermined codes can impact this.

Notes
Clinical notes include any unstructured text information such as radiography reports, electrocardiogram reports, discharge summaries, admission notes, and daily notes made by the care team.

Procedure Codes
Procedure codes indicate timed medical services and surgeries. PI staging is a core element of the HAPI deterioration status from admission to discharge. Most modern clinical information systems, including the Metavision and CareVue clinical information systems in the MIMIC-III and Emory Healthcare’s clinical data warehouse, contain PI staging events and notes. Multimedia Appendix 1 [12] summarizes the details of the PI data sources in the MIMIC-III.

Ideal HAPI Criteria Based on Guidelines
Regulatory authorities identify HAPI using many elements, including the presence of PIs at admission and discharge, changes in stages, unit transfers during admission, and patient death. CMS provides several inclusion and exclusion criteria for HAPI [13-15]. One inclusion criterion is the presence of one or more new or worsened PIs at discharge compared with admission. This includes stages 2 to 4, or PIs not staged owing to slough or eschar, nonremovable dressing or device, or deep tissue injury. Another inclusion criterion is an unstageable PI on admission that is later staged. This is coded on discharge assessment as “present on admission,” with the earliest assessed numerical stage. A patient stay is excluded if data on new or worsened stages 2, 3, and 4, or unstageable pressure ulcers, including deep tissue injuries, are missing on the planned or unplanned discharge assessment. In addition, a patient stay is excluded when the patient died during the hospital stay.

The standard practice for newly admitted patients is the completion of admission assessment, as close as possible to the time of admission and within 24 hours. AHRQ also suggests “performance of comprehensive skin assessment within 24 hours of admission” to accurately assess PI rates [16]. The National Pressure Injury Advisory Panel (NPIAP) reference guide [17] defines the facility-acquired rate as the “percentage of individuals who did not have a pressure injury on admission who acquire a pressure injury during their stay in the facility.”

Existing MIMIC-III HAPI Case Definitions and Their Limitations
There are 4 existing HAPI definitions for MIMIC-III, which are summarized in the subsequent section. Detailed flowcharts for the various definitions are provided in Multimedia Appendix 1.

Recurrent additive network for temporal risk prediction (CANTRIP) [10] focused on predicting HAPI 48 to 96 hours before its first appearance, or date of event (DOE). The DOE was defined as the first occurrence of either mention of PI-related keywords in time-stamped hospital notes or a PI staging chart event (>stage 1) >48 hours after admission. Other stays without a DOE were marked as controls. Unfortunately, the CANTRIP case definition included deceased patients and healed or improved PIs.

Cramer et al [6] sought to develop a screening tool for PI by using the first 24 hours of data. They identified HAPI cases using only the PI staging chart events occurring 24 hours after admission. It excluded stage 1 PIs and “unable to stage” and deep tissue injury PIs. Similar to CANTRIP, the Cramer case definition included deceased patients and healed or improved PIs. Other stays constituted the control group.

Sotoodeh et al [9] explored the use of negation preprocessing on clinical text to detect PI. Case patients were defined using International Classification of Diseases (ICD)-9 codes or PI-specific keywords in the clinical notes. Similar to CANTRIP and Cramer definitions, deceased, healed, or improved PIs were included in the case definition. However, in contrast to the CANTRIP and Cramer definitions, they did not consider PI staging chart events. Control stays were defined as the absence of both ICD-9 codes and PI-specific keywords.

Cox et al [7] focused on identifying appropriate risk factors for PI by using selected variables from the existing literature. They identified a subset of patients who did not have preexisting PI on admission. However, the inclusion and exclusion criteria for identifying HAPI were not explicitly mentioned and are therefore not presented here.

Other studies have focused on predicting HAPI by using other EHR databases. Ranzani et al [8] focused on predicting PI within 30 days of intensive care unit admission in the first 24 hours. They excluded patients who had a preexisting PI on admission or developed PI within the first 48 hours. The case definition was similar to that of CANTRIP, except that notes were not used. Song et al [18] also proposed an early assessment tool for PI risk using 28 relevant features from existing literature. However, the case definition was not discussed in detail. Finally, Hyun et al [19] developed a machine learning model to predict the HAPI. HAPI cases were defined as those containing an ICD-9 code associated with a PI.
EHAPI Case Definition in MIMIC-III

On the basis of existing and ideal HAPI criteria, we identified several essential elements to create a HAPI case definition using EHR data and applied it to MIMIC-III. MIMIC-III has limitations, that is, incongruence of data sources regarding PIs presence and complexity of extracting stage data to verify PI deterioration criteria from admission to discharge, for stays with only comments about PIs in nursing notes and not as timed structured data (Figure 1 and section Limitations and Future Work). These limitations inform this exemplar MIMIC-III HAPI case definition.

Figure 1. Medical Information Mart for Intensive Care III (MIMIC-III) data sources consistency for pressure injury (PI) hospital stays.

For the HAPI criteria, we can include or exclude deceased patients, set a minimum age, and consider either 24 or 48 hours from admission to determine admission PIs status. Further decisions for HAPI criteria include the set of clinical events related to PI hospital stays, the minimum numerical stage for HAPI, numerical stage values assigned to deep tissue injury and unstageable PIs, and the inclusion or exclusion of healed or improved PIs at discharge. Moreover, in addition to staging events, to determine HAPI labels, we considered the presence of certain keywords or diagnosis codes in the notes. We propose a more comprehensive version of the previous definitions, EHAPI. The EHAPI definition is based on the updated version of the HAPI criteria as determined by the CMS, NPIAP, and AHRQ guidelines [2,14-16]. We extracted data from the admission, patient, and intensive care unit stay tables in the MIMIC-III to construct features and remove irrelevant stays from our analysis.

Our case definition considered only new PIs or PIs that deteriorated by discharge, which required determining the PI stages at admission and discharge for each hospital stay. If a patient had multiple hospital stays, we treated each stay separately. Moreover, in MIMIC-III, a hospital stay encompasses ≥1 intensive care unit stays. Staging occurred within 24 hours of admission. Stage 4 is deep PI, and “unable to stage” was coded as 0. In the absence of the PI stage information at admission, the stage was set to 0. Discharge stage was set as the last recorded stage above 2 occurring later than 24 hours of admission, considering deep tissue injury as stage 3 and “unable to stage” as stage 5. “Unable to stage” was set to stage 5 to capture all possible HAPI irrespective of the admission stage. On the basis of NPIAP documentation [20], deep tissue injury was either stage 3 or 4 PI. Therefore, to allow exclusion from the HAPI criteria because of stage improvement during the stay, we coded deep tissue injury as stage 4 at admission and stage 3 at discharge.

We excluded the stays that did not meet the common inclusion criteria. The common inclusion criteria across the four definitions were as follows: (1) presence of at least one clinical note, (2) documented discharge time as after admission time, (3) the patients aged <15 years, and (4) no admission documentation of a PI. EHAPI excluded patients who died in the hospital. We excluded deceased patients for three reasons: (1) adherence to CMS guidelines (including the need for a discharge PI stage that is not available in deceased patients), (2) potential bias of the computational model toward learning characteristics of deceased patients instead of HAPI, and (3) weakness and fragility in patients who have terminal illness result in PI occurrence and do not reflect poor nursing care quality. We conducted an experiment that included deceased patients and observed that some deceased patients were classified as HAPI when they were not HAPI cases.

We found that some HAPI cases lacked PI staging events and yet contained PI keywords in their notes. Thus, the EHAPI also checked the PI-related keywords that occurred in notes later than 24 hours after admission. We scanned all stays for PI-related keywords mentioned in notes later than 24 hours after admission; if present, we considered these cases to be HAPI cases. We used negation detection and analyzed the notes of these cases to ensure that the keywords were not spurious (Multimedia Appendix 1). Other stays constituted the control
Figure 2 provides the flowchart for the EHAPI case definition process. To ensure the generalizability of the EHAPI definition, Multimedia Appendix 1 shows the HAPI-related SNOMED and ICD-10 codes used in many clinical information systems. However, the keywords for notes may need to be tailored to each hospital system. For details of the PI lists, keywords, and mappings in MIMIC-III used for the EHAPI, CANTRIP, Cramer, and Sotoodeh definitions, we refer the reader to Multimedia Appendix 1.

Figure 2. Flowchart for the Emory hospital-acquired pressure injury (EHAPI) definition. Common inclusion criteria across existing definitions and EHAPI are the presence of notes, patients aged 15 years and older, discharge time after admission time, and no pressure injury (PI) diagnosis on admission. D: dimension; DTI: deep tissue injury; HAPI: hospital-acquired pressure injury; MIMIC III: Medical Information Mart for Intensive Care III.

Assessing Impact of HAPI Labels on Classification Performance

We compared the 3 existing HAPI definitions for HAPI classification in MIMIC-III with our EHAPI definition. One systematic review [21] study looked at data-driven models for PI prediction and risk assessment and concluded that many of these predictive models were difficult to compare because they were not externally validated and did not use the same data set.

Event Time Stamp Definition

For each hospital stay, we identified an event time stamp for feature construction. The idea is that for HAPI cases, HAPI-related information is not directly or indirectly present in the features (ie, target or label leakage). Similarly, for non-HAPI cases, we prevented biasing the classifier from predicting longer note durations that would be associated with non-HAPI stays. The event time stamp for HAPI cases is the time of the first PI stage assessment that occurred later than 24 hours after admission. The assessment is the earliest time stamp of either the PI staging chart event or the mention of any of the defined PI keywords in the notes.

For control stays, we matched the non-HAPI duration distribution with the HAPI duration distribution. We modeled the duration of the notes in case stays (the time difference between the earliest note and event time stamp) as a random variable. We used the scikit-learn [22] package to learn the density distribution for this random variable (ie, the estimated distribution with the smallest chi-square score). We then sampled the duration from this estimated distribution for the allowed note length for control stays. Each sampled duration pairs with a true duration length by preserving the ranked order (ie, the fifth smallest duration of sampled length and true length paired with each other). The minimum sampled length and true duration length then serve as the event time stamp (ie, earliest note+sampled length) for the control stay. Thus, if the sampled event time stamp exceeds the stay duration, then the event time stamp is the original stay discharge time.

Notes for HAPI Classification

Hospital stay features are based on patient notes. Machine learning models then only use notes with a time stamp before the event time stamp, or notes of interest, as features. If there were no notes of interest, the stay was excluded from the experiments. The notes of interest were then concatenated into a single document. This minimized the potential for label or target leakage, where HAPI-related information was directly or indirectly present in the features. For instance, defined PI-related keywords do not appear in the concatenated document. Similarly, the feature construction excludes notes after the first staging assessment, thereby preventing implicit PI-related words.
Thus, feature construction excludes all elements discussed in the definition of the HAPI.

**Classifiers**

We chose 2 classifiers to demonstrate that the relative impacts of HAPI labels from the case or control definitions are independent of classifier choice. We chose gradient boosting, a tree-based classifier, and a sequential neural network–based classifier. The latter consisted of input word embedding learned from the features of each definition, a global max pooling layer, and several dense layers. The selected classifiers offer superior performance compared with other tested classifiers (ie, decision tree, logistic regression, support vector machine, multilayer perceptron, random forest, or AdaBoost).

The term frequency–inverse document frequency vector of the abridged notes (described in the aforementioned section) is the feature vector of each stay for the tree-based classifier (with a 5000-word vocabulary). The sequential neural network model uses a sequence of 800 words for each document. The 4 different HAPI definitions (ie, CANTRIP, Cramer, Sotoodeh, and EHAPI) used the same features to yield unbiased model performance comparisons with different definitions.

**Train-Test Compositions and Evaluation Metrics**

Because HAPI criteria differed across definitions, the samples for the prior papers were different (eg, EHAPI discards deceased patients, but others have it either as a case or control). Nevertheless, there were considerable case overlaps across the samples (Figure 3). For a valid comparison, we created 10 different test sets consisting of three parts: (1) consensus HAPI case stays where the definitions agreed, (2) randomly subsampled consensus HAPI control stays, and (3) manually annotated stays where the definitions disagreed. For the latter stays, our nursing experts, a coauthor (WZ), and her nurse colleague (Deborah Silverstein, DNP) assessed and labeled 97 patient stays for HAPI based on the EHR data. The 97 stays constructed each constituent subset proportional to the total size of the differently labeled stay subset (Table S1 in Multimedia Appendix 1). Annotation relies not only on nursing guidelines but also on nursing experience and case discussion between the 2 nurses. Furthermore, one of our nurse annotators (Deborah Silverstein, DNP) was unaware of the EHAPI criteria and labeled the samples from a clinical practitioner’s perspective. Out of the 97 admissions, our nursing experts marked 19 as HAPI.

The manually annotated subset was augmented with consensus stays. There were 219 HAPI cases identified by the 4 definitions, which were included in the 10 test samples. The remaining 3620 non-HAPI stays were randomly sampled from the 41,241 admissions where all 4 definitions agreed on the label. Each test set contained 3936 stays and 7% HAPI prevalence. The main difference between each test set was the 3620 randomly sampled non-HAPI consensus stays, as the 97 manually annotated stays and 219 consensus HAPI cases were present in all test sets.

For each definition, the training samples were the remaining eligible stays that were not in the shared test set. As an example, because CANTRIP did not exclude stays with deceased patients, we included these in the training sample. The training labels were set using definition-specific HAPI criteria. Therefore, although the test samples and labels were the same, the classifier trained for each definition had a different training set and definition-specific label. Figure 3 illustrates the overall process for training the classifier, including the feature construction and label determination.
For the experiments, 5-fold cross-validation of the training set determined the best classifier hyperparameters, as shown in Table S2 in Multimedia Appendix 1. The test performance was the average of 10 different test and training data partitions. Given the unbalanced classes, we report both area under the precision-recall curve (AUPRC) and area under the receiver operating characteristic curve (AUROC).

### Ethical Considerations

The patients were not explicitly recruited to acquire the data used in this work. The MIMIC-III data set has been deidentified through elimination of attributes revealing patients’ identity.

Approval for data collection, processing, and release for the MIMIC-III database was granted by the Institutional Review Boards of the Beth Israel Deaconess Medical Center (Boston, United States) and Massachusetts Institute of Technology (Cambridge, United States).

### Results

#### Consistency of MIMIC-III Data Sources for PI Hospital Stays

Even without consideration of patient attributes or timing, evidence shows conflicts among data sources for identifying PI case hospital stays in MIMIC-III. Figure 1 presents an UpSet plot that summarizes the intersection of PI-related information across the 4 data sources (ie, procedure codes, diagnosis codes, notes, and chart events) for stays with at least one data source indication of PIs (7908 total stays). The data source bar charts (bottom left side) plot the cardinality (size of the number of stays with the data source indication. Chart events were the most common indicator (6243/7908, 78.95%), whereas procedure codes only appeared in 3.53% (279/7908) of the stays. The data source bar charts across the 4 data sources (ie, procedure codes, diagnosis codes, notes, and chart events) for stays with at least one data source indication of PI keywords present in notes are considered an HAPI event.

An UpSet plot capturing the overlap between HAPI stays across the 4 definitions is shown in Figure 4. Only 4.63% (219/4731) of HAPI stays shared 4 definitions. CANTRIP had the highest number of unique positive stays (n=1134), arising from considering PI stages above 1 and deep tissue injury and unstageable events as positives. We observed 315 stays unique to EHAPI, attributed to the cutoff period (24 vs 48 with CANTRIP). EHAPI had the highest overlap, 53.98% (2554/4731) with CANTRIP, followed by Cramer at 22.53% (1066/4731) and Sotoodeh at 17.84% (844/4731). The details of the number of PIs identified using notes, staging, and ICD-9 codes for each definition are provided in Multimedia Appendix 1.

#### Analyzing Differences in HAPI Case Definitions

On the basis of the 4 HAPI case definitions, there are 8 dimensions in which the criteria diverge. The exclusion criteria encompass deceased patients (D1), minimum age (D2), and the amount of time to ascertain PIs on admission (D3). The determination of HAPI includes the minimum PI stage (D4), consideration of deep tissue injury or unstageable events (D5), use of PI-specific keywords in the notes (D6), calculation of deteriorating or new PIs (D7), and use of ICD-9 codes (D8).

Table 1 summarizes the decisions along these 8 dimensions for the 4 different definitions. As can be observed from the table, EHAPI definition excludes the deceased entirely from case or control and ascertains whether the PI deteriorated or newly developed. Both the Cramer and Sotoodeh definitions yielded substantially lower estimates of HAPI prevalence, whereas CANTRIP had the highest prevalence at 8.46% (4261/50,376).

### Table 1. Definition properties and compositions along the 8 criteria dimensions (Ds).

<table>
<thead>
<tr>
<th>Definition</th>
<th>D1(^a)</th>
<th>D2(^b)</th>
<th>D3(^c)</th>
<th>D4(^d)</th>
<th>D5(^e)</th>
<th>D6(^f)</th>
<th>D7(^g)</th>
<th>D8(^h)</th>
<th>Cases, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EHAPI(^i) (n=44,823)</td>
<td>Yes</td>
<td>15</td>
<td>24 h</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>2976 (6.64)</td>
</tr>
<tr>
<td>CANTRIP(^j) (n=50,376) [10]</td>
<td>No</td>
<td>15</td>
<td>48 h</td>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>4261 (8.46)</td>
</tr>
<tr>
<td>Cramer (n=50,276) [6]</td>
<td>No</td>
<td>18</td>
<td>24 h</td>
<td>2</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>1572 (3.13)</td>
</tr>
<tr>
<td>Sotoodeh (n=50,276) [9]</td>
<td>No</td>
<td>18</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>1027 (2.04)</td>
</tr>
</tbody>
</table>

\(^a\)D1 denotes the decision of whether to exclude deceased.

\(^b\)D2 refers to the minimum age in years.

\(^c\)D3 indicates the cutoff period for determining preexisting pressure injury (PI).

\(^d\)D4 characterizes the minimum numerical PI stage.

\(^e\)D5 signifies whether deep tissue injury or unstageable PI staging chart events are hospital-acquired pressure injury (HAPI).

\(^f\)D6 represents whether PI keywords present in notes are considered an HAPI event.

\(^g\)D7 designates whether the criteria captured worsening or newly developed PI.

\(^h\)D8 captures whether International Classification of Diseases 9 codes use HAPI for identification.

\(^i\)EHAPI: Emory hospital-acquired pressure injury.

\(^j\)CANTRIP: recurrent additive network for temporal risk prediction.

\(^k\)N/A: not applicable.
Impact of HAPI Labels on Classification Performance

We evaluated the performance of the gradient boosting and sequential neural network classifiers trained on labels determined by the 4 HAPI definitions by using AUPRC and AUROC for each case. Table 2 presents the results for the 10 described test sets. Table S3 in Multimedia Appendix 1 summarizes the performance based on the test label source (ie, nurse or consensus). Classifiers trained on the EHAPI criteria performed better than those trained on other 3 criteria with an improvement in AUROC up to 0.03 and in AUPRC up to 0.11.

A 1-sided paired t test (1-tailed) between EHAPI and the next best performing definition (CANTRIP) resulted in a P value of <.001 for AUPRC and AUROC for the better-performing gradient boosting classifier and machine epsilon for other classifiers and definitions (except neural networks and CANTRIP), demonstrating the merits of the EHAPI definition. Further analysis of the models’ performance stability and the most important words in each setting are provided in Figure S6 and Table S4 in Multimedia Appendix 1. A GitHub repository [23] contains scripts for these experiments, the generation of the stay labels, and the other presented results.

Table 2. Classifiers’ performance for the 4 hospital-acquired pressure injury definitions in Medical Information Mart for Intensive Care III over 10 test sets. The results represent the average across test sets with the SD in parenthesis and the P value.

<table>
<thead>
<tr>
<th>Definition</th>
<th>Gradient boosting</th>
<th>Neural networks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUPRCA</td>
<td>AUROCb</td>
</tr>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>P value</td>
</tr>
<tr>
<td>EHAPI</td>
<td>0.46 (0.015)</td>
<td>N/A</td>
</tr>
<tr>
<td>CANTRIP [10]</td>
<td>0.44 (0.017)</td>
<td>≤.001</td>
</tr>
<tr>
<td>Cramer [6]</td>
<td>0.35 (0.015)</td>
<td>≤.001</td>
</tr>
<tr>
<td>Sotoodeh [9]</td>
<td>0.33 (0.015)</td>
<td>≤.001</td>
</tr>
</tbody>
</table>

aAUPRC: area under the precision-recall curve.
bAUROC: area under the receiver operating characteristic curve.
cEHAPI: Emory hospital-acquired pressure injury.
dN/A: not applicable.
eCANTRIP: recurrent additive network for temporal risk prediction.
Discussion

Principal Findings

Given the low concurrence of PI between data sources, any HAPI classification requires careful reconciliation of conflicts between data sources. On the basis of discussions with our nursing collaborators (WZ, RN, PhD; Deborah Silverstein, RN, DNP; and RLS, RN, DNP), we prioritized data source reliability as (1) chart events, (2) notes, and (3) diagnosis codes. Charting events were least likely to have a false positive and had better coverage than the other 2 data sources. The nurses indicated that PI indications from notes have false positives, as keywords are preceded by a negative word (ie, no PI), or denote suggestions for PI prevention. Diagnosis codes include only the most prominent diagnoses and might include diagnoses of earlier admissions. In addition, their lack of time stamps prevents investigation of the deterioration condition of the HAPI. Because procedure codes are not specific and inconsistent with other PI sources, we excluded them from the EHAPI definition.

As shown in Table 2, the classifiers trained using the EHAPI definition achieved the best performance. Moreover, the AUROC of the resulting classifiers from 4 definitions were consistently high (≥0.86). The high AUROC is consistent with the CANTRIP results (AUROC of 0.87) [10] and Sotoodeh results (AUROC of 0.95) [9]. However, the AUPRC remains unacceptable, with the highest performance achieved by gradient boosting (0.46). These values are consistent with the existing literature, as CANTRIP reported precision and recall of 0.42 and 0.71, respectively [10], and Cramer reported precision and recall of 0.09 and 0.71, respectively [6]. This illustrates that to identify the HAPI cases, the computational model generates a sizeable portion of false positives.

Limitations and Future Work

CMS-defined guidelines specify that HAPI are only newly developed, unhealed, or deteriorated PIs. Unfortunately, this involves matching admission and discharge PIs, as a patient may be admitted with >1 PI and discharged with more or fewer PIs. The deterioration condition describes each PI individually. However, given the limited data in the event table of MIMIC-III, our case criteria assume that stays are associated with only one PI. Further analysis of multiple possible PI locations yielded better grouping. However, unless skin assessments at admission and discharge are documented in a structured format, matching PIs is difficult. Ideally, the “deteriorated PI” criterion applies to positive PI samples using patient notes as well. However, information on the PI stage is difficult to obtain from notes and, thus, is not implemented in the current case definition. We plan to study the HAPI in other data sets that have better PI documentation practices to fully understand the impact of multiple PIs.

Another limitation of our study is the use of a simple negation detection algorithm to identify false positives occurring with positive PI mentions in the clinical notes. The keyword list disregarded structure matches such as “bedsore: none,” and the negation detection mainly captures instances of text that mentioned “no bed sore observed.” However, instances of negation in more complex textual descriptions may be missed, thus creating false positives in the identified 2976 HAPI stays. A manual inspection of the 1175 case stays labeled through the PI keyword mentions route is left for future work.

Enhancing the manually labeled samples in the 10 test sets beyond the 97 randomly selected ones is another avenue for future research. The small curated set was not large enough for stand-alone analysis, as it yielded large performance variations across the test sets. Unfortunately, it was labor intensive for our nursing annotators to annotate the samples; thus, further annotation is beyond the scope of this work.

In addition, we note that our assessment of the impact of the HAPI definition is based only on MIMIC-III. Furthermore, MIMIC-III contains data collected only in critical care settings. To better understand the performance implications of the HAPI definition, applying implications to other settings, such as the general care units, as well as other health care systems, is needed. We plan to apply these EHAPI criteria to define HAPI in more data sets.

In addition to the focus on critical care stays, the MIMIC-III has unique demographic characteristics, such as predominantly Caucasian. We plan to test the generalizability and impact of the EHAPI case definition against more data sets with diverse demographics including higher percentages of African American, Asian, and Hispanic individuals or different insurance compositions.

A recent systematic review on the utility of decision support systems for PI management concluded that their adoption in practice has clinical significance in terms of reducing PI incidence and prevalence, but statistical significance was not observed [24]. This emphasizes the importance of studying practical challenges in the adoption of data-driven PI methods by nurses. Moreover, the practical deployment of a computational model necessitates a higher AUPRC to prevent false alarms. Thus, an open question is whether the integration of other patient information in addition to clinical notes, such as physiological measurements, patient demographics, and medications, yields better predictive performance.

Conclusions

An accurate definition of HAPI based on clinical data is critical for automating nursing quality metrics and for valid comparisons of HAPI machine learning models. However, one of the major challenges is the inconsistency of the PI indicators across various data sources. We demonstrate the lack of congruency between the 3 existing HAPI definitions for MIMIC-III and highlight the gaps between each definition and the CMS and AHRQ regulatory guidelines. We then created a refined definition, the EHAPI, that more closely reflects the regulatory guidelines. Our experimental results using 2 different classifiers illustrate the impact of the definition on the predictive performance when evaluated on an unseen combination of a small, manually labeled set by 2 nurse annotators and a random sample of the consensus set (ie, all 4 definitions agree on the labels). This reinforces the need for a high-quality standardized HAPI definition, as the EHAPI achieves a better predictive performance across multiple test sets.
Acknowledgments
The authors express their deepest gratitude to the following nurse collaborators: Cynthia A Oster, PhD, RN, APRN, MBA, ACNS-BC, ANP, FAAN, Nurse Scientist for Patient Safety—Emory Healthcare, Adjunct Assistant Professor—Nell Hodgson Woodruff School of Nursing, Emory University; and Deborah Silverstein, DNP, APRN, FNP-C, Instructor, Nell Hodgson Woodruff School of Nursing, Emory University. The authors also thank the anonymous reviewers for their suggestions and comments.

This research was supported by the National Library of Medicine of the National Institutes of Health under the award number R01LM013323-01.

Data Availability
Medical Information Mart for Intensive Care III (MIMIC-III) data can be downloaded from the PhysioNet webpage after completing the required Collaborative Institutional Training Initiative, data or human subjects research training. The scripts used to preprocess the data and obtain the presented results can be found in the GitHub repository.

Authors' Contributions
All authors conceptualized the study. MS designed and conducted the data analysis and generated the results. MS conducted the data preprocessing. MS is responsible for the integrity of the work. MS drafted the paper. WZ and RLS provided nursing expertise throughout the study and provided recommendations for study parameters. All authors participated in the writing and revising of the manuscript. All aspects of the study (design; management, analysis, and interpretation of data; writing of the report; and decision to publish) were led by the authors. All authors have read and approved the final manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
The supplementary material containing the Medical Information Mart for Intensive Care III details for the Emory hospital-acquired pressure injury definition, additional comparisons between the 4 case definitions, and further experimental results related to this study.

References


14. CMS guideline for LTCH Quality reporting. CMS. URL: https://tinyurl.com/26ct2yfs [accessed 2021-04-07]


23. GitHub. URL: https://github.com/manisci/EHAPI [accessed 2023-02-07]


Abbreviations

AHRQ: Agency for Healthcare Research and Quality
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
CANTRIP: recurrent additive network for temporal risk prediction
CMS: Center for Medicare and Medicaid
DOE: date of event
EHAPI: Emory hospital-acquired pressure injury
EHR: electronic health record
HAPI: hospital-acquired pressure injury
ICD: International Classification of Diseases
MIMIC-III: Medical Information Mart for Intensive Care III
NPIAP: The National Pressure Injury Advisory Panel
PI: pressure injury
Identification of Postpartum Depression in Electronic Health Records: Validation in a Large Integrated Health Care System

Jeff Slezak¹, MS; David Sacks¹,², MD; Vicki Chiu¹, MS; Chantal Avila¹, MA; Nehaa Khadka¹, MPH; Jiu-Chiuan Chen², MD, SCD; Jun Wu³, PhD; Darios Getahun¹,⁴, MD, PhD

¹Kaiser Permanente Southern California, Pasadena, CA, United States
²Keck School of Medicine, University of Southern California, Los Angeles, CA, United States
³Program in Public Health, Susan and Henry Samuei College of Health Sciences, University of California, Irvine, CA, United States
⁴Kaiser Permanente Bernard J. Tyson School of Medicine, Pasadena, CA, United States
*these authors contributed equally

Corresponding Author:
Jeff Slezak, MS
Kaiser Permanente Southern California
100 S. Los Robles Ave
Pasadena, CA, 91101
United States
Phone: 1 626 564 3477
Email: Jeff.M.Slezak@kp.org

Abstract

Background: The accuracy of electronic health records (EHRs) for identifying postpartum depression (PPD) is not well studied.

Objective: This study aims to evaluate the accuracy of PPD reporting in EHRs and compare the quality of PPD data collected before and after the implementation of the International Classification of Diseases, Tenth Revision (ICD-10) coding in the health care system.

Methods: Information on PPD was extracted from a random sample of 400 eligible Kaiser Permanente Southern California patients’ EHRs. Clinical diagnosis codes and pharmacy records were abstracted for two time periods: January 1, 2012, through December 31, 2014 (International Classification of Diseases, Ninth Revision [ICD-9] period), and January 1, 2017, through December 31, 2019 (ICD-10 period). Manual chart reviews of clinical records for PPD were considered the gold standard and were compared with corresponding electronically coded diagnosis and pharmacy records using sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Kappa statistic was calculated to measure agreement.

Results: Overall agreement between the identification of depression using combined diagnosis codes and pharmacy records with that of medical record review was strong (κ=0.85, sensitivity 98.3%, specificity 83.3%, PPV 93.7%, NPV 95.0%). Using only diagnosis codes resulted in much lower sensitivity (65.4%) and NPV (50.5%) but good specificity (88.6%) and PPV (93.5%). Separately, examining agreement between chart review and electronic coding among diagnosis codes and pharmacy records showed sensitivity, specificity, and NPV higher with prescription use records than with clinical diagnosis coding for PPD, 96.5% versus 72.0%, 96.5% versus 65.0%, and 96.5% versus 65.0%, respectively. There was no notable difference in agreement between ICD-9 (overall κ=0.86) and ICD-10 (overall κ=0.83) coding periods.

Conclusions: PPD is not reliably captured in the clinical diagnosis coding of EHRs. The accuracy of PPD identification can be improved by supplementing clinical diagnosis with pharmacy use records. The completeness of PPD data remained unchanged after the implementation of the ICD-10 diagnosis coding.

(JMIR Med Inform 2023;11:e43005) doi:10.2196/43005

KEYWORDS
validation; postpartum depression; electronic health records; pregnancy; health care system; diagnosis codes; pharmacy records; health data; data collection; implementation; eHealth record; depression; mental well-being; women’s health
Introduction

Postpartum depression (PPD), major or minor depressive episodes occurring within 12 months after childbirth, is a common obstetric complication in the United States, with a prevalence of 13.2% in 2018 [1]. The American College of Obstetricians and Gynecologists recommends all obstetric care providers conduct comprehensive screening for PPD and anxiety disorders using a validated instrument for each patient separately during their postpartum visit [2]. Meanwhile, the American Academy of Pediatrics recommended routine PPD screening to be integrated at well-child visits (1-, 2-, 4-, and 6-month infant visits) [3]. The US Preventive Services Task Force also supports the provision of depression screening during postpartum visits, citing moderate net benefits for identifying those affected and recommending referrals to counseling interventions [4]. It is important to identify those with PPD because undetected or untreated depressive episodes can negatively impact the patient and their infant’s health and well-being. For instance, about 9% of pregnancy-related deaths were due to mental health conditions [5]. Early PPD was also associated with increased behavior disturbances in the infant [6]. Moreover, other potential risk factors, including a prior history of depression, depression and anxiety episodes during pregnancy, preterm birth and lower infant birth weight, traumatic birth experience, stressful life events during early postpartum, and low social support, have been linked with PPD [7-9].

Health systems previously used the International Classification of Diseases, Ninth Revision (ICD-9), an official coding system to identify hospital-related diagnoses and procedures in the United States [10]. However, the Kaiser Permanente health systems shifted to using the International Classification of Diseases, Tenth Revision (ICD-10) codes after October 1, 2015, which has significant improvements over ICD-9 for many clinical codes [11]. However, Stewart et al [12] concluded that there is a need to perform a validation of diagnosis codes for each mental health condition following the ICD-10 transition. Colvin et al [13] used a data linkage of national pharmacy records and hospital admission information to identify patients with major depressive episodes in pregnancy but found the use of either source alone to be inadequate.

While there are multiple validated scales to screen for PPD, like the Patient Health Questionnaire (9-item) and the Edinburgh Postnatal Depression Scale, validation of these measures has been performed using ICD-9 or ICD-10 diagnostic codes as the gold standard [14,15]. Several studies have also developed machine learning algorithms using electronic health record (EHR) data to create risk-based models and examined whether they can predict PPD in large health care systems, relying on PPD ascertainment using ICD-9 or ICD-10 codes [16,17]. However, the accuracy of ICD-9 and ICD-10 codes as the gold standard in ascertaining PPD has not been established previously. Prior validation of ICD-9 and ICD-10 found high positive predictive values (PPVs) for ascertaining general depression (89.7% and 89.5%, respectively), but these were not specific to the postpartum period [18]. This study aimed to assess the validity of ascertaining PPD diagnosis using the EHR from a large integrated health care delivery system, Kaiser Permanente Southern California (KPSC).

Methods

Cohort Selection

We identified a random sample of 400 women with live birth records in the Air Pollution and Pregnancy Complications in Complex Urban Environments (APPCUE) study [19] between January 1, 2008, and December 31, 2018, within KPSC, a large integrated health system. The APPCUE study was a retrospective cohort study conducted in collaboration between KPSC and the University of California, Irvine with access to KPSC’s comprehensive EHRs. The APPCUE study included all singleton births at KPSC facilities. The EHRs contain patient-level data from out- and inpatient clinical care, including ICD-9, Clinical Modification or ICD-10, Clinical Modification diagnosis and procedure codes, as well as pharmacy and laboratory test records. From 236,759 pregnancies during the study period, we excluded pregnancies resulting in nonlive births (n=8422) and patients who were not members from the start of their pregnancy through a 1-year postpartum period (n=70,836) to have a complete medical history for this validation study. Of the remaining 157,501 pregnancies, we selected a random sample of 400. Simple random sampling was used to select 100 patients from groups based on EHR data: those without any diagnostic or pharmacy use record for PPD, those with only a diagnostic code for PPD, those with only a pharmacy record indicating treatment for PPD, and those with both diagnostic and pharmacy indications. Additionally, each sample was evenly split (50 each) between the ICD-9 diagnosis code era (date of delivery 2012-2014) and the ICD-10 era (2017-2019).

Outcomes

EHR outcomes were determined by the presence of PPD diagnosis codes in inpatient or outpatient encounters in the 12 months after delivery, new prescription order, or pharmacy dispense for the treatment of PPD. Diagnosis codes during the ICD-9 coding period were 300.4, 309.0, and 311 and during the ICD-10 period were F32.9, F33.0, F33.2, F33.3, F33.41, F33.9, F34.1, F43.21, and F53.0. Medications included were bupropion, Celexa, citalopram, Cymbalta, desvenlafaxine, duloxetine, Effexor, escitalopram, fluoxetine, Lexapro, paroxetine, Paxil, Pristiq, Prozac, sertraline, venlafaxine, Wellbutrin, and Zoloft.

Gold standard PPD outcomes were determined by review of health records by trained research personnel, who documented any diagnosis or finding of PPD in the record, including in free-text encounter notes, as well as any prescription given for the treatment of PPD. These included new prescriptions for the treatment of PPD. PPD diagnosis and medication were documented independently, both for the EHR data and the chart review. A mother was considered to have PPD if she had either a diagnosis or a prescription noted in the EHR within 1 year postpartum.

Quality Assurance

Multiple individuals were trained on reviewing charts, and a double chart review was performed at the beginning of data collection, with agreement confirmation after 12 months.
collection as a training exercise and near the middle and at the end of data collection to verify data quality and consistency. At each point, eight charts were randomly selected for review by two abstractors. In case of disagreement on the findings, abstractors met with the trainer to determine the correct result.

**Statistical Analysis**

The patient population was described in terms of demographics, smoking status, prenatal care, and birth weight using percentages. These characteristics were also described for the study population of the APPCUE study [19] and all live births among KPSC members and the state of California during the study timeframe. The chi-square test was used to compare the distribution of characteristics in the study sample to the APPCUE population, all KPSC births, and the California birth cohort.

Manual chart review findings were treated as the true PPD status. The sensitivity, specificity, PPV, and negative predictive value (NPV) of the electronic records to identify true PPD status were calculated and presented as a percentage and 95% exact binomial CI. Agreement between electronic records and manual review was calculated using the kappa statistic, which adjusts for agreement expected due to random chance, and its 95% CI. The area under the receiver operating characteristic curve was calculated. Each measure was calculated overall and within the ICD-9 and ICD-10 coding eras separately. There was no missing data for PPD status; those without documented PPD diagnosis or medication were taken to not have PPD. For patient characteristics, a missing category was included when presenting the data.

The primary analysis focused on the ability of EHRs to capture PPD, while secondary analyses examined the agreement of diagnosis and prescription records separately. The sample size was selected so that the expected width of the CIs for sensitivity and PPV would be at most 10% for the full sample and 13% for the ICD-9 and ICD-10 periods if the true sensitivity and PPV were 80%. Higher sensitivity and PPV would yield narrower CIs. The STARD (Standards for Reporting Diagnostic Accuracy Studies) guidelines were followed. All analyses were performed in SAS version 9.4 (SAS Institute).

**Ethics Approval**

The study was approved by the institutional review board of KPSC and received a waiver for informed consent (IRB 12110).

**Results**

**Cohort Selection**

Table 1 shows the distribution of the APPCUE study cohort as well as the overall KPSC birth cohort during the study period. Nearly half (194/400, 48.5%) were Latina, most (379/400, 94.8%) received prenatal care starting in the first trimester, and most (354/400, 88.5%) delivered at 37 weeks of gestation or later. The study sample generally has very similar characteristics to the APPCUE study cohort overall and all KPSC births during the period, though there are some differences relative to all deliveries in the state of California, notably a higher percentage of non-Hispanic White mothers (113/400, 28.3% vs 372,037/2,874,396, 12.9%), older mothers (259/400, 64.8% age ≥30 years vs 1,465,998/2,874,396, 50.0%), and generally higher educational attainment (199/400, 49.8% with at least a college degree vs 1,047,594/2,874,396, 36.5%).
Table 1. Characteristics of the study sample and women delivered in all Kaiser Permanente Southern California (KPSC) hospitals and the state of California (2012-2014 and 2017-2019).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Chart review sample&lt;sup&gt;a&lt;/sup&gt; (N=400), n (%)</th>
<th>APPCUE&lt;sup&gt;b&lt;/sup&gt; study population (N=157,501), n (%)</th>
<th>P value</th>
<th>All KPSC births&lt;sup&gt;c&lt;/sup&gt; (N=236,759), n (%)</th>
<th>P value</th>
<th>All California State births&lt;sup&gt;c&lt;/sup&gt; (N=2,874,396), n (%)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Maternal age (years)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;20</td>
<td>10 (2.5)</td>
<td>4665 (3.0)</td>
<td>.42</td>
<td>6804 (3.0)</td>
<td>.01</td>
<td>144,945 (5.0)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>20-29</td>
<td>131 (32.8)</td>
<td>56,679 (36.0)</td>
<td></td>
<td>92,203 (40.4)</td>
<td></td>
<td>1,263,453 (44.0)</td>
<td></td>
</tr>
<tr>
<td>30-34</td>
<td>153 (38.3)</td>
<td>54,810 (34.8)</td>
<td></td>
<td>75,633 (33.1)</td>
<td></td>
<td>843,010 (29.3)</td>
<td></td>
</tr>
<tr>
<td>≥35</td>
<td>106 (26.5)</td>
<td>41,347 (26.3)</td>
<td></td>
<td>53,697 (23.5)</td>
<td></td>
<td>622,988 (21.7)</td>
<td></td>
</tr>
<tr>
<td><strong>Race/ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>113 (28.3)</td>
<td>39,219 (24.9)</td>
<td>.31</td>
<td>55,218 (24.2)</td>
<td>.29</td>
<td>372,037 (12.9)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>32 (8.0)</td>
<td>10,862 (6.9)</td>
<td></td>
<td>16,207 (7.1)</td>
<td></td>
<td>68,195 (2.4)</td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>194 (48.5)</td>
<td>78,853 (50.1)</td>
<td></td>
<td>117,162 (51.3)</td>
<td></td>
<td>1,356,354 (47.2)</td>
<td></td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>47 (11.8)</td>
<td>22,783 (14.5)</td>
<td></td>
<td>31,318 (13.7)</td>
<td></td>
<td>213,499 (7.4)</td>
<td></td>
</tr>
<tr>
<td>Others/unknown</td>
<td>14 (3.5)</td>
<td>5784 (3.7)</td>
<td></td>
<td>8432 (3.7)</td>
<td></td>
<td>864,311 (30.1)</td>
<td></td>
</tr>
<tr>
<td><strong>Educational attainment</strong></td>
<td></td>
<td></td>
<td>.36</td>
<td></td>
<td>.20</td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Less than high school</td>
<td>9 (2.2)</td>
<td>4355 (2.8)</td>
<td></td>
<td>6925 (3.0)</td>
<td></td>
<td>435,360 (15.1)</td>
<td></td>
</tr>
<tr>
<td>High school graduate</td>
<td>83 (20.8)</td>
<td>35,411 (22.5)</td>
<td></td>
<td>55,598 (24.4)</td>
<td></td>
<td>694,118 (24.1)</td>
<td></td>
</tr>
<tr>
<td>Some college</td>
<td>99 (24.8)</td>
<td>32,616 (20.7)</td>
<td></td>
<td>50,153 (22.0)</td>
<td></td>
<td>558,288 (19.4)</td>
<td></td>
</tr>
<tr>
<td>Bachelor’s/associate’s degree</td>
<td>126 (31.5)</td>
<td>54,293 (34.5)</td>
<td></td>
<td>75,849 (33.2)</td>
<td></td>
<td>729,896 (25.4)</td>
<td></td>
</tr>
<tr>
<td>Master’s degree/above</td>
<td>73 (18.3)</td>
<td>27,388 (17.4)</td>
<td></td>
<td>34,556 (15.1)</td>
<td></td>
<td>317,698 (11.1)</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>10 (2.5)</td>
<td>3438 (2.2)</td>
<td></td>
<td>5256 (2.3)</td>
<td></td>
<td>139,036 (4.8)</td>
<td></td>
</tr>
<tr>
<td><strong>Household income (US $)</strong></td>
<td></td>
<td></td>
<td>.64</td>
<td></td>
<td>.25</td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>&lt;30,000</td>
<td>16 (4.0)</td>
<td>5194 (3.3)</td>
<td></td>
<td>8318 (3.6)</td>
<td></td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>30,000-49,999</td>
<td>90 (22.5)</td>
<td>39,969 (25.4)</td>
<td></td>
<td>61,562 (27.0)</td>
<td></td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>50,000-69,999</td>
<td>124 (31.0)</td>
<td>47,864 (30.4)</td>
<td></td>
<td>69,844 (30.6)</td>
<td></td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>70,000-89,999</td>
<td>82 (20.5)</td>
<td>32,486 (20.6)</td>
<td></td>
<td>45,469 (19.9)</td>
<td></td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>≥90,000</td>
<td>88 (22.0)</td>
<td>31,925 (20.3)</td>
<td></td>
<td>42,782 (18.7)</td>
<td></td>
<td>—</td>
<td></td>
</tr>
<tr>
<td><strong>Prenatal care initiation</strong></td>
<td></td>
<td></td>
<td>.52</td>
<td></td>
<td>&lt;.001</td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>First trimester</td>
<td>379 (94.8)</td>
<td>147,017 (93.3)</td>
<td></td>
<td>199,866 (87.5)</td>
<td></td>
<td>2,386,232 (83.0)</td>
<td></td>
</tr>
<tr>
<td>No or late care</td>
<td>20 (5.0)</td>
<td>9860 (6.3)</td>
<td></td>
<td>26,966 (11.8)</td>
<td></td>
<td>442,493 (15.4)</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>1 (0.2)</td>
<td>624 (0.4)</td>
<td></td>
<td>1505 (0.7)</td>
<td></td>
<td>45,671 (1.6)</td>
<td></td>
</tr>
<tr>
<td>Smoking during pregnancy</td>
<td>23 (5.8)</td>
<td>6420 (4.1)</td>
<td>.09</td>
<td>10,256 (4.5)</td>
<td>.16</td>
<td>46,977 (1.6)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Gestational age (weeks)</strong></td>
<td></td>
<td></td>
<td>.13</td>
<td></td>
<td>.08</td>
<td></td>
<td>.16</td>
</tr>
<tr>
<td>&lt;34</td>
<td>14 (3.5)</td>
<td>3412 (2.2)</td>
<td></td>
<td>4779 (2.1)</td>
<td></td>
<td>66,099 (2.3)</td>
<td></td>
</tr>
<tr>
<td>34-36</td>
<td>32 (8.0)</td>
<td>9865 (6.3)</td>
<td></td>
<td>13,933 (6.1)</td>
<td></td>
<td>180,352 (6.3)</td>
<td></td>
</tr>
<tr>
<td>≥37</td>
<td>354 (88.5)</td>
<td>144,192 (91.5)</td>
<td></td>
<td>209,553 (91.8)</td>
<td></td>
<td>2,624,620 (91.3)</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>0 (0.0)</td>
<td>32 (0.0)</td>
<td></td>
<td>72 (0.0)</td>
<td></td>
<td>3325 (0.1)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Sample is based on data from KPSC electronic health records 2012-2014 and 2017-2019.

<sup>b</sup>APPCUE: Air Pollution and Pregnancy Complications in Complex Urban Environments.

<sup>c</sup>Data from the natality information of the Center for Disease Control and Prevention [20].

<sup>d</sup>Data not available.
Outcomes
The overall agreement of EHR-identified PPD (based on either a diagnosis or a prescription) with medical record review was high, with a kappa of 84.7% (95% CI 78.8%-90.6%). The EHR identified 281 of 286 cases (sensitivity 98.3%, 95% CI 96.0%-99.4%) while maintaining high specificity (95.0%, 95% CI 88.7%-98.4%), PPV (93.7%, 95% CI 90.3%-96.1%), and NPV (95.0%, 95% CI 88.7%-98.4%). There was little difference in the overall agreement between the ICD-9 coding era (κ=86.0%, 95% CI 78.0%-94.0%) and the ICD-10 era (κ=83.4%, 95% CI 74.8%-92.1%; Table 2).

Table 2. Identification of postpartum depression using diagnostic codes and/or pharmacy records–based data sources before and after implementation of the ICD-10 code in the Kaiser Permanente Southern California system in 2015 (N=400).

<table>
<thead>
<tr>
<th></th>
<th>TP, n</th>
<th>TN, n</th>
<th>FP, n</th>
<th>FN, n</th>
<th>Sensitivity, % (95% CI)</th>
<th>Specificity, % (95% CI)</th>
<th>PPV, % (95% CI)</th>
<th>NPV, % (95% CI)</th>
<th>Kappa (95% CI)</th>
<th>AUC(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Combined electronic diagnosis codes and pharmacy records</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>281</td>
<td>95</td>
<td>19</td>
<td>5</td>
<td>98.3 (96.0-99.4)</td>
<td>83.3 (75.2-89.7)</td>
<td>93.7 (90.3-96.1)</td>
<td>95.0 (88.7-98.4)</td>
<td>0.85 (0.79-0.91)</td>
<td>0.91</td>
</tr>
<tr>
<td>2012-2014</td>
<td>141</td>
<td>48</td>
<td>9</td>
<td>2</td>
<td>98.6 (95.0-99.8)</td>
<td>84.2 (72.1-92.5)</td>
<td>94.0 (88.9-97.2)</td>
<td>96.0 (86.3-99.5)</td>
<td>0.86 (0.78-0.94)</td>
<td>0.91</td>
</tr>
<tr>
<td>2017-2019</td>
<td>140</td>
<td>47</td>
<td>10</td>
<td>3</td>
<td>97.9 (94.0-99.6)</td>
<td>82.5 (70.1-91.3)</td>
<td>93.3 (88.1-96.8)</td>
<td>94.0 (83.5-98.7)</td>
<td>0.83 (0.75-0.92)</td>
<td>0.90</td>
</tr>
<tr>
<td><strong>ICD-9(^b)/ICD-10(^i) diagnosis codes only</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>187</td>
<td>101</td>
<td>13</td>
<td>99</td>
<td>65.4 (59.6-70.9)</td>
<td>88.6 (81.3-93.8)</td>
<td>93.5 (89.1-96.5)</td>
<td>50.5 (43.4-57.6)</td>
<td>0.44 (0.36-0.52)</td>
<td>0.77</td>
</tr>
<tr>
<td>2012-2014</td>
<td>94</td>
<td>51</td>
<td>6</td>
<td>49</td>
<td>65.7 (57.3-73.5)</td>
<td>89.5 (78.5-96.0)</td>
<td>94.0 (87.4-97.8)</td>
<td>51.0 (40.8-61.1)</td>
<td>0.45 (0.34-0.56)</td>
<td>0.78</td>
</tr>
<tr>
<td>2017-2019</td>
<td>93</td>
<td>50</td>
<td>7</td>
<td>50</td>
<td>65.0 (56.6-72.8)</td>
<td>87.7 (76.3-94.9)</td>
<td>93.0 (86.1-97.1)</td>
<td>50.0 (39.8-60.2)</td>
<td>0.43 (0.32-0.54)</td>
<td>0.76</td>
</tr>
<tr>
<td><strong>Pharmacy records only</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>194</td>
<td>108</td>
<td>6</td>
<td>92</td>
<td>67.8 (62.1-73.2)</td>
<td>94.7 (88.9-98.0)</td>
<td>97.0 (93.6-98.9)</td>
<td>54.0 (46.8-61.1)</td>
<td>0.51 (0.43-0.59)</td>
<td>0.81</td>
</tr>
<tr>
<td>2012-2014</td>
<td>97</td>
<td>54</td>
<td>3</td>
<td>46</td>
<td>67.8 (59.5-75.4)</td>
<td>94.7 (85.4-98.9)</td>
<td>97.0 (91.5-99.4)</td>
<td>54.0 (43.7-64.0)</td>
<td>0.51 (0.40-0.62)</td>
<td>0.81</td>
</tr>
<tr>
<td>2017-2019</td>
<td>97</td>
<td>54</td>
<td>3</td>
<td>46</td>
<td>67.8 (59.5-75.4)</td>
<td>94.7 (85.4-98.9)</td>
<td>97.0 (91.5-99.4)</td>
<td>54.0 (43.7-64.0)</td>
<td>0.51 (0.40-0.62)</td>
<td>0.81</td>
</tr>
</tbody>
</table>

\(^a\)TP: true positive.
\(^b\)TN: true negative.
\(^c\)FP: false positive.
\(^d\)FN: false negative.
\(^e\)PPV: positive predictive value.
\(^f\)NPV: negative predictive value.
\(^g\)AUC: area under the receiver operating characteristic curve.
\(^h\)ICD-9: International Classification of Diseases, Ninth Revision.
\(^i\)ICD-10: International Classification of Diseases, Tenth Revision.

Electronic diagnosis records alone were not able to accurately identify PPD, only identifying 187 of 286 cases (sensitivity 65.4%, 95% CI 59.6%-70.9%), with low NPV (50.5%, 95% CI 43.4%-57.6%). PPV (93.5%, 95% CI 89.1%-96.5%) and specificity (88.6%, 95% CI 81.3%-93.8%) were high, however (Table 2). Results were similar when using EHR prescription records alone (sensitivity 67.8%, 95% CI 62.1%-73.2%; specificity 94.7%, 95% CI 88.9%-98.0%; PPV 97.0%, 95% CI 93.6%-98.9%; NPV 54.0%, 95% CI 46.8%-61.1%).

Considering only medication data, the reliability of EHR data for identifying prescriptions for PPD was high, with an overall kappa of 92.5% (95% CI 88.8%-96.2%). Agreement was very high in both the ICD-9 (κ=92.0%, 95% CI 86.6%-97.4%) and ICD-10 eras (κ=93.0%, 95% CI 87.9%-98.1%; Table 3). Sensitivity, specificity, PPV, and NPV were all at or above 96% (Table 3).

Agreement for ICD diagnostic codes between EHR and manual chart review was much lower overall (κ=55.0%, 95% CI 47.1%-62.9%; Table 3). The PPV was high (90.0%, 95% CI 85.0%-93.8%), with sensitivity lower (72.0%, 95% CI 66.0%-77.5%) and specificity and NPV much lower (both 65.0%, 95% CI 58.0%-71.6%; Table 3). Agreement was similar
between the ICD-9 (κ=58.0%, 95% CI 47.1%-68.9%) and ICD-10 (κ=52.0%, 95% CI 40.5%-63.5%) eras (Table 3).

Table 3. Identification of postpartum depression based on individual data sources before and after implementation of the ICD-10 code in the Kaiser Permanente Southern California system in 2015 (N=400).

<table>
<thead>
<tr>
<th></th>
<th>TP, n</th>
<th>TN, n</th>
<th>FP, n</th>
<th>FN, n</th>
<th>Sensitivity, % (95% CI)</th>
<th>Specificity, % (95% CI)</th>
<th>PPV, % (95% CI)</th>
<th>NPV, % (95% CI)</th>
<th>Kappa (95% CI)</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ICD-9</strong>/ICD-10 diagnosis codes only**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>180</td>
<td>130</td>
<td>20</td>
<td>70</td>
<td>72.0 (66.0-77.5)</td>
<td>86.7 (80.2-91.7)</td>
<td>90.0 (85.0-93.8)</td>
<td>65.0 (58.0-71.6)</td>
<td>0.55 (0.47-0.63)</td>
<td>0.79</td>
</tr>
<tr>
<td>2012-2014</td>
<td>92</td>
<td>66</td>
<td>8</td>
<td>34</td>
<td>73.0 (64.4-80.5)</td>
<td>89.2 (79.8-95.2)</td>
<td>92.0 (84.8-96.5)</td>
<td>66.0 (55.8-75.2)</td>
<td>0.58 (0.47-0.69)</td>
<td>0.78</td>
</tr>
<tr>
<td>2017-2019</td>
<td>88</td>
<td>64</td>
<td>12</td>
<td>36</td>
<td>71.0 (62.1-78.8)</td>
<td>84.2 (74.0-91.6)</td>
<td>88.0 (80.0-93.6)</td>
<td>64.0 (53.8-73.4)</td>
<td>0.52 (0.41-0.64)</td>
<td>0.81</td>
</tr>
<tr>
<td><strong>Pharmacy records only</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>192</td>
<td>193</td>
<td>8</td>
<td>7</td>
<td>96.5 (92.9-98.6)</td>
<td>96.0 (92.3-98.3)</td>
<td>96.0 (92.3-98.3)</td>
<td>96.5 (92.9-98.6)</td>
<td>0.93 (0.89-0.96)</td>
<td>0.96</td>
</tr>
<tr>
<td>2012-2014</td>
<td>96</td>
<td>96</td>
<td>4</td>
<td>4</td>
<td>96.0 (90.1-98.9)</td>
<td>96.0 (90.1-98.9)</td>
<td>96.0 (90.1-98.9)</td>
<td>96.0 (90.1-98.9)</td>
<td>0.92 (0.87-0.97)</td>
<td>0.97</td>
</tr>
<tr>
<td>2017-2019</td>
<td>96</td>
<td>97</td>
<td>4</td>
<td>3</td>
<td>97.0 (91.4-99.4)</td>
<td>96.0 (90.2-98.9)</td>
<td>96.0 (90.1-98.9)</td>
<td>97.0 (91.5-99.4)</td>
<td>0.93 (0.88-0.98)</td>
<td>0.96</td>
</tr>
</tbody>
</table>

**aTP**: true positive.  
**bTN**: true negative.  
**cFP**: false positive.  
**dFN**: false negative.  
**ePPV**: positive predictive value.  
**fNPV**: negative predictive value.  
**gAUC**: area under the receiver operating characteristic curve.  
**hICD-9**: International Classification of Diseases, Ninth Revision.  
**iICD-10**: International Classification of Diseases, Tenth Revision.

Quality Assurance

During the training process, 8 charts were independently reviewed by two chart abstractors. Their assessments of medication use for PPD agreed for all 8 records (100%), while the assessment of a diagnostic finding agreed for 7 (88%). After training was complete, another 8 records were independently reviewed. All 8 (100%) agreed in their findings for both medications and diagnoses.

Discussion

Principal Findings

This validation study demonstrated the potential to improve the accuracy of PPD case identification from an EHR when using diagnosis codes in conjunction with pharmacy records. The combination of clinical codes and prescription pharmacy records yielded much greater sensitivity and NPV, with no notable loss in specificity or PPV, compared with using either the diagnosis codes or pharmacy records alone. Using either record alone would result in significant undercounting, each missing about one-third of those with PPD, compared to the 95% identified using both together. Furthermore, we observed no significant difference in the ICD-9 and ICD-10 codes in terms of ascertaining PPD cases.

We found that electronic records of PPD diagnosis were not a reliable indicator of PPD diagnostic findings identified through chart review, relative to pharmacy records. Pharmacy records have both a sensitivity and specificity much higher than that seen for diagnosis codes.

The quality of data extracted from EHRs for pharmacoepidemiologic research has been proven to be valuable. Although using clinical diagnosis codes for perinatal epidemiology studies has limitations, the use of KPSC’s comprehensive pharmacy use records enhances the identification of PPD cases (sensitivity 98.3%, specificity 95.0%, PPV 93.7%, and NPV 95.0%).

While switching from ICD-9 to ICD-10 coding created some complexity, we did not see a significant difference in the accuracy of the electronic diagnosis records between the ICD-9 and ICD-10 coding eras. This is reassuring, as studies would not need to be limited to one era or the other for the sake of accuracy. Additionally, the prevalence of PPD identified in both periods is essentially the same, suggesting that both ICD-9 and ICD-10 coding systems identify patients with PPD at the same...
rate, negating any need to adjust prevalence estimates to account for the difference.

Accurate characterization of those with PPD is crucial to performing valid research on this condition. Many researchers rely on electronic records due to a lack of access to detailed patient histories or a lack of time to review these records. Our study suggests that researchers can accurately identify PPD from EHRs using both diagnosis and pharmacy records.

Comparison to Prior Work
Prior research validating diagnosis codes for identifying general depression found the PPV to be similar to that seen in our study (89.7% for ICD-9 and 89.5% for ICD-10), but these were not specific to the postpartum period [18]. These findings highlight the continuing debate regarding the use of diagnosis codes alone for epidemiological studies. Our study concurs with prior findings that the sensitivity and specificity of case ascertainment can be improved by concurrently using both diagnosis and pharmacy records [13]. Therefore, researchers should not rely exclusively on either diagnostic codes or pharmacy records for PPD case ascertainment.

Strengths and Limitations
There are some potential limitations to this study. First, while the KPSC EHR is comprehensive, it may not capture care received outside the system if it is not submitted for reimbursement. Specifically, members may receive mental health counseling from non-KPSC providers, and a PPD diagnosis made in that setting may not be entered into the KPSC medical record, resulting in a potentially missed PPD diagnosis and an underestimate of the sensitivity of diagnosis coding. However, these diagnoses may still be identified during regular clinical care within KPSC, hence limiting the number of potentially missed diagnoses.

Second, misclassification is also possible as variables were ascertained from clinical diagnosis codes and pharmacy record notes. In addition, there is the potential for misclassification of PPD within the data sources if women are unaware of the condition, do not seek medical care, or the diagnosis or treatment is not recorded in the clinical notes. Any completely undocumented cases would result in an underestimate of PPD in the population, though its potential effect on our validation is unknown. Finally, due to the small number of records reviewed in some groups, we were not able to look for differences in medical record accuracy within subsets of the population, including by age and race/ethnicity. If differences are present, this will limit the generalizability of these findings to other populations with different demographics.

Strengths of this study include the comprehensive medical record and chart review conducted to identify PPD in this patient population. The training and validation of the chart review process helped to ensure that the gold standard PPD identification was accurate.

Conclusions
This validation study of PPD that was carried out in a large integrated health care system in Southern California has demonstrated that PPD data ascertainment based on a combination of diagnosis codes and prescription medication records from the EHR is highly accurate for pharmacoepidemiologic studies. Neither diagnosis codes alone nor prescription records alone are sufficient to capture PPD cases.

Acknowledgments
Funding for this research was provided by National Institute of Health grant R01 ES030353-01 to DG (Kaiser Permanente Southern California) and JW (University of California, Irvine). The opinions expressed are solely the responsibility of the authors and do not necessarily reflect the official views of the funding agency. The Air Pollution and Pregnancy Complications in Complex Urban Environments (APPCUE) study team would like to thank Kaiser Permanente members who contributed electronic health information to this study.

Data Availability
Most of the data that support the findings of this study are available on request from the corresponding author. The complete data set is not publicly available due to privacy, institutional approval, and/or ethical restrictions. Study data come from patient electronic health records and birth certificates from the state of California. Data from patient health records cannot be shared without signed confidentiality agreements. Some of the data that support the findings of this study are available from the state of California. Restrictions apply to the availability of these data, which were used under license and approval for this study. Data can be made available by the authors provided that all required approvals are obtained from the departments in the state that oversees the use of state vital records data. To obtain California birth certificate data, researchers can email cphs@chhs.ca.gov or visit their website [21]. Requests for data may be sent to JS (Jeff.M.Slezak@kp.org) and DG (Darios.T.Getahun@kp.org).

Conflicts of Interest
The Air Pollution and Pregnancy Complications in Complex Urban Environments (APPCUE) study team led the design of the study and interpretation of the results. NK, CA, and JCC have no competing interests. JS receives research support from the National Institutes of Health (NIH), Pfizer Inc, Dynavax Technologies, and ALK. DG receives research support from NIH, National Institute of Environmental Health Sciences (NIEHS), Department of Health and Human Services, National Institute of Child Health and Human Development, Patient-Centered Outcomes Research Institute, Garfield Memorial Fund, Bayer AG, and...
Hologic, Inc. JW receives research support from NIEHS, the California Air Resources Board, and the Health Effects of Air Pollution Foundation. Kaiser Permanente Southern California (KPSC) led the design of the study and interpretation of the results in collaboration with study team members from the University of California, Irvine (UCI) and the University of Southern California (USC). JS conducted the analyses, which were reviewed by study team members from KPSC, UCI, and USC.

References


### Abbreviations

- **APPCUE**: Air Pollution and Pregnancy Complications in Complex Urban Environments
- **EHR**: electronic health record
- **ICD-9**: International Classification of Diseases, Ninth Revision
- **ICD-10**: International Classification of Diseases, Tenth Revision
- **KPSC**: Kaiser Permanente Southern California
- **NPV**: negative predictive value
- **PPD**: postpartum depression
- **PPV**: positive predictive value
- **STARD**: Standards for Reporting Diagnostic Accuracy Studies

---

© Jeff Slezak, David Sacks, Vicki Chiu, Chantal Avila, Nehaa Khadka, Jiu-Chiuan Chen, Jun Wu, Darios Getahun. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 01.03.2023. 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.
Perspectives on Challenges and Opportunities for Interoperability: Findings From Key Informant Interviews With Stakeholders in Ohio

Abstract

Background: Interoperability—the exchange and integration of data across the health care system—remains a challenge despite ongoing policy efforts aimed at promoting interoperability.

Objective: This study aimed to identify current challenges and opportunities to advancing interoperability across stakeholders.

Methods: Primary data were collected through qualitative, semistructured interviews with stakeholders (n=24) in Ohio from July to October 2021. Interviewees were sampled using a stratified purposive sample of key informants from 4 representative groups as follows: acute care and children’s hospital leaders, primary care providers, behavioral health providers, and regional health information exchange networks. Interviews focused on key informant perspectives on electronic health record implementation, the alignment of public policy with organizational strategy, interoperability implementation challenges, and opportunities for health information technology. The interviews were transcribed verbatim followed by rigorous qualitative analysis using directed content analysis.

Results: The findings illuminate themes related to challenges and opportunities for interoperability that align with technological (ie, implementation challenges, mismatches in interoperability capabilities across stakeholders, and opportunities to leverage new technology and integrate social determinants of health data), organizational (ie, facilitators of interoperability and strategic alignment of participation in value-based payment programs with interoperability), and environmental (ie, policy) domains.

Conclusions: Interoperability, although technically feasible for most providers, remains challenging for technological, organizational, and environmental reasons. Our findings suggest that the incorporation of end user considerations into health information technology development, implementation, policy, and standard deployment may support interoperability advancement.

(JMIR Med Inform 2023;11:e43848) doi:10.2196/43848

KEYWORDS

interoperability; health information exchange; health information technology; electronic health record; usability
Introduction

Background

Starting with the Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009, the United States has invested over US$ 36 billion to promote interoperability—the ability of 2 or more systems to exchange and use information [1-3]—through health information exchange (HIE) networks and electronic health records (EHRs). HITECH promoted the adoption and implementation of certified EHRs by providing financial incentives through its “meaningful use” programs and funded grants that helped establish regional HIE networks [4]. Some of these financial incentive programs, such as the state Medicaid Provider Incentive Program (MPIP) were phased out in 2021. However, subsequent legislation such as the Patient Protection and Affordable Care Act reinforced this financing by advancing payment and care delivery models that use risk-based contracts to incentivize quality of care and patient outcomes, such as accountable care organizations, value-based care [5], and patient-centered medical homes [6], which stand to benefit from enhanced electronic data exchange. Other population health policy programs, including Comprehensive Primary Care (CPC) [7], CPC-Plus [8], and Primary Care First [9], further rely on robust data exchange among regional providers. Ongoing federal policy development has continued and enhanced support for meaningful use programs through the Medicare Access and Children’s Health Insurance Program Reauthorization Act of 2015, leading to the Promoting Interoperability Program in 2018.

Recent reports show that 96% of acute care hospitals and 80% of primary care providers (PCPs) have implemented certified EHRs with interoperability capabilities [10,11]. However, actual use of data from an HIE network in clinical care encounters remains low [12]. Researchers have identified several barriers to the use of HIE networks, including poor user interfaces and lack of leadership support [13-15]. Although interoperable health information technology (HIT) is theorized to address these barriers, it continues to elude the health care system [16,17]. For instance, a recent study found that only 45% of all US hospitals engaged in the 4 core elements of interoperability, such as the capability of different EHR systems to find, send, receive, and use or integrate clinical information with one another [16]. Further, rural and smaller hospitals, ambulatory practices, as well as those ineligible for meaningful use incentives (eg, rehabilitation, long-term care, or behavioral health providers [BHPs]) lag behind large, integrated systems in adopting interoperability [18,19].

To promote interoperability, the 21st Century Cures Act (21CCA; 21CCA 2016) mandated the sharing of certain data elements, placed restrictions on information blocking, and promoted the use of application programming interfaces (APIs; eg, Fast Healthcare Interoperability Resources [FHIR]). The 21CCA also established the Trusted Exchange Framework Common Agreement (TEFCA) that provides an infrastructure model and governing approach for HIE networks [20]. However, in the sixth national survey of HIEs, 56% (84/151) of the regional HIEs planned to participate in TEFCA [21], which points to ongoing challenges that continue to limit data exchange. HIEs that intended to participate in TEFCA already had connections with HIEs in other states and participated in at least one national network. Therefore, the decision to participate in TEFCA may be determined by the alignment of HIE processes with existing data sharing rather than convincing HIEs of the benefits of participation [21].

Purpose

In alignment with policy efforts geared toward the promotion of interoperability, this study aimed to explore provider perspectives on the current state of interoperability challenges. Given the confluence of timing around the conclusion of state MPIPs and the ramp up of 21CCA and TEFCA, questions remain about the progress and remaining challenges related to achieving an interoperable health system. Moreover, previous research on interoperability typically focuses on a single perspective rather than those of multiple stakeholders. This multistakeholder lens is particularly relevant to consider when examining interoperability, as a key goal of this innovation is to connect disparate parts of the health system. Specifically, we conducted semistructured interviews with a stratified sample of key informants, including providers and individuals in leadership positions representing diverse organizations in Ohio, to identify barriers and facilitators to interoperability. The study’s findings add to the body of knowledge about interoperability and may contribute to the efforts of state agencies and federal policy makers, such as the Office of the National Coordinator for Health Information Technology and the Centers for Medicare and Medicaid Services, to advance interoperability. For instance, our findings may provide evidence that supports alignment between the Health Insurance Portability and Accountability Act (HIPAA) with HITTECH and 21CCA. Finally, technology vendors may benefit from increased understanding of the end user perspective of their applications to develop user-friendly software.

Methods

Study Setting and Design

We used a cross-sectional qualitative design to solicit multistakeholder perspectives on the state of interoperability in Ohio. Ohio provided financial incentives for the adoption of interoperable EHRs for eligible professionals and hospitals through the MPIP. Eligible professionals included those with an active Ohio Medicaid Provider Agreement, such as physicians, optometrists, dentists, certified nurse-midwives, nurse practitioners, and physician assistants practicing in a federally qualified health center or a rural health center led by a physician assistant. Eligible hospitals were also required to have an active Ohio Medicaid Provider Agreement and include acute care hospitals, critical access hospitals, cancer hospitals, and children’s hospitals. Ineligible providers included most behavioral and mental health, long-term care, and home health providers. MPIP operated through 7 incentive cycles, with the final cycle occurring in 2021. On the basis of the 2021 Ohio Medicaid Electronic Health Records Survey for Practices and Hospitals [22], 23.87% (5593/23,435) of the eligible providers and hospitals had received at least one MPIP payment, 96.06%
(5280/5593) of the MPIP recipient providers and hospitals had adopted and used an EHR, and 90.73% (16,188/17,842) of those who did not receive an MPIP payment had adopted and used an EHR. Among the 735 ineligible providers, 72.1% (530/735) reported adopting and using an EHR. Epic Systems Corporation is the most prevalent EHR vendor in the state, with 36.80% (2058/5593) of MPIP recipient providers and hospitals and 56.56% (10,092/17,842) of non-MPIP recipient providers and hospitals using an Epic EHR. However, though Epic remained the EHR of choice for group practices (across multiple or single sites) and hospitals, individual practices were more likely to choose other EHR vendors (eg, NextGen and eClinicalWorks). Among the ineligible providers, there was substantial variation in EHR vendors, and CareLogic was the single most prevalent EHR vendor, with 7.8% (57/735) of providers adopting it. Presently, Ohio has 2 large-scale regional HIEs that facilitate electronic exchange of patient data. Almost 31.64% (7416/23,435) of MPIP eligible providers and hospitals had existing arrangements with the regional HIEs to share electronic patient-level clinical data, whereas 21.5% (158/735) of MPIP ineligible providers participated in the regional HIEs.

Data collection and analysis were guided by the technology-organization-environment (TOE) framework [23]. TOE is an organization-level theory that has been applied to explain how the 3 interacting contextual domains (ie, technology-organization-environment) influence a health care organization’s technology-related decision-making [24].

Ethics Approval
This study was considered to have minimal risk and was approved by the Ohio State University institutional review board (2021B0378).

Sample Selection
We used a stratified purposive sampling approach to gather diverse perspectives on interoperability based on four representative groups: (1) acute care and children’s hospital leaders, (2) PCPs, (3) BHPs (ie, providers or organizations that provide care for mental health, substance use disorders, stress-related physical symptoms, and life stressors and crises), and (4) regional HIEs (ie, organizations that facilitate information exchange within a network of facilities within a geographic boundary). In addition, both rural and urban subsamples within each of the 3 provider groups were interviewed to ensure geographic representation. We focused on key informants with first-hand perspectives on HIT adoption and its future directions. These key informants were administrative leaders with clinical and nonclinical backgrounds within organizations with decision-making capacity regarding HIT (eg, chief medical information officers, executive directors, chief executive officers, and strategy officers). The study leads received a list of potential key informants from the Ohio Department of Medicaid. Emails were sent to gauge interest in participation; key informants who agreed to participate (20 of 38 organizations invited) were then interviewed. To eliminate any potential conflict of interest, the Ohio Department of Medicaid was not notified of who agreed or refused to participate, and did not participate in any interviews.

Data Collection
We conducted semistructured interviews with key informants from July to October 2021. The interview guide was developed to ask about EHR implementation in general, the alignment of public policy with organizational strategy, interoperability implementation challenges, and opportunities to improve HIT effectiveness. The interview guide was piloted with administrative leaders (n=2; eg, chief information officer) at an urban hospital in Ohio. This process yielded two versions of the semistructured interview guide for (1) providers (ie, those from hospitals, primary care, and behavioral health) and (2) representatives of HIEs (Multimedia Appendix 1). All interviews were conducted remotely via Zoom and audio recorded.

Data Analysis
Interviews were transcribed verbatim and analyzed using directed content analysis—an approach that begins with the a priori codes from the TOE framework yet is permissive of emergent themes [25]. The coding team (DMW, WLT, and LP) met weekly to discuss the interviews and preliminary findings throughout the data collection phase. A preliminary codebook was developed by reviewing transcripts and identifying broad themes that emerged from the interview transcripts to organize the data into the 3 TOE domains. Next, to build consensus on the coding guide, the analysis team collectively reviewed common transcripts (n=3) to compare results, refine the codebook, and reconcile any coding discrepancies. The remaining transcripts were divided by organization type for coding. The team continued to compare findings throughout the analysis phase to achieve thematic saturation. The trustworthiness of our findings is ensured by our rigorous and iterative approach to analysis [26]. NVivo software (version 12) was used to support data coding and analysis.

Results
Overview
Overall, 24 key informants were interviewed representing 20 distinct organizations: 5 hospitals, 7 PCPs, 6 BHPs, and 2 HIEs. Of the 20 organizations, 15 (75%) were located in urban areas (Table 1).

Below, we report on the key themes and subthemes of the analysis organized by the domains of the TOE framework, including comparing and contrasting those that cut across organizational types as well as those specific to each individual organization.
Table 1. Key informant interview sample characteristics\textsuperscript{a}.

<table>
<thead>
<tr>
<th>Organization type</th>
<th>Interviewees (n=24), n (%)</th>
<th>Organizations (n=20), n (%)</th>
<th>Urban area organizations (n=15), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute care and children’s hospital</td>
<td>5 (21)</td>
<td>5 (25)</td>
<td>4 (27)</td>
</tr>
<tr>
<td>Primary care provider</td>
<td>7 (29)</td>
<td>7 (35)</td>
<td>5 (33)</td>
</tr>
<tr>
<td>Behavioral health provider</td>
<td>10 (42)</td>
<td>6 (30)</td>
<td>4 (27)</td>
</tr>
<tr>
<td>Regional health information exchange</td>
<td>2 (8)</td>
<td>2 (10)</td>
<td>2 (13)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Some interviews (n=3) included multiple key informants.

**Technology Domain**

**Overview**

Three themes within the technology domain that related to the usability and technological aspects of interoperability were identified: (1) implementation challenges, (2) interoperability capabilities, and (3) opportunities (Table 2).
<table>
<thead>
<tr>
<th>Theme and primary subtheme</th>
<th>Representative quote</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Implementation challenges</strong></td>
<td>“There’s been an explosion of health IT [information technology] applications, vendors and products over the past decade, many of whom overlap and functionality intersect in ways that don’t really allow for great interoperability, so just the challenge of sort of how do you meet all the needs, using all the various products out there, and still have a cohesive, reliable, safe experience is a challenge.” [Hospital representative #15]</td>
</tr>
<tr>
<td>Maintaining growing number of applications</td>
<td>“Our community health centers are on a lot of different EHR [electronic health record] platforms, and those platforms don’t talk to each other, and so the interoperability that we all desire is still not really there. So, we...are utilizing a health population, a population health tool that can sit over any EHR [electronic health record] platform, and that is allowing us to get some of the data aggregated, in spite of the lack of interoperability and communication between different EHRs.” [Primary care provider #22]</td>
</tr>
<tr>
<td>Integrating diverse sources of data into unified medical record</td>
<td>“Every hospital has this issue, is that there’s not really a good way to leverage the data being collected by the state vital statistic[s] for our use...We really get no automatic notification that a patient has died. They die elsewhere and the state help out because there’s a death certificate somewhere, but our HIM [health information management] department is sort of stuck almost to the point of reading obituaries trying to figure out what patients to mark...It’s really hard, from our perspective to reach out to a family with an appointment reminder about a patient who died, and it’s just, not only is it horrible customer service and patient experience or family experience to do that...I think more connection points with actually the state for some of this basic stuff like birth records and death records and marriage certificates where names are changing and that information’s sitting there, but it seems to be behind this kind of either bureaucratic or policy firewall.” [Hospital representative #15]</td>
</tr>
<tr>
<td>Exchanging data across the continuum of care</td>
<td>“In our long-term care systems, it would be nice for us to be able to exchange information about those patients, especially with medications, make sure their medications are set up and they know everything that the patient’s on. One thing, too, is provide our providers with information, a little bit more timely from the long-term care facilities to keep them from being readmitted or admitted to the hospital. So that’s something we’d like to be able to do. Another thing would be the merging of medical records for both mental health and their regular healthcare.” [Primary care provider #6]</td>
</tr>
<tr>
<td>Reliance on regional HIEs4 for data exchange</td>
<td>“I see a lot of potential for us to be able to really be part of that health information exchange network, and use our EHR [electronic health record] system to do a lot of that in the background, as opposed to currently what we are doing is we have access to [regional HIE] portal to get the community health record, the information, but that takes staff time. You know, a lot of training, and so it’s not really fully integrated into our EHR [electronic health record], and it’s not fully integrated into our clinical practice processes.” [Behavioral health provider #17]</td>
</tr>
<tr>
<td><strong>Interoperability capabilities</strong></td>
<td>“Every hospital has this issue, is that there’s not really a good way to leverage the data being collected by the state vital statistic[s] for our use...We really get no automatic notification that a patient has died. They die elsewhere and the state help out because there’s a death certificate somewhere, but our HIM [health information management] department is sort of stuck almost to the point of reading obituaries trying to figure out what patients to mark...It’s really hard, from our perspective to reach out to a family with an appointment reminder about a patient who died, and it’s just, not only is it horrible customer service and patient experience or family experience to do that...I think more connection points with actually the state for some of this basic stuff like birth records and death records and marriage certificates where names are changing and that information’s sitting there, but it seems to be behind this kind of either bureaucratic or policy firewall.” [Hospital representative #15]</td>
</tr>
<tr>
<td>Connecting to state-hosted registries and databases (ie, state immunization registry)</td>
<td>“In our long-term care systems, it would be nice for us to be able to exchange information about those patients, especially with medications, make sure their medications are set up and they know everything that the patient’s on. One thing, too, is provide our providers with information, a little bit more timely from the long-term care facilities to keep them from being readmitted or admitted to the hospital. So that’s something we’d like to be able to do. Another thing would be the merging of medical records for both mental health and their regular healthcare.” [Primary care provider #6]</td>
</tr>
<tr>
<td><strong>Opportunities</strong></td>
<td>“That’s the beauty of it. So, I...mentioned computer visionb, so what that’s compared to if you wanted to do that in the past, you would have had to have a pretty labor-intensive interface between the platform and each individual practice’s EHR [electronic health record], right. And that’s a huge level of effort that most people can’t really get to, and that’s why the computer vision piece of that really makes sense. You don’t have to have that, well, it’s still fancy, it’s fancy in a different way, you don’t need a fancy interface, you’re using the computer vision to match the patients.” [Hospital representative #14]</td>
</tr>
<tr>
<td>Using new population health software to improve interoperability</td>
<td>“Back to the social determinants...the opportunity to connect to external things like Aunt Bertha or NowPow or HealthyI, one of those products that helps kind of do closed loop referral, and whether we’ll do that within Epic...I think those...are helping us kind of reach our goals around reducing health disparities.” [Primary care provider #14]</td>
</tr>
<tr>
<td>Integrating care coordination programs to improve social needs referrals</td>
<td>“We are just implementing our FHIR [Fast Healthcare Interoperability Resources] layer right now. What I will tell you is while FHIR [Fast Healthcare Interoperability Resources] is definitely a direction of the future, it is not broadly deployed in the marketplace and not broadly deployed in the workflow or business applications to great extent. But it is, definitely will be an important factor as we move into the future. But it also will not be the silver bullet that everybody’s hoping it was going to be.” [HIE representative #8]</td>
</tr>
<tr>
<td>Using FHIR-based applications to advance interoperability</td>
<td>“I see a lot of potential for us to be able to really be part of that health information exchange network, and use our EHR [electronic health record] system to do a lot of that in the background, as opposed to currently what we are doing is we have access to [regional HIE] portal to get the community health record, the information, but that takes staff time. You know, a lot of training, and so it’s not really fully integrated into our EHR [electronic health record], and it’s not fully integrated into our clinical practice processes.” [Behavioral health provider #17]</td>
</tr>
</tbody>
</table>

---

aHIE: health information exchange.
bComputer vision is a field of technology that enables devices such as smart cameras to acquire, process, analyze, and interpret text, images, and videos.
cFHIR: Fast Healthcare Interoperability Resources.
Implementation Challenges

For hospitals, a primary concern focused on the growing number and proliferation of application types (eg, EHR, personal health records, HIE, and population health platforms) that contain health information and can be used in clinical encounters. These applications were viewed as difficult and costly to maintain. Similarly, these applications do not use a common data structure and storage format, resulting in too many places for data to be located and for clinicians to search for useful information. The lack of interoperability necessitates that providers leave the EHR to access other applications. As one hospital representative stated as follows:

...It’s one of those things where now you’ve got to subscribe to it [eg, a regional HIE], and it’s another place [eg, application] for you to go to look for more information [eg, patient clinical data]...There’s too many places for data to land and get sent. People just stop looking. [Hospital representative #10]

HIE leadership echoed this perspective, adding that a lack of interoperability was an impediment to creating a complete medical record. PCPs noted similar issues related to the number and type of applications needed to overcome gaps in interoperability, such as population health tools (eg, Innovaccer). In contrast, for BHPs, interoperability across applications was not discussed; instead, remarks focused on off-the-shelf EHR systems being misaligned to the specific requirements for BHPs, such as lacking additional protections for substance abuse data.

Interoperability Capabilities

Hospitals described advanced interoperability functionality and attributed their advanced capabilities to their EHR vendor rather than the regional HIEs. The regional HIEs were helpful for exchanging continuity of care documents [27] but did not facilitate the integration of information across EHRs. Typically, hospitals only expressed limitations with interoperability functionality as being a function of data recipients’ capacities. For instance, a major concern for hospitals was connecting and being able to integrate EHR data with state-hosted information systems such as the state immunization registry, vital statistics, death certificates, or birth records.

Hospitals and PCPs both noted limitations of exchanging data across the continuum of care, such as with long-term care and BHPs. PCPs additionally noted interoperability challenges with hospitals in their own health systems even when they all used the same EHR vendor, as some vendors are not capable of data exchange in different instances of the same EHR.

BHPs described much more basic data exchange capabilities relative to hospitals and PCPs. For instance, they mentioned that their data exchange is primarily focused on billing. In contrast to hospitals, BHPs mentioned a greater reliance on the regional HIEs for access to health records from other providers and event notifications. The regional HIEs echoed this relationship and discussed how BHPs lag in their interoperability capabilities.

Opportunities

Hospitals described opportunities related to new population health platforms that may be able to improve interoperability without the costs associated with interfacing with different EHR systems or HIEs. Both hospitals and PCPs felt that increasing analytic rigor and predictive modeling with artificial intelligence and machine learning would support their population health efforts. API-based data exchange was generally identified as an opportunity to promote interoperability. Hospitals mentioned that the use of applications with this technology will benefit remote patient monitoring and chronic care management. API-based data sharing was also identified as an opportunity to improve interoperability in behavioral health despite the focus on more basic technological opportunities, such as increasing EHR functionality.

The regional HIEs also shared these perspectives on API-based data sharing and were particularly attentive to FHIR APIs for the development and implementation of EHR-integrated applications to advance interoperability. FHIR-based applications were also expected to improve patient access to their medical records. Although most hospitals and PCPs offered patient access to their medical records through patient portals at the time of our study, these were not unified across providers. FHIR-based applications could potentially enable broader patient access through a unified patient portal that collects information from disparate providers. However, the regional HIEs tempered this enthusiasm, recognizing that FHIR-based applications are not currently broadly deployed in existing technology builds.

Hospitals, PCPs, and HIEs also discussed using care coordination programs (eg, Aunt Bertha) that can track referrals to social service agencies that address social determinants of health (SDoH). These providers described collecting SDoH data in discrete data fields but also noted that the lack of standardization of SDoH data fields remains problematic and results in questionable data quality and limited ability to combine data across sources. Efforts to develop SDoH data standards, such as the Gravity Project [28], were raised as opportunities to improve the interoperability of SDoH data.

Organization Domain

Overview

We identified two themes within the organizational domain that affected interoperability: (1) facilitators and (2) strategic alignment (Table 3).
Table 3. Themes and primary subthemes in the organization domain.

<table>
<thead>
<tr>
<th>Themes and primary subthemes</th>
<th>Representative quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relationship with EHR vendor</td>
<td>“They [EHR vendor] are as interested in making sure that interoperability happens as what we are and so when we start to look at different interfaces that we need to have built, whether it’s to another vendor, like Cerner or eClinicalWorks, then Epic builds that interface, so that it makes it easier on both ends, to make that connection.” [Hospital representative #3]</td>
</tr>
<tr>
<td>Data standards adoption</td>
<td>“I think standards, the general and the meaningful use did quite a bit of pushing this sort of embracing of standards around things like nomenclatures, terminologies that allow for transmission of information. Prior to this we’re pretty much stuck with HL7 [Health Level 7] and custom specifications, but now with just CPT [current procedural technology] or ICD [International Classification of Disease] but, between SNOMED [Systematized Nomenclature of Medicine] and LOINC [Logical Observation Identifiers, Names and Codes] and RxNorm and CVX [vaccine administered] codes for immunizations and it’s gotten a lot better.” [Hospital representative #15]</td>
</tr>
<tr>
<td>Senior leadership support</td>
<td>“The biggest thing [to support data gathering and integration] is the addition of scribes. Adding on that expense of additional manpower to do that data entry for the providers to get them to where they’re comfortable with...what is pertinent to that visit. It may be an ER [emergency room] visit, is it a recent CT [computerized tomography scan], so that way staff isn’t trying to print the last X-amount of things and really all they wanted was one thing, so trying to streamline that to get the physicians the information they need, but having a team around them, to help them put the information back in and alleviate that work from them.” [Primary care provider #16]</td>
</tr>
</tbody>
</table>

**Strategic alignment**

| Payment program participation impacts technology purchasing decisions | “We are in CPC [Comprehensive Primary Care]-Ohio and CPC+ [Comprehensive Primary Care Plus]. We’re also doing Primary Care First and I have a number of value-based commercial contracts that we deal with as well. We are not an ACO [accountable care organization]. Our new software with our population health software that’s been added to our regular EMR [electronic medical record] should help greatly with that and that’s the reason we did it is because we’re getting into more value-based contracts. I think that’s something that will improve our outcomes for our patients and improve our financial return as well.” [Primary care provider #6] |
| Using interoperability to develop cross-sector alignment and stakeholder consensus | “I’d love to see us in the state of Ohio come together at kind of a developer’s conference or something...How can we come together and figure out how to make this work better for Ohio? And I know that sounds really altruistic because everyone’s trying to run a business and all that, but it just seems like there’s so much overlap and you think—I’ll use the example with us: I’m sitting on a mountain of data. Right, and so it just drives me bonkers to hear of a small mom-and-pop startup software company, who has to go out and buy a big giant data warehouse, you know, big giant SQL server and pay licenses and then they contact all the hospitals in the doctor’s office and say, give me all your data, right. And we’re just duplicating these silos. Not too long ago, I was giving a presentation, I said, how many of your hospital systems have a population health strategy? And, of course, 100% of them raised their hand, right. And I said so you’ve invested millions into giant data warehouses to support population health, you know. Right? And they all go, yeah. Like, well so did all the HIEs[b] [health information exchanges], so did the state of Ohio, you know. ODH [Ohio Department of Health] is trying to, like we’re all we’re all spending—Microsoft and Oracle and all those guys are making money hand over fist. Because we can’t get ourselves organized.” [HIE representative #4] |
| Different perspectives on the value of interoperability | “They didn’t see how interoperability would help them take care of their patients any better. And our team even said, ‘Well, you can get lab results like instantaneously.’ ‘Yeah, you know, but I get them a day late. It’s fine.’ ...I think that’s the other challenge is really, is that one example, or is that some X percent of providers in the state of Ohio who don’t see value in that interoperability.” [HIE representative #4] |

**Facilitators**

A consistent subtheme related to the importance of relationships with the EHR vendor to support interoperability emerged. Hospitals placed considerable value on their relationship with Epic and perceived the high concentration of state-wide Epic institutions and the integration between HIEs and Epic as a benefit. Hospitals also mentioned the push for data standards through meaningful use as facilitating interoperability.

PCPs described the important role of senior leadership, who can designate sufficient human resources for tasks that typically increase clinical workload, such as data gathering before appointments. These staff resources can help physicians access and use information from other sources.

**Strategic Alignment**

A critical driver for interoperability among hospitals was their participation in value-based payment programs and population health initiatives. To facilitate the data exchange and integration...
for care coordination, billing, and reporting required to support these programs, hospitals reported purchasing EHRs for network partners that lacked these advanced capabilities, thus promoting interoperability among their clinical partners. Likewise, hospitals preferred to develop their own in-house population health analytics platforms rather than outsource this function to HIEs. PCPs also felt that interoperability is central to achieving strategic goals such as population health management, which is integral to participation in alternative payment models (ie, accountable care organizations, CPC-Ohio, CPC-plus, and Primary Care First). They noted the benefits of event notifications made possible by admission, discharge, and transfer feeds that allow PCPs to be notified when their patients have received care in other settings and follow-up accordingly with them to meet their needs.

Conversely, BHPs indicated that they were not participating in value-based purchasing programs to the same degree. However, similar to hospitals and PCPs, their technology investment decisions were guided by their participation in payment programs with sponsors (eg, Health Resources and Services Administration and Substance Abuse and Mental Health Services Administration) that require specific reports, although they may not necessarily aid in interoperability.

The regional HIEs viewed the development of their exchange networks as an opportunity to advocate for cross-sector alignment, particularly as it pertains to streamlining duplicate efforts toward population health management. They remarked that regional HIEs are in a unique position to negotiate partnerships that address the concerns of different stakeholders. Finally, the regional HIEs noted challenges related to some organizations, particularly BHPs, not viewing interoperability as valuable to their organization or aligning with their strategy.

**Environment Domain**

Within the environmental domain, a policy theme was identified that focused on how policy may hinder or facilitate interoperability (Table 4).

Hospitals felt that there was room for a greater policy push for managed care plans to initiate value-based payment contracts and distribute incentives to providers. Hospitals also felt the costs of interoperability, such as establishing admission, discharge, and transfer feeds or registry reporting, fall predominantly on hospitals, yet there are no corresponding changes to reimbursement. Similarly, BHPs also desired additional funding to support their adoption of advanced EHR systems.

All providers, including the regional HIEs, also noted the considerable impact of 42 Code of Federal Regulations (CFR) Part 2 (ie, Substance Abuse Confidentiality Regulation) as limiting interoperability and connectivity with BHPs. Despite the general recognition of the policy’s good intentions, it was viewed as negatively impacting clinical care. Some key informants felt that this policy was incongruent with new information blocking rules as part of the 21CCA. The regional HIEs also saw consequences of this policy with respect to responding to public health emergencies such as the opioid epidemic. BHPs, PCPs, and the regional HIEs all advocated for aligning 42 CFR Part 2 with the HIPAA to clarify what is protected and not shareable versus what can be shared for continuity of care.
The regional HIEs noted some specific concerns, given their experience in the COVID-19 response. First, they described a need for greater enforcement of information-blocking rules. The implications of information blocking were particularly notable during the pandemic when withheld information about addresses prevented HIE from tracking COVID-19 at the local level. Second, the regional HIEs felt that HIPAA guidance on reporting on geographic areas with less than 20,000 patients is both vague and confusing, limiting population health efforts.

Finally, the regional HIEs also mentioned that TEFCA helps to establish standards but did not include any information about FHIR-based standards. They noted that this update would be helpful to promote adoption of FHIR-based applications [30].
Discussion

Principal Findings

The near-ubiquitous adoption of certified EHRs over the past decade has resulted in the capture of vast amounts of data across the care continuum. Recent policy efforts such as the 21CCA aim to promote vendor-agnostic integration of external data into the EHRs of all provider types and address information blocking to mitigate challenges. However, questions persist about how well these efforts assist health care organizations in achieving interoperability. Our study examined perspectives from a variety of provider types in Ohio on the state of interoperability. We identified important barriers and facilitators to interoperability among hospitals, PCPs, BHPs, and regional HIEs.

Our findings related to implementation issues within the technology domain suggest that the proliferation of applications that address various use cases promotes capture of rich data. However, from an end user perspective, this approach can create inefficiencies because of excess information. User interfaces that do not embed multiple discrete applications within the EHR may ultimately create a fragmented medical record, which makes it harder to find relevant information. These types of information silos of patient data can potentially jeopardize patient safety, care quality, and organizational efficiency [31,32]. Fragmented or siloed information may also contribute to provider burnout [33]. Thus, our findings highlight a need for user-centric approaches in technology design and implementation to translate increased information access to use.

Providers did note that technological advances such as computer vision-based population health software (eg, Innovaccer) can help overcome these barriers by pulling data directly into the EHR without requiring back-end integration. Likewise, the growing support for API-based data sharing can further support interoperable information exchange between dissimilar EHR vendors. APIs can extend EHR capabilities [34], although this potential remains unrealized to date. Recent initiatives such as TEFCA that mandate sharing standardized sets of data and promote the use of FHIR APIs are expected to facilitate interoperability in the coming years [30]. Moreover, the API approach has the potential to facilitate population health analytics using machine learning techniques [35].

Within the technology domain, our findings reiterated a gap between hospitals and PCPs at one end of the spectrum and BHPs at the other with respect to their interoperability capabilities. This gap may be a result of most BHPs being ineligible for federal incentive programs [18]. This omission likely not only disincentivizes the adoption of advanced EHRs capable of interoperability among BHPs but also discourages investment and development by EHR vendors of tools designed to meet the needs of BHPs [36].

In addition, across the technology, organization, and environment domains and across provider types, a notable issue that emerged from the key informant interviews was the limitation on interoperability imposed by 42 CFR Part 2, a rule that restricts sharing substance use and behavioral health data. Our findings suggest that despite its intentions, 42 CFR Part 2 effectively operates to prevent, or severely limit, BHPs from participating in exchange. In addition to potential medical errors, this restriction may result in missing data from analytic data sets used by providers, insurers, or researchers [37]. Further, providers expressed concern that this rule may no longer be in alignment with 21CCA information-blocking rules. Moving forward, modifications to 42 CFR Part 2 may be necessary to support further interoperability. Consideration of end user needs and incorporating perspectives of BHPs is essential in any policy changes to carefully consider the need for privacy related to substance use and behavioral health data balanced against the benefits of interoperability [38]. To this effect, the United States Department of Health and Human Services issued a Notice of Proposed Rulemaking in November 2022 [39]. The Notice of Proposed Rulemaking proposes to permit the use and disclosure of Part 2 records based on a single prior signed consent, to expand prohibitions on the use and disclosure of Part 2 records in legal proceedings and to expand patient rights that align with the HIPAA Privacy Rule. The importance of this issue has recently become more visible because of concerns around the protections of reproductive health information following increased abortion restrictions because of the overturning of the Roe versus Wade supreme court case [40,41]. The revisions to Part 2, if enacted, would not only help BHPs engage in interoperability but would also provide greater protection for how sensitive health data can be used in legal proceedings.

Interestingly, hospitals attributed their interoperability capabilities more to their relationship with EHR vendors as opposed to the regional HIEs. This finding is likely a consequence of the strong foothold of Epic in Ohio. From an operational and governance standpoint, providers may face fewer barriers to participating in Epic’s Care Everywhere vendor-mediated HIE network. Conversely, participation with a regional HIE may require further effort to establish data exchange policies that require buy-in from multiple stakeholders. Vendor-mediated HIEs may create a divide in HIE engagement among providers with different vendors [42]. Indeed, as the HITECH funding period drew to a close, the number of state and regional HIEs declined, partly because of the mergers of regional HIEs, funding challenges, and competition from vendor-mediated HIEs [43].

Our findings related to strategic alignment may offer useful policy recommendations; provider participation in value-based payment programs plays a critical role in how providers are considering (or not considering) investments in interoperability. To the extent that most of a provider’s patients are beneficiaries of these programs, providers may expand the breadth of their interoperability functionality, such as participation in regional HIEs or use of population health platforms to meet the needs
of that particular patient population [44]. Key informants also described the use of social needs referral platforms; however, developing exchange with non-HIPAA–covered social service or community-based organizations can be challenging without properly aligned incentives [45]. In addition to clarifying HIPAA rules around exchange with noncovered entities, the expansion of value-based payment programs may leverage HIT investment. Further, regional HIEs may be well positioned to advocate for cross-sector strategic alignment.

The providers in our study reported capacity-related challenges in interoperability with public health agencies. Chronically underfunded public health systems impeded efficient and timely electronic information exchange during the COVID-19 pandemic [46]. In response, the Centers for Disease Control and Prevention launched the Data Modernization Initiative, resulting in changes to core data sources and facilitating access to electronic case reports and the COVID electronic laboratory reporting that makes test results available. Other initiatives such as TEFCA, through their emphasis on interoperability, are also expected to mitigate challenges, particularly from differing vocabulary standards. Other barriers to interoperability in public health arise from the complex legal and regulatory environment [47]. Even though 21CCA established a legal framework to address information blocking [48], the HIEs participating in our study reported information blocking that prevented tracking cases of COVID-19. Moving forward, it will be critical to monitor the impact of the Data Modernization Initiative and TEFCA on the interoperability of public health data.

Limitations
We purposely sampled from multiple stakeholders to gain a representative perspective on interoperability. The design focused on breadth across providers rather than depth within a specific provider type or within a single health care organization. Similarly, the study only included Ohio stakeholders, which minimizes variation in the policy environment and may limit generalizability.

The sampling approach was designed to include individuals with decision-making authority with respect to HIT. These perspectives may differ from those of other end users. Finally, owing to time and resource constraints, the interview guide may not have probed all issues relevant to interoperability but was purposefully open-ended to allow participants to discuss topics they deemed important.

Conclusions
Our findings suggest that despite the ubiquity of data and applications, seamless interoperability into a comprehensive medical record, both within and across providers, remains out of reach. Technological solutions offer promise to overcome these challenges. Likewise, the expansion of value-based payment programs can further incentivize interoperability. Although policy initiatives to expand interoperability existed, they were often misaligned to operational needs and may not be sufficient to overcome market forces. A policy focus toward embracing user-centric design to incorporate end user experience into HIT development may overcome barriers associated with achieving interoperability.

Acknowledgments
This project was supported by a contract from the Ohio Department of Medicaid. The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsor. The authors would like to thank Lauren Phelps for her assistance with project management.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Interview guides.
[DOCX File, 41 KB - medinform_v11i1e43848_app1.docx ]

References


The Effect of Implementation of Guideline Order Bundles Into a General Admission Order Set on Clinical Practice Guideline Adoption: Quasi-Experimental Study

Justine Mrosak¹, MD; Swaminathan Kandaswamy², PhD; Claire Stokes³,⁴, MD, MPH; David Roth⁵, MD, MSPH; Jenna Gorbaktin², MD; Ishaan Dave², MSPH; Scott Gillespie², MS, MSPH; Evan Orenstein²,⁴, MD

¹Hennepin Healthcare, Minneapolis, MN, United States
²Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, United States
³Division of Pediatric Hematology/Oncology, Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, United States
⁴Children's Healthcare of Atlanta, Atlanta, GA, United States
⁵Department of Pediatrics, University of Pittsburgh School of Medicine, Pittsburgh, PA, United States

Corresponding Author:
Justine Mrosak, MD
Hennepin Healthcare
701 Park Avenue
Minneapolis, MN, 55415
United States
Phone: 1 6128891839
Email: jmrosak@gmail.com

Abstract

Background: Clinical practice guidelines (CPGs) and associated order sets can help standardize patient care and lead to higher-value patient care. However, difficult access and poor usability of these order sets can result in lower use rates and reduce the CPGs’ impact on clinical outcomes. At our institution, we identified multiple CPGs for general pediatrics admissions where the appropriate order set was used in <50% of eligible encounters, leading to decreased adoption of CPG recommendations.

Objective: We aimed to determine how integrating disease-specific order groups into a common general admission order set influences adoption of CPG-specific order bundles for patients meeting CPG inclusion criteria admitted to the general pediatrics service.

Methods: We integrated order bundles for asthma, heavy menstrual bleeding, musculoskeletal infection, migraine, and pneumonia into a common general pediatrics order set. We compared pre- and postimplementation order bundle use rates for eligible encounters at both an intervention and nonintervention site for integrated CPGs. We also assessed order bundle adoption for nonintegrated CPGs, including bronchiolitis, acute gastroenteritis, and croup. In a post hoc analysis of encounters without order bundle use, we compared the pre- and postintervention frequency of diagnostic uncertainty at the time of admission.

Results: CPG order bundle use rates for incorporated CPGs increased by +9.8% (from 629/856, 73.5% to 405/486, 83.3%) at the intervention site and by +5.1% (896/1351, 66.3% to 509/713, 71.4%) at the nonintervention site. Order bundle adoption for nonintegrated CPGs decreased from 84% (536/638) to 68.5% (148/216), driven primarily by decreases in bronchiolitis order bundle adoption in the setting of the COVID-19 pandemic. Diagnostic uncertainty was more common in admissions without order bundle use, we compared the pre- and postintervention frequency of diagnostic uncertainty at the time of admission.

Conclusions: The integration of CPG-specific order bundles into a general admission order set improved overall CPG adoption. However, integrating only some CPGs may reduce adoption of order bundles for excluded CPGs. Diagnostic uncertainty at the time of admission is likely an underrecognized barrier to guideline adherence that is not addressed by an integrated admission order set.

(JMIR Med Inform 2023;11:e42736) doi:10.2196/42736

KEYWORDS

clinical practice guideline; user-centered design; clinical decision support; diagnostic uncertainty; diagnostic; decision support; CPG; clinical guideline; order bundle
It remains unknown what CDS designs best address these barriers and most improve CPG adherence.

Rationale/Specific Aims

Our purpose for this project was to provide higher-value care through improved adherence to evidence-based CPGs at our institution. In preliminary data described elsewhere [18], we found that common reasons among local frontline providers for not adopting CPG order sets in eligible populations included lack of awareness (32%) and forgetting to use the stand-alone CPG-specific order set (20%). We therefore implemented a new CDS system in our admission process in the form of embedded CPG order bundles integrated into the general pediatrics admission order set that were identical to CPG-specific order bundles in the existing stand-alone CPG-associated order sets. Our primary aim was to increase CPG-specific order bundle use for eligible patients admitted on the general pediatrics service by 20% from July 2019 to May 2021. The primary aim was determined by the project stakeholder team to likely be an achievable improvement based on initial data that many CPGs demonstrated an <50% adherence rate, as well as an improvement that would justify the anticipated effort to complete the project. Secondary aims included determining if there were differences in order bundle use between specific CPGs at the intervention site and comparing CPG order bundle adherence between the intervention site and another hospital within the same health system where the intervention was not implemented.

Methods

Context/Setting

This study was performed on the general pediatrics service in an academic urban children’s hospital within a 3-hospital, 638-bed pediatric health system serving the greater Atlanta, Georgia, area. Over 90 pediatric and family medicine residents rotate through the general pediatrics service each year and are overseen by 16 pediatric hospital medicine faculty members at the intervention hospital. Our institution uses an official electronic health record supplied by Epic Systems. Currently, there are over 20 general pediatric-specific CPGs customized to local workflows that are available for use through the institutional intranet. Prior to the intervention, 15 of the CPGs had their own stand-alone order set to facilitate adherence.

The intervention was implemented at 1 of 3 freestanding children’s hospitals; this was an academic tertiary care center staffed primarily by resident teams with pediatric hospital medicine attendings. The nonintervention site was a hybrid community-academic hospital primarily staffed directly by pediatric hospital medicine attending within the same health system.

Intervention

Planning the Intervention

Stakeholders included representatives from the Pediatric Hospital Medicine service, the Department of Clinical Effectiveness, and the Department of Quality and Safety, as well as a clinical informaticist, a human factors engineer, a
medical student, and a quality improvement methodology expert. This team met formally multiple times in the planning stages of the project and while formal problem analysis was underway before the intervention.

Preintervention problem analysis has been previously described [18]. Briefly, we identified patients eligible for a CPG order set for whom it was not ordered; we contacted the admitting provider within 2 weeks, inquired about reasons for CPG nonuse from a predefined list (we also added categories as needed), and asked for narrative comments. Based on these results, we created a Pareto chart that identified the most common barriers to CPG order set use: (1) lack of awareness or forgetting to use the CPG, (2) eligibility for multiple CPG order sets at the time of admission, and (3) use of a similarly named order set that was not the intended CPG order set.

The Intervention

CPG-specific order bundles were integrated into the general pediatrics admission order set for better visibility and more efficient usability. Order bundles for 6 CPGs were chosen for the intervention, including asthma, complicated pneumonia, heavy menstrual bleeding, migraine, musculoskeletal infection, and uncomplicated pneumonia. These incorporated CPGs were chosen because they either (1) demonstrate low guideline order set use, (2) are very common, or (3) represent important improvement areas in antimicrobial stewardship. Order bundles were added to a section titled “Common Guidelines and Pathways—General Pediatrics” (Figure 1).

Orders in each CPG order bundle were identical to the existing stand-alone CPG-associated order sets. Within each order bundle, embedded hyperlinks referenced the published CPG and relevant literature from which recommendations were made and referenced common target disease pathogens for bundles that recommended empiric first-line antibiotics. For patients that qualified for multiple CPG order sets, the integrated order set also allowed for the selection of multiple relevant order bundles within the order set.

Prior to implementation of the integrated order set into a live production environment, formative usability testing and summative usability testing were both completed. Usability testing aimed to test the effectiveness and iteratively improve the intervention in a simulated environment. Results are reported elsewhere [18].

To address the barrier of similarly named but non-CPG order sets being used accidentally, we identified order set “mimics” by searching the system with common clinical synonyms for each CPG. Similarly named order sets were retired from the production environment after obtaining approval from both the order set owners and the corresponding CPG owners. In total, 9 mimics were identified, and all of these were subsequently retired after owner approval. Additionally, all relevant CPG-associated order sets were reviewed and updated, if necessary, for both naming consistency and related search terms.

After the integrated order set was implemented into a live production environment, an update outlining the new CDS tool and its capabilities and expected use was emailed to all current and incoming residents to review and presented at the weekly resident educational conference.

Figure 1. Integrated general pediatrics admission order set with clinical practice guideline order bundles.

Study of the Intervention and Measures

All patients aged 0 to 21 years who were admitted to the general pediatrics service and met the eligibility criteria for any one of the incorporated CPG order sets based on preexisting computable population definitions were included in this study. This study used a quasi-experimental design, analyzing pre- and postintervention CPG order set adherence at both the intervention site and a nonintervention site. Our primary exposure was intervention period, with the preintervention period defined as July 1, 2019, to June 3, 2020, and the postintervention period as June 4, 2020, to May 28, 2021, as the integrated order set went live on June 4, 2020, at the intervention site.
Our primary outcome was the proportion of appropriate CPG order set use for eligible patients at the time of admission. To evaluate the impact of our intervention, we adopted existing automated queries to assess whether the clinician used the appropriate CPG order set, a wrong but similarly named order set, or the available general pediatrics admission order set. We also reviewed whether the “CPG guideline initiation order” was signed, which is a prechecked order in all our included CPG order sets. Through the query, demonstrated use of the CPG-associated order bundle and the presence of the guideline initiation order were assumed to represent appropriate guideline order set use. Encounters where the clinician appeared to use the appropriate guideline order set or bundle but the guideline initiation order was absent were manually chart reviewed to confirm appropriate order set use. All encounters that appeared eligible but where the clinician did not appear to demonstrate appropriate guideline order set use were manually chart reviewed to ensure CPG eligibility at admission throughout the study period. Eligibility was based on defined eligibility criteria in each published CPG. All manual chart review was completed by a pediatric hospital medicine fellow using both the Epic electronic health record and the Phrase Health system.

The pre- and postintervention proportion of CPG-eligible admissions for which the CPG order set or bundle was used was compared at the intervention hospital, where the integrated order set was implemented into the production environment. Data in this context were considered our “intervention cohort.” The proportion of appropriate CPG order set use for eligible patients was also compared in the same study period at the nonintervention hospital within the same health system, where the integrated order set was not implemented. This hospital uses the same CPGs and associated order sets and serves a similar patient population in the greater Atlanta area; it was thus considered our “nonintervention cohort.” The purpose of having both intervention and nonintervention cohorts in this study was to better assess whether the observed outcomes were directly related to our intervention rather than secular trends.

In this intervention, there was concern that surfacing some guidelines in the integrated order set but not others could lead to the unintended consequence of reducing order set use of CPGs that were not included in the intervention. Therefore, pre- and postintervention use of CPG order sets that were not initially included in the integrated admission order set (acute gastroenteritis, croup, bronchiolitis) was assessed as a “balancing cohort.”

Evaluation of Diagnostic Uncertainty

As our intervention was created to address lack of knowledge and awareness of guidelines, we hypothesized that it would not address diagnostic uncertainty, an underrecognized barrier that may influence order set adoption. In a post hoc analysis, we therefore aimed to evaluate the presence of diagnostic uncertainty at the time of admission to determine if this barrier accounted for a larger proportion of CPG order bundle nonuse after the intervention.

All eligible encounters where the associated CPG order set was not used were manually chart reviewed to assess the presence of diagnostic uncertainty throughout the entire study period. Diagnostic uncertainty was defined based on an algorithm (Multimedia Appendix 1) adapted from the approach of Bhise et al [15] to measuring diagnostic uncertainty in primary care. In the algorithm, encounters needed to include direct or indirect markers of uncertainty in documentation, initial definitive treatment had to have been withheld while awaiting further diagnostic workup or observation, and an operational definition of diagnostic uncertainty had to be met. Two members of the research team, a pediatric hospital medicine fellow and a pediatric resident, completed the manual chart review based on information available in the initial history and physical documentation and reported the presence or absence of diagnostic uncertainty. Interrater reliability was assessed to confirm reliability between the 2 researchers’ assessments. After chart review, the number of eligible encounters where the CPG-associated order set was not used that demonstrated the presence of diagnostic uncertainty was compared before and after implementation to determine the change in proportion after the intervention.

Analysis

Data were summarized using counts and percentages by site (intervention and nonintervention), period (pre- and postintervention), and guideline (eg, asthma or heavy menstrual bleeding). Binary logistic regression was used to analyze overall and by-guideline associations between use of appropriate CPG-specific order bundle (yes vs no) and period (pre- and postintervention) across sites via statistical interactions. We further ran binary logistic regression models evaluating the association between use of appropriate CPG-specific order bundle and period in the balancing cohort and relevant guidelines. Results are presented as contingency tables with odds ratios (ORs), 95% CIs, and corresponding P values. All analyses were conducted using SAS (version 9.4; SAS Institute), and significance was assessed at the .05 level. Percent adherence for eligible encounters by month for the intervention cohort was tracked and plotted on a statistical process P chart with annotations for the order set clean up and integrated order set go-live interventions.

Ethical Considerations

This study was deemed by the Children’s Healthcare of Atlanta Institutional Review Board to be nonhuman-subjects research as a quality improvement study (STUDY00000367).

Results

The integrated order set went live on June 4, 2020. From January 1, 2019, to May 28, 2021, a total of 1664 encounters were identified as eligible for a CPG order set based on preexisting computable population definitions. Of these encounters, 1052 were preimplementation (Figure 2) and 612 were postimplementation (Figure 3).
The number of encounters was unbalanced, partially due to lower-than-average admission volumes as a result of the COVID-19 pandemic. We manually reviewed all encounters that appeared eligible by computable definitions for a CPG order bundle where the CPG order bundle was not used: 423/1052 (40.2%) encounters before the intervention and 207/612 (33.8%) after the intervention. Of the encounters that were reviewed, 188/1052 (17.9%) before the intervention and 125/612 (20.4%) after the intervention were excluded by manual review for not meeting eligibility criteria. Overall rates of exclusion were similar when comparing the difference before and after the intervention (17.9% before and 20.4% after the intervention; 95% CI −6.62% to 1.51%, \(P = .22\)).

CPG order set use rates for included CPGs were tracked over time (Figure 4).

The trend in monthly adherence was positive following implementation and demonstrated special cause variation.
beginning in August 2020, 8 weeks after the integrated order set went live. The rate of order set use at the intervention site for integrated CPGs increased from 73.5% before the intervention to 83.3% after the intervention (OR 1.80, 95% CI 1.36-2.39). Order set use rate at the nonintervention site, where the integrated order set was not implemented but mimics were also deleted, increased from 66.3% to 71.4% during the same study period (OR 1.27, 95% CI 1.04-1.54). Of note, this increase in the nonintervention cohort appeared driven by musculoskeletal infection (OR 2.84, 95% CI 1.49-5.40) and asthma (OR 2.15, 95% CI 1.22-3.79), as seen in Table 1. When comparing ORs between the intervention and nonintervention cohorts, the intervention cohort had significantly improved order set use from before to after the intervention relative to the nonintervention cohort (intervention OR 1.80 (95% CI 1.36-2.39) vs nonintervention OR 1.27 (95% CI 1.04-1.54; \( P = .045 \)).

When broken down by disease-specific CPGs, all integrated CPGs showed positive adherence trends after implementation in the intervention cohort but with different effect sizes. Heavy menstrual bleeding and pneumonia had more improvement than musculoskeletal infection or migraine (Table 1). Adherence in asthma, for which the CPG order set has historically high use rates, remained excellent after the intervention (92.1%-95.5%; OR 1.81, 95% CI 0.95-3.45). Adoption of CPG order bundles that were not included in the integrated admission order set (including bronchiolitis, acute gastroenteritis, and croup) decreased from 84% to 68.5% following the intervention (OR 0.41, 95% CI 0.29-0.59). Of note, this was largely driven by bronchiolitis, where adoption changed from 86.9% to 75.7% after the intervention (OR 0.47, 95% CI 0.28-0.78), as seen in Table 2.

In a post hoc analysis, based on the observation that improvements were lower for musculoskeletal infection and migraine, we reviewed 308 eligible encounters where a CPG order bundle was not used to evaluate the presence of diagnostic uncertainty at admission. One reviewer (a pediatric hospital medicine fellow) completed manual chart review on all charts and a second (a pediatric resident) reviewed a random subsample of 50 encounters (16%), with interrater reliability measured by the Cohen \( \kappa \) (\( \kappa = 0.73, \ P < .001 \)). The proportion of eligible encounters where the CPG order set was not used that demonstrated diagnostic uncertainty increased from 12.3% (28/227) before implementation to 23.4% (19/81) after implementation (OR 2.18, 95% CI 1.12-4.16).

**Figure 4.** Statistical process control chart of percentage guideline order set adherence for eligible encounters at the intervention site between July 2019 to December 2020. OS: order set.
**Table 1.** Order set bundle use before and after implementation on June 4, 2020, in the intervention and nonintervention cohorts.

<table>
<thead>
<tr>
<th>Interaction</th>
<th>Intervention cohort</th>
<th>Nonintervention cohort</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No bundle, n (%)</td>
<td>Bundle, n (%)</td>
<td>OR (95% CI)</td>
</tr>
<tr>
<td>Overall(^b)</td>
<td>&lt;.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before</td>
<td>227 (26.5)</td>
<td>629 (73.5)</td>
<td>Reference</td>
</tr>
<tr>
<td>After</td>
<td>81 (16.7)</td>
<td>405 (83.3)</td>
<td>1.80 (1.36-2.39)</td>
</tr>
<tr>
<td>Asthma(^c)</td>
<td>.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before</td>
<td>39 (7.9)</td>
<td>454 (92.1)</td>
<td>Reference</td>
</tr>
<tr>
<td>After</td>
<td>13 (4.5)</td>
<td>274 (95.5)</td>
<td>1.81 (0.95-3.45)</td>
</tr>
<tr>
<td>Heavy menstrual bleeding(^d)</td>
<td>.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before</td>
<td>8 (42.1)</td>
<td>11 (57.9)</td>
<td>Reference</td>
</tr>
<tr>
<td>After</td>
<td>5 (11.6)</td>
<td>38 (88.4)</td>
<td>5.53 (1.50-20.35)</td>
</tr>
<tr>
<td>Musculoskeletal infection(^e)</td>
<td>.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before</td>
<td>22 (75.9)</td>
<td>7 (24.1)</td>
<td>Reference</td>
</tr>
<tr>
<td>After</td>
<td>21 (63.6)</td>
<td>12 (36.4)</td>
<td>1.80 (0.59-5.44)</td>
</tr>
<tr>
<td>Migraine(^f)</td>
<td>.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before</td>
<td>25 (30.1)</td>
<td>58 (69.9)</td>
<td>Reference</td>
</tr>
<tr>
<td>After</td>
<td>19 (27.9)</td>
<td>49 (72.1)</td>
<td>1.11 (0.55-2.26)</td>
</tr>
<tr>
<td>Complicated pneumonia(^g)</td>
<td>.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before</td>
<td>7 (87.5)</td>
<td>1 (12.5)</td>
<td>Reference</td>
</tr>
<tr>
<td>After</td>
<td>5 (62.5)</td>
<td>3 (37.5)</td>
<td>4.20 (0.33-53.11)</td>
</tr>
<tr>
<td>Uncomplicated pneumonia(^h)</td>
<td>.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before</td>
<td>126 (56.3)</td>
<td>98 (43.7)</td>
<td>Reference</td>
</tr>
<tr>
<td>After</td>
<td>18 (38.3)</td>
<td>29 (61.7)</td>
<td>2.07 (1.09-3.95)</td>
</tr>
</tbody>
</table>

\(^a\) OR: odds ratio.

\(^b\) Intervention cohort: n=1342; nonintervention cohort: n=2064.

\(^c\) Intervention cohort: n=780; nonintervention cohort: n=897.

\(^d\) Intervention cohort: n=62; nonintervention cohort: n=80.

\(^e\) Intervention cohort: n=62; nonintervention cohort: n=161.

\(^f\) Intervention cohort: n=151; nonintervention cohort: n=378.

\(^g\) Intervention cohort: n=16; nonintervention cohort: n=52.

\(^h\) Intervention cohort: n=271; nonintervention cohort: n=496.
Discussion

Summary
The integration of CPG order bundles into a general pediatric admission order set improved CPG adoption in a stand-alone academic pediatric hospital compared to a control hospital within the same health system. In a post hoc analysis, the disease processes with lower diagnostic uncertainty at the time of admission saw the greatest improvement from this intervention.

Interpretation
CPG adoption improved both in relation to preintervention encounters at the same hospital and in relation to encounters at a similar hospital within the same institution where the integrated order set was not released. This suggests that the increase in CPG adherence was directly related to the implementation of the integrated order set at the study site. While CPG adherence also significantly improved at the nonintervention hospital, the improvement seen at the intervention site was significantly more than that at the nonintervention site. Additionally, improvement was only seen for 2 of the 6 guidelines (asthma and musculoskeletal infection) at the nonintervention site, compared to all 6 guidelines showing a trend toward improvement at the intervention site. This change may reflect the removal of known CPG order set “mimics” at both locations prior to the integrated order set implementation, as this was identified as a barrier to CPG adherence in prior work. While some CPGs demonstrated improvements of close to 20%, overall improvement did not meet our initial primary aim of 20% increased adherence after the intervention. This is likely due to finding a higher than anticipated preintervention overall adherence rate, largely driven by the CPG for asthma, which has historically high adherence rates.

Nonincorporated CPGs demonstrated a reduction in order set use following implementation of the integrated order set. This finding was largely attributable to a decrease in bronchiolitis guideline adherence. The timing of this intervention, in June 2020, correlated to multiple surges of SARS-CoV-2 infections. We are unable to distinguish whether the reduction in adherence was due to our intervention or the change in the management of respiratory infections during this time. Future studies that incorporate a more comprehensive list of CPGs may elucidate how this decision support design affects nonincorporated CPGs.

The presence of diagnostic uncertainty was not initially identified as a primary barrier to guideline adherence based on frontline clinician queries during this study [18]. In our analysis, the proportion of eligible encounters without CPG adherence that demonstrated diagnostic uncertainty increased following the implementation of the integrated order set. Our intervention addressed other drivers, but not diagnostic uncertainty, which may explain a higher fraction of diagnostic uncertainty in encounters without a CPG order bundle after implementation. Alternative designs that account for the change in diagnostic certainty across a hospitalization may demonstrate a better job of improving CPG adherence.

While previous literature has shown that CPGs and associated order sets can successfully decrease variation in care delivery and improve patient outcomes [19-22], the context in which decision support is aligned into the workflow remains of utmost importance for the success of these interventions [4]. Combining alerts with order sets has been shown to have success in specific contexts [23,24] but risks generating alert fatigue and requires considerable disease-specific logic behind the alert. Alternatively, automating order suggestions through machine learned patterns of order use was shown to influence ordering behavior [4,25] but may not reflect evidence-based recommendations and can be resource intensive.

Furthermore, increasing patient medical complexity and diagnostic uncertainty leads to a workflow mismatch when coupled with single diagnoses or simplified guidelines. This gap was largely underrecognized by clinicians when self-reporting barriers to guideline adherence [18] and likely requires decision support in a different context or format to overcome than admission order sets. Mehta et al [16] sought to integrate CDS into documentation workflows through problem-oriented templates aimed at improving documentation.

Table 2. Pre- and postimplementation (before and on or after June 4, 2020, respectively) order set bundle use in the balancing cohort.

<table>
<thead>
<tr>
<th>CPG Category</th>
<th>Overall (n=854)</th>
<th>Bronchiolitis (n=563)</th>
<th>Acute gastroenteritis (n=109)</th>
<th>Croup (n=184)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Preimplementation</td>
<td>Postimplementation</td>
<td>Preimplementation</td>
<td>Postimplementation</td>
</tr>
<tr>
<td></td>
<td>No bundle, n (%)</td>
<td>Bundle, n (%)</td>
<td>Odds ratio (95% CI)</td>
<td>P value</td>
</tr>
<tr>
<td>Overall</td>
<td>102 (16)</td>
<td>536 (84)</td>
<td>Reference</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Bronchiolitis</td>
<td>68 (31.5)</td>
<td>148 (68.5)</td>
<td>0.41 (0.29-0.59)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Acute gastroenteritis</td>
<td>19 (36.5)</td>
<td>3 (10.5)</td>
<td>Reference</td>
<td>&lt;.004</td>
</tr>
<tr>
<td>Croup</td>
<td>6 (12.5)</td>
<td>42 (87.5)</td>
<td>0.57 (0.19-1.65)</td>
<td>.29</td>
</tr>
</tbody>
</table>

Reference124 (92.5)10 (7.5) Preimplementation
Reference19 (36.5)33 (63.5) Preimplementation
Reference393 (86.9)59 (13.1) Preimplementation
Reference536 (84)102 (16) Preimplementation
Reference0.290.57 (0.19-1.65)42 (87.5)6 (12.5) Postimplementation
Reference0.831.09 (0.50-2.37)22 (38.6)35 (61.4) Postimplementation
Reference0.0040.47 (0.28-0.78)84 (75.7)27 (24.3) Postimplementation
Reference<.0010.41 (0.29-0.59)148 (68.5)68 (31.5) Postimplementation
Reference
for patients with multiple problems and to provide greater evidence-based prompts and organization. While demonstrating potential to provide recommendations during the documentation process, potentially a better context to address diagnostic uncertainty, clinicians are still called upon to label diagnoses for their patients at a point in time when this may still be unclear. Further research into the most effective format and context for CDS to address diagnostic uncertainty is needed.

Limitations
This study has several limitations. First, results may not be generalizable, as this was a multisite, single-system study focused on a single service. Different contexts, organizational cultures, and electronic health record vendors could affect the feasibility and impact of this intervention. Second, due to the COVID-19 pandemic, hospital admission volumes were significantly lower, and admissions consisted of fewer respiratory illnesses in the postimplementation period, potentially creating pre- and postintervention cohorts that were less similar. Additionally, due to time and resource constraints, a single chart reviewer was used for charts flagged as nonadherent to confirm eligibility. Despite this limitation, pre- and postimplementation rates of exclusion were similar, suggesting this did not have a significant influence on results. Lastly, in a post hoc analysis of diagnostic uncertainty, only one reviewer reviewed all nonadherent charts to determine the presence of uncertainty. The development of an algorithm for uncertainty and an assessment of interrater reliability for a subset of charts attempted to address this limitation and minimize subjectivity.

Conclusion
The integration of CPG-specific order bundles into a general pediatrics admission order set improved overall CPG adoption by addressing the most commonly reported barriers to CPG adherence by clinicians. Further improvement in guideline adherence could be seen with integration of a more comprehensive list of available guidelines for a particular service. Diagnostic uncertainty at the time of admission is likely an underrecognized barrier to guideline adherence that is not fully addressed with an integrated admission order set. Further work is needed to determine the impact of an integrated admission order set on clinical outcomes and what types of clinical decision support could better address the presence of diagnostic uncertainty.

Acknowledgments
We would like to acknowledge Christy Bryant for her assistance in building the redesigned order set as well as the pediatric residents and attendings who participated in all stages of this study. JM, SK, SG, and ID’s participation was supported by the Emory Department of Pediatrics and Children’s Healthcare of Atlanta through the Warshaw Fellow Research Award.

Authors’ Contributions
JM, SK, CS, DR, and EO conceptualized and designed the project. JM performed manual chart review for nonadherence. JM and JG created the algorithm for diagnostic uncertainty and performed chart review for diagnostic uncertainty. JG structured and wrote Multimedia Appendix 1 with input from JM. ID and SG performed statistical analysis and contributed to interpretation of results. All authors discussed the results and conclusions of the study. JM wrote the manuscript with contribution and input from all authors.

Conflicts of Interest
EO is a cofounder of and has equity in Phrase Health, a clinical decision support analytics company. He receives no direct revenue from this relationship. No other authors declare any conflicts.

Multimedia Appendix 1
Overview of algorithm to evaluate for diagnostic uncertainty with case examples.

References


Interoperable, Domain-Specific Extensions for the German Corona Consensus (GECCO) COVID-19 Research Data Set Using an Interdisciplinary, Consensus-Based Workflow: Data Set Development Study

Gregor Lichtner\textsuperscript{1,2,3}, PhD; Thomas Haese\textsuperscript{1}, MSc; Sally Brose\textsuperscript{1}, MSc; Larissa Röhrig\textsuperscript{1,4}, MSc; Liudmila Lysyakova\textsuperscript{5,6}, PhD; Stefanie Rudolph\textsuperscript{5,6}, PhD; Maria Uebe\textsuperscript{5,6}, BA; Julian Sass\textsuperscript{1}, MSc; Alexander Bartschke\textsuperscript{1}, MSc; David Hillus\textsuperscript{7}, MD; Florian Kurth\textsuperscript{7,8,9}, MD; Leif Erik Sander\textsuperscript{7}, MD; Falk Eckart\textsuperscript{10}, MD; Nicole Toepfner\textsuperscript{10}, MD; Reinhard Berner\textsuperscript{10}, MD; Anna Frey\textsuperscript{11}, MD; Marcus Dörr\textsuperscript{12}, MD; Jörg Janne Vehreschild\textsuperscript{13,14,15}, MD; Christof von Kalle\textsuperscript{5,6}, MD; Sylvia Thun\textsuperscript{1}, MD

Corresponding Author: Gregor Lichtner, PhD

Abstract

Background: The COVID-19 pandemic has spurred large-scale, interinstitutional research efforts. To enable these efforts, researchers must agree on data set definitions that not only cover all elements relevant to the respective medical specialty but also are syntactically and semantically interoperable. Therefore, the German Corona Consensus (GECCO) data set was developed as a harmonized, interoperable collection of the most relevant data elements for COVID-19–related patient research. As the GECCO data set is a compact core data set comprising data across all medical fields, the focused research within particular medical domains demands the definition of extension modules that include data elements that are the most relevant to the research performed in those individual medical specialties.

Objective: We aimed to (1) specify a workflow for the development of interoperable data set definitions that involves close collaboration between medical experts and information scientists and (2) apply the workflow to develop data set definitions that include data elements that are the most relevant to COVID-19–related patient research regarding immunization, pediatrics, and cardiology.

Methods: We developed a workflow to create data set definitions that were (1) content-wise as relevant as possible to a specific field of study and (2) universally usable across computer systems, institutions, and countries (ie, interoperable). We then gathered medical experts from 3 specialties— infectious diseases (with a focus on immunization), pediatrics, and cardiology—to select data elements that were the most relevant to COVID-19–related patient research in the respective specialty. We mapped the data elements to international standardized vocabularies and created data exchange specifications, using Health Level Seven International (HL7) Fast Healthcare Interoperability Resources (FHIR). All steps were performed in close interdisciplinary collaboration with medical domain experts and medical information specialists. Profiles and vocabulary mappings were syntactically and semantically validated in a 2-stage process.
Results: We created GECCO extension modules for the immunization, pediatrics, and cardiology domains according to pandemic-related requests. The data elements included in each module were selected, according to the developed consensus-based workflow, by medical experts from these specialties to ensure that the contents aligned with their research needs. We defined data set specifications for 48 immunization, 150 pediatrics, and 52 cardiology data elements that complement the GECCO core data set. We created and published implementation guides, example implementations, and data set annotations for each extension module.

Conclusions: The GECCO extension modules, which contain data elements that are the most relevant to COVID-19–related patient research on infectious diseases (with a focus on immunization), pediatrics, and cardiology, were defined in an interdisciplinary, iterative, consensus-based workflow that may serve as a blueprint for developing further data set definitions. The GECCO extension modules provide standardized and harmonized definitions of specialty-related data sets that can help enable interinstitutional and cross-country COVID-19 research in these specialties.

**KEYWORDS** interoperability; research data set; Fast Healthcare Interoperability Resources; FHIR; FAIR principle; COVID-19; interoperable; SARS-CoV-2; pediatric; immunization; cardiology; standard

### Introduction

The COVID-19 pandemic has led to unprecedented, strong efforts in connecting nationwide and international research to help manage the disease and its effects on public health. To enable research across different health care providers, institutions, or even countries, interoperability between medical data systems is essential [1]. Therefore, early in the pandemic, the German Corona Consensus (GECCO) data set was developed in a collaborative effort to provide a standardized, unified core data set for interinstitutional COVID-19–related patient research [2]. The GECCO data set specifies a set of 81 essential clinical data elements from 13 domains, such as anamnesis and risk factors, symptoms, and vital signs, that have been selected by expert committees from university hospitals, professional associations, and research initiatives. Since its development, the GECCO data set has been implemented in a large number of institutions, most notably in virtually all German university hospitals, which now provide access to the GECCO data set in the context of the German COVID-19 Research Network of University Medicine (“Netzwerk Universitätsmedizin”) [3,4].

The GECCO data set was developed to contain as many relevant data elements as possible but few enough to keep the effort of implementing the data set manageable. Therefore, the data set contains mostly data elements of general research interest, excluding data elements that are only of interest for particular medical specialties or use cases. These data items are considered part of domain-specific extension modules of the GECCO data set, which are introduced in this paper.

We aimed to develop domain-specific extensions to the GECCO data set that cover the most relevant data elements for COVID-19–related patient research in the infectious disease (with a focus on immunization), pediatrics, and cardiology medical specialities. To that end, we first developed a workflow that aims at providing data set definitions that (1) contain the most relevant data elements for the research aims of the end users and (2) can be applied universally across institutions and countries. We then followed that workflow with different groups of medical experts from different medical specialties to define extension modules that are relevant for research regarding immunization, pediatrics, and cardiology.

These extension modules complement the GECCO core data set and use the same international health IT standards and terminologies as those in the GECCO data set, such as the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [5], the Logical Observation Identifiers Names and Codes (LOINC) [6,7], and the Fast Healthcare Interoperability Resources (FHIR) [8,9] standard. The extension modules were developed in close alignment with the GECCO data set to ensure interoperability and compatibility with existing definitions.

We herein describe the consensus-based data element selection and data format definition workflow that we applied in close collaboration with medical experts from 3 specialties—infected diseases (with a focus on immunization), pediatrics, and cardiology (ie, for content definition)—as well as medical information specialists and FHIR developers (ie, for technical aspects). This workflow may serve as a blueprint for the further development of consensus-based data set definitions.

### Methods

#### Workflow Definition

We aimed to develop a workflow to create data set definitions that are (1) content-wise as relevant as possible to a specific field of study and (2) universally usable across computer systems, institutions, and countries (ie, interoperable). We based the specification of the workflow on our experience with the definition of the GECCO data set, during which health professionals from 50 institutions (university hospitals, professional associations, and other relevant organizations) participated to define the most relevant data elements for general-scope, COVID-19–related research [2]. To fulfill the first requirement (relevancy), we decided to leave the full responsibility of data element selection to groups of medical professionals of the respective specialty, with only minimal interference by the medical information specialists. We deliberately did not specify the exact process of how the group of medical experts could select the data elements (eg, literature...
review, focus groups, and consensus-based processes) to allow for the maximal flexibility of the data set definition workflow, with respect to the medical experts’ values and preferences. To fulfill the second requirement (interoperability), we adopted a model that was loosely based on the data FAIRification workflow of Jacobsen et al [10]; the mapping, quality assurance, and publication steps are outlined in detail below.

Selection of Data Items

The content of the domain-specific research data sets was defined by medical domain experts in a transparent workflow (Figure 1). The involvement of the medical domain experts as the end users of the data to be provided ensured that the contents of the data sets were aligned to the actual research needs. In our project, the so-called subject- and organ-specific working groups of the National Pandemic Cohort Network (“Nationales Pandemie Kohorten Netz” [NAPKON]) served as the domain-specific groups of medical experts. These groups were established by a voluntary association of medical experts from the medical specialties within the nationwide NAPKON project in Germany. Each of the subject- and organ-specific working groups elected a board, and all communication between the data set developers and the working groups was organized and carried out via the working groups’ boards. In preparation for the GECCO extension modules, we invited the subject- and organ-specific groups for infectious diseases (with a focus on immunization), pediatrics, and cardiology to provide up to 50 data elements (with up to 10 response items each) that were, in the view of the medical experts, the most relevant to patient-related COVID-19 research in these medical specialties and not already included in the GECCO core data set. If necessary, more data items or response options could be provided in coordination with the medical information specialists. The provided data items were then reviewed by the medical information specialists, and a first definition of the contents of the extension module was returned to the respective subject- and organ-specific working group for approval or change requests. After approval by the subject- and organ-specific working group, the definition of the extension module content was considered finalized.
Figure 1. Flowchart of the consensus-based, interdisciplinary data set definition and mapping workflow for the domain-specific COVID-19 research data sets. FHIR: Fast Healthcare Interoperability Resources.

Development of the Standardized Data Formats
To map the data items selected by the subject- and organ-specific working groups to international standard vocabularies, we performed a consensus-based mapping procedure, wherein every concept was mapped to appropriate vocabularies—the SNOMED CT for general concepts [11]; LOINC for observations [7]; International Statistical Classification of Diseases and Related Health Problems, 10th Revision, German Modification for diagnoses [12]; Anatomical Therapeutic Chemical Classification System for Germany for drugs and active ingredients [13]; and Unified Code for Units of Measurement for measurement units [14]—by 2 medical information specialists independently. Ambiguities and nonmatching mappings were then discussed among the medical information specialists and in close collaboration with the medical experts of the subject- and organ-specific working groups until consensus was achieved. The data item–to-concept mappings were annotated on ART-DECOR, an open-source collaboration platform for creating and maintaining data set element descriptions [15].

As for the GECCO data set, the format for data exchange was specified by using Health Level Seven International (HL7) FHIR resources. The mapping of data items to FHIR resources was performed in an iterative, consensus-based workflow among the medical information specialists. Wherever possible, published FHIR profiles from the GECCO data set, the Medical
The profiles and value sets were specified by using the FHIR Shorthand (FSH) language (version 1.2.0) and translated to Structure Definition JSON files by using the HL7 FSH SUSHI software package (version 2.2.3) [18,19]. We required that at least one exemplary instance be defined for every profile. The syntactic validation of the profile and value set definitions was performed through the error-free conversion of the FSH files to JSON via SUSHI, and the subsequent validation of each profile and their defined instances was performed by using the HL7 FHIR validator as implemented in the FSH Validator Python package (version 0.2.2) [20]. After the successful syntactic validation of a set of profiles, the profiles were subjected to a 2-stage review process, as follows. First, the profiles and the corresponding value sets and extensions were internally reviewed for semantic appropriateness with the GECCO core developer (JS). After all necessary changes and approval by the internal reviewer, the profiles were subjected to the second review round by an external FHIR development expert. Subsequent to necessary corrections and the approval of the external reviewer, the respective profiles, together with their value sets and, optionally, extensions and code systems, were considered finalized and published to the main branch of the Git repository. The subsequent and ongoing maintenance phase of the data set definitions involves inviting implementers and users to report any issues that they encounter with the definitions, in order to ensure their accuracy and relevance over time.

The whole development process was performed collaboratively on GitHub. The syntactic validation of the profiles was performed via continuous integration/continuous development workflows, which were implemented as GitHub actions. Semantic validation during the internal and external review rounds was performed by using pull requests to 2 different Git branches. After the final approval, profiles and value sets were merged into the main branch of the respective extension module’s repository, which served as the publication branch of that module. Since then, maintenance requests and updates of the extension modules have been handled via GitHub issues. All kinds of relevant changes have become subjects of the internal review, as defined above; major changes (eg, nontechnical corrections) are additionally exposed to the external review.

Implementation guides were created for all 3 extension modules, using the FHIR IG Publisher tool and a customized template for the implementation guides’ HTML pages [21]. The implementation guides were published to GitHub pages, where they remain automatically synchronized with the main branch of the respective repository via continuous integration/continuous development workflows.

Ethics Approval

This study did not involve any human or animal experiments. No permissions were required to access any data used in this study.

Results

Data Set Definition Workflow

We developed an interdisciplinary, iterative, expert consensus–based workflow for the initial definition of domain-specific COVID-19 research data sets based on 2 key requirements. The first key requirement for the content of the data sets was that the content definition (ie, selection of data elements) was to be performed under the full responsibility of a group of medical experts to ensure that the selected data elements were truly those that are required for research in the respective medical specialty. The second key requirement was to produce FAIR (Findable, Accessible, Interoperable, Reusable) digital assets [22], that is, the data set definitions should be represented in FHIR profiles and implementation guides, and these should be registered on open platforms (ie, findable); they should be retrievable through open, free, standard protocols (ie, accessible); they should use only standard, international medical terminologies, such as SNOMED CT and LOINC (ie, interoperable); and they should be released with rich usage guides and examples (FHIR implementation guide) and under a permissive license (ie, reusable).

To fulfill these requirements, the data set definition workflow consists of the following 6 phases: content definition, mapping, quality assurance, publication, an optional public review, and maintenance (Figure 1). In the content definition phase, a group of medical experts from a particular medical specialty are approached by the medical information specialists and asked to provide a list of the data elements that are the most relevant to patient-related COVID-19 research in the respective medical specialty. How the medical expert group compiles the list in detail is left to their discretion (eg, based on systematic literature review or Delphi consensus processes). The medical information specialists only review the provided lists for consistency and redundancy and compile the final content definition in agreement with the medical expert group. In the mapping phase, all data elements are then mapped to international terminologies in consultation with the group of medical experts. Based on these, a logical model and the mappings of data elements to FHIR resources are established. In the quality assurance phase, the FHIR specifications are syntactically validated by using the HL7 FHIR validator as implemented in the FSH Validator Python package (version 0.2.2) [20] and then subjected to a 2-stage review process, during which 2 individual data interoperability and harmonization experts validate the specifications semantically, that is, they validate that the data elements defined by the group of medical experts are appropriately mapped to international standards. After any required changes, the logical model and the FHIR implementation guide are published and are openly accessible to the research community in repositories that fulfill the FAIR criteria as closely as possible, such as ART-DECOR [15] for the logical model and GitHub or the FHIR Implementation Guide registry for the implementation guide [23]. If desired, the initial release of the data set definition can be subjected to public review and balloting processes, which allow stakeholders to provide feedback and suggest changes. The public review and balloting processes provide an opportunity to obtain broader
input from and facilitate consensus building among the research community and stakeholders. Any changes resulting from the review and balloting processes can then be incorporated into the data set definition according to the herein presented workflow, and the updated version is released and maintained according to the same workflow. In the maintenance phase, the medical information specialists invite implementers and users of the data set definitions to report any issues they encounter with the definitions via GitHub issues or email, in order to ensure their accuracy and relevance over time. During the maintenance phase, requests for changes or updates to the data set definition should generally be limited to minor issues or corrections, as adding new data elements or making significant modifications to the definition would require running the entire workflow from the beginning.

Data Set Contents

Groups of Medical Experts

In the context of the NAPKON project of the German COVID-19 Research Network of University Medicine [24], so-called subject- and organ-specific working groups were established by the voluntary association of medical experts from different medical specialties. In preparation for the domain-specific data set definitions that extend the GECCO core data set, the working groups for infectious diseases (with a focus on immunization), pediatrics, and cardiology were invited by the data set development group to provide up to 50 data elements (with up to 10 response items each) that were of particular interest to their field, concerned patient-related COVID-19 research, and were not already included in the GECCO core data set. For the immunization data set definition, physicians from the COVIM (Collaborative Immunity Platform of the Netzwerk Universitätsmedizin) study for the determination and use of SARS-CoV-2 immunity [25-27] assumed the role of the organ-specific working group, as no such working group had been established previously.

Overview

We extended the GECCO core data set by developing domain-specific data set definitions for a total of 250 new data items—48 for the immunization extension module, 150 for the pediatrics extension module, and 52 for the cardiology extension module. These data items were collected, via an iterative consensus-based approach, from the subject- and organ-specific working groups, and they fall under 10 of the 13 data categories of the GECCO data set (Table 1). Data elements and the number of items for each individual extension module are shown in Tables 2, 3, and 4. The full lists of items are shown in the Tables S1-S3 in Multimedia Appendix 1.

Table 1. Number of data items per GECCO extension module.

<table>
<thead>
<tr>
<th>GECCO data category</th>
<th>GECCO extension module</th>
<th>Pediatrics data items (N=150), n</th>
<th>Cardiology data items (N=52), n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anamnesis and risk factors</td>
<td>13</td>
<td>21</td>
<td>6</td>
</tr>
<tr>
<td>Complications</td>
<td>24</td>
<td>47</td>
<td>7</td>
</tr>
<tr>
<td>Demographics</td>
<td>—</td>
<td>6</td>
<td>—</td>
</tr>
<tr>
<td>Epidemiological factors</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Imaging</td>
<td>—</td>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td>Laboratory values</td>
<td>1</td>
<td>27</td>
<td>2</td>
</tr>
<tr>
<td>Medication</td>
<td>1</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>Onset of illness and admission</td>
<td>6</td>
<td>2</td>
<td>—</td>
</tr>
<tr>
<td>Outcome at discharge</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Study enrollment and inclusion criteria</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Symptoms</td>
<td>—</td>
<td>9</td>
<td>—</td>
</tr>
<tr>
<td>Therapy</td>
<td>2</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>Vital signs</td>
<td>1</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

aGECCO: German Corona Consensus.
bNot available.
<table>
<thead>
<tr>
<th>Category and data element</th>
<th>FHIR resource</th>
<th>Items (N=48), n</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Anamnesis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemotherapy</td>
<td>Procedure</td>
<td>1</td>
</tr>
<tr>
<td>Immunosuppressive therapy</td>
<td>Procedure</td>
<td>1</td>
</tr>
<tr>
<td>Regular alcohol intake</td>
<td>Observation</td>
<td>2</td>
</tr>
<tr>
<td><strong>COVID-19 infection and treatment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease course</td>
<td>Encounter, Procedure</td>
<td>5</td>
</tr>
<tr>
<td>SARS-CoV-2 infection</td>
<td>Condition</td>
<td>1</td>
</tr>
<tr>
<td>SARS-CoV-2 variant</td>
<td>Observation</td>
<td>1</td>
</tr>
<tr>
<td><strong>Immunization</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contraindications to immunization</td>
<td>Immunization</td>
<td>2</td>
</tr>
<tr>
<td>Immunizations performed</td>
<td>Immunization</td>
<td>3</td>
</tr>
<tr>
<td>Reason for immunization</td>
<td>Immunization</td>
<td>5</td>
</tr>
<tr>
<td>Willingness to receive additional immunization doses</td>
<td>Observation</td>
<td>1</td>
</tr>
<tr>
<td><strong>Immunization reactions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analgesic or antipyretic drug intake</td>
<td>MedicationStatement</td>
<td>1</td>
</tr>
<tr>
<td>Body temperature</td>
<td>Observation</td>
<td>1</td>
</tr>
<tr>
<td>Complications after immunization</td>
<td>Observation</td>
<td>5</td>
</tr>
<tr>
<td>Medical treatment for adverse reactions</td>
<td>Encounter</td>
<td>3</td>
</tr>
<tr>
<td>Symptoms after vaccination</td>
<td>Condition</td>
<td>16</td>
</tr>
</tbody>
</table>

aGECCO: German Corona Consensus.
bFHIR: Fast Healthcare Interoperability Resources.
<table>
<thead>
<tr>
<th>Category and data element</th>
<th>FHIR resource</th>
<th>Items (N=150), n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complications</td>
<td>Complications to COVID-19</td>
<td>47</td>
</tr>
<tr>
<td>Demographics</td>
<td>Body measures</td>
<td>6</td>
</tr>
<tr>
<td>Imaging</td>
<td>Echocardiography</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>PET-CT&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1</td>
</tr>
<tr>
<td>Immunization</td>
<td>Immunizations performed</td>
<td>2</td>
</tr>
<tr>
<td>Laboratory values</td>
<td>Laboratory values</td>
<td>27</td>
</tr>
<tr>
<td>Medical history</td>
<td>Chronic hematologic diseases</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Chronic kidney diseases</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Congenital disease</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Gastrointestinal diseases</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Medical history stem cells transplant</td>
<td>2</td>
</tr>
<tr>
<td>Medication</td>
<td>Medication</td>
<td>35</td>
</tr>
<tr>
<td>Symptoms</td>
<td>COVID-19 symptoms</td>
<td>9</td>
</tr>
<tr>
<td>Therapy</td>
<td>Hospitalization</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Thoracic drainage</td>
<td>1</td>
</tr>
</tbody>
</table>

<sup>a</sup>GECCO: German Corona Consensus.

<sup>b</sup>FHIR: Fast Healthcare Interoperability Resources.

<sup>c</sup>PET-CT: positron emission tomography–computed tomography.
Table. Types of data elements in the cardiology extension module extending the GECCO\(^a\) core data set. Shown are the data elements and the FHIR\(^b\) resource they have been mapped to, as well as the number of items for each data element (ie, different response options).

<table>
<thead>
<tr>
<th>Category and data element</th>
<th>FHIR resource</th>
<th>Items (N=52), n</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Anamnesis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic cardiologic diseases</td>
<td>Condition</td>
<td>6</td>
</tr>
<tr>
<td>COVID-19–related complications</td>
<td>Condition</td>
<td>7</td>
</tr>
<tr>
<td><strong>Echocardiography</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Echocardiography findings</td>
<td>Observation</td>
<td>20</td>
</tr>
<tr>
<td>Echocardiography procedure</td>
<td>Procedure</td>
<td>3</td>
</tr>
<tr>
<td><strong>Electrocardiography</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electrocardiography findings</td>
<td>Observation</td>
<td>11</td>
</tr>
<tr>
<td>Electrocardiography procedure</td>
<td>Procedure</td>
<td>2</td>
</tr>
<tr>
<td><strong>Laboratory values</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laboratory values</td>
<td>Observation</td>
<td>2</td>
</tr>
<tr>
<td><strong>Medication</strong></td>
<td>MedicationStatement</td>
<td>1</td>
</tr>
<tr>
<td>Angiotensin receptor antagonist</td>
<td>MedicationStatement</td>
<td>1</td>
</tr>
</tbody>
</table>

\(^a\)GECCO: German Corona Consensus.

\(^b\)FHIR: Fast Healthcare Interoperability Resources.

All data items were mapped to the appropriate FHIR resources (Observation, Condition, Procedure, MedicationStatement, Encounter, Questionnaire, QuestionnaireResponse, Immunization, ImagingStudy, List, and Specimen), and 26, 14, and 18 profiles (25, 17, and 12 value sets) were created for the immunization, pediatrics, and cardiology extension modules, respectively. The data items that were already part of the GECCO data set and not removed during the data selection step were taken from the GECCO data set and referenced as such in the implementation guides.

The implementation guides for the three extension modules have been published on GitHub pages [28-30]. The source FSH files have been published on GitHub [31-33]. Logical models and data set descriptions are hosted on ART-DECOR, an open collaboration platform for modeling data set definitions, their descriptions, and their terminology bindings [34-36].

**Discussion**

**Principal Findings**

We herein present an interdisciplinary, iterative, consensus-based workflow for the definition of research data sets, focusing on creating data sets with the most relevant data elements for a particular field of study and on creating universally usable data sets according to the FAIR principles [22]. We applied the workflow to develop 3 GECCO extension modules that contain data items that are relevant for COVID-19–related patient research on infectious diseases (with a focus on immunization), pediatrics, and cardiology. These extension modules complement the GECCO core data set for domain-specific research. The data items are represented in HL7 FHIR profiles and use international terminologies to ensure a harmonized, standardized, and interoperable data set definition for these medical domains. The provision of data according to the extension modules introduced in this paper will enable cross-institutional and cross-country data collection and collaborative research with a particular focus on immunization, pediatrics, and cardiology.

We have specified and implemented an interdisciplinary, iterative, consensus-based workflow for the selection of data items and the development of the data set definition. Close collaboration and constant feedback loops with domain experts from various medical specialties right from the beginning of a project, as performed in this study, are key for the successful development of a useful data set definition. Indeed, since the selection of relevant data items in this study was driven by the end users of the data set, who are the researchers that later will be using the data for their specialized areas of research, the semantic usability of the data sets is guaranteed. Likewise, having medical information specialists develop the formal data set specification ensures the technical interoperability and usability of the data set definition. In this study, we focused on the initial development of interoperable data set definitions for COVID-19–related patient research on infectious diseases (with a focus on immunization), pediatrics, and cardiology. To ensure the continued accuracy and relevance of the data set definitions, such data set definitions should be regularly subjected to public review and balloting processes following the initial development. For example, a revised version of the GECCO data set will undergo HL7 balloting, pending stakeholders’ approval.

Although general interoperability in health care and clinical research is difficult to achieve, we focused on achieving syntactic and semantic interoperability of the data set definitions,
which are 2 of the 4 levels into which interoperability can be distinguished, alongside technical and organizational interoperability [8]. We pursued semantic interoperability by using international standardized vocabularies, such as those provided by the LOINC and SNOMED CT vocabularies, to ensure that the meanings of the data elements and their interpretations were unambiguous. We pursued syntactic interoperability by using an open standard for data representation, namely the HL7 FHIR standard, which provides a flexible and extensible framework for exchanging data elements and resources between different systems and applications. We did not focus on organizational interoperability in our work, as this requires coordination and alignment between different health care organizations and stakeholders, which can be challenging in practice. Although we did not specifically address organizational interoperability in our study, we believe that our approach to achieving semantic and syntactic interoperability can contribute to broader efforts toward achieving organizational interoperability over time.

In addition to the successful development of data set definitions, several factors determine a successful deployment or the use of the developed extension modules [37]. First and most importantly, clear and concise documentation of how to implement and provide data using the data set definition is required. For FHIR-based data set definitions, so-called implementation guides are used to provide a narrative overview as well as technical details on the data set definition [38]. Thus, we have created and published implementation guides for each of the here developed extension modules. Second, the example implementations of the extension modules serve as a blueprint for developers and data engineers who implement the extension modules for their clinical databases. From our experience with the implementation of the GECCO data set, well-defined example data items may be of equal if not higher importance than the technical description of the data set specification, as developers and engineers tend to use the examples as blueprints for their implementation. Thus, we equipped every FHIR profile defined in the extension modules with at least one example. These examples are incorporated and issued within the implementation guides of the modules. Specifically, we aimed to provide 1 example for each different category of response option per profile. Third, the actual implementation of the extension modules should be part of follow-up infrastructure projects to supply funding and resources for filling the data set definition with actual data. For the GECCO data set, this is ensured by follow-up projects of the German COVID-19 Research Network of University Medicine (“Netzwerk Universitätsmedizin”), such as CODEX+ (Collaborative Data Exchange and Usage), which includes several implementation tasks that are actively using the GECCO data set items [39] and further projects [40-43]. Fourth, once the data set definitions are implemented and leveraged in use cases, additional demands to the data set are likely raised, or issues with existing definitions are revealed. The maintenance of existing definitions (eg, performing technical corrections, evolving the definitions, or adding new items) is, therefore, necessary and must be organized and funded. Last, successful use of the extension modules is also highly dependent on the degree of interoperability of the data set definitions [1,44,45]. For example, the use of questionnaires to assess certain features is common in clinical research. However, depending on the exact wording of the question and the number and wording of response options, results from different studies might not be directly comparable even if they assessed the same features, as the questions and response options differ between studies. In the presented extension modules, several items were at first specified in a questionnaire-like fashion, and the direct implementation of these as Questionnaire resources in FHIR would have limited the applicability of such data elements, especially when aiming to map these elements from an electronic health record system. In these cases, we revised the data element specification to use interoperable concepts rather than questions. Here, repeated consultation with and final approval of the group of medical experts were key to being able to convert questions into interoperable concepts that convey the same information as that intended by the content definition of the group of medical experts. In general, we recommend not to use Questionnaire and QuestionnaireResponse FHIR profiles in cases where the information to be represented can be modeled by using more general, interoperable concepts and FHIR resources.

The challenges of creating and harmonizing COVID-19 data sets are not unique to our work, and although initiatives, such as the Clinical Data Interchange Standards Consortium (CDISC), have released guidance on how to represent COVID-19 research data in a standardized format [46], the actual selection of the relevant biomedical concepts to be represented is left to the implementers. We explicitly selected the data elements for COVID-19–related patient research that are the most relevant for further characterizing patients with respect to research in infectious diseases (with a focus on immunization), pediatrics, and cardiology. However, we recognize the need for ongoing collaboration and standardization efforts to maximize interoperability and facilitate data sharing and analysis. Such efforts include integrating the GECCO data set with other COVID-19–related data sets and standards, both within and between countries. For example, we are currently harmonizing the GECCO data set with the ORCHESTRA (Connecting European Cohorts to Increase Common and Effective Response to SARS-CoV-2 Pandemic) project, which intends to create a harmonized and standardized data set for a pan-European cohort for COVID-19 research [40]. To facilitate the mapping of the data items that were developed in our work and represented in HL7 FHIR to the CDISC Study Data Tabulation Model standard, the organizations behind the two standards have collaboratively developed a comprehensive implementation guide, thereby enabling mapping between the different standards, ensuring compatibility, and facilitating interoperability across systems [47]. Moving forward, we encourage developers of tools and resources to facilitate the mapping and harmonization of different data standards, and we look forward to continued collaboration with the wider research community to address these challenges and advance COVID-19 research.

Conclusion
We herein introduce the development workflow and the resulting data set definitions for GECCO extension modules for the immunization, pediatrics, and cardiology domains. We have defined and implemented a workflow in which interdisciplinary
teams of medical domain experts, medical information specialists, and FHIR developers closely collaborate in an iterative, consensus-based fashion for the successful development of useful and interoperable data set definitions. This workflow may serve as a blueprint for further data set definition projects, such as the further development of data set definitions for extending the GECCO core data set. The extension modules described in this work have been validated and published. Their implementation and active use are anticipated in the context of current nationwide COVID-19 research networks in Germany.

Acknowledgments
The NAPKON (“Nationales Pandemie Kohorten Netz”; German National Pandemic Cohort Network) project is funded under a scheme issued by the Network of University Medicine (Nationales Forschungsnetzwerk der Universitätsmedizin [NUM]) by the Federal Ministry of Education and Research of Germany (Bundesministerium für Bildung und Forschung [BMBF]; grant number 01KK2021). The funding body did not take a role in the design of the study, in the development of the data set, or in the writing of the manuscript. We thank Yannick Börner for his valuable contribution to the definition of the Fast Healthcare Interoperability Resources (FHIR) profiles. We thank all members of the subject- and organ-specific working groups.

Data Availability
The implementation guides for the three extension modules have been published on GitHub pages [28-30]. The source Fast Healthcare Interoperability Resources Shorthand (FSH) files have been published on GitHub [31-33]. Data set descriptions can be found on ART-DECOR [34-36].

Authors' Contributions
All authors contributed to the development of the extension modules. GL, TH, SB, LR, JS, AB, and ST performed terminology mapping, FHIR profiling, and critical review of the concept and resource mappings. TH, SB, and LR defined the data sets in ART-DECOR. DH, FK, LES, FE, NT, RB, AF, and MD developed and compiled the list of data items for the data sets. SR, LL, and MU coordinated the project and the consensus finding process within and between working groups. JJV, CvK, and ST conceived the work. GL drafted the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest
ST is the vice chair of Health Level Seven International (HL7) Germany. The other authors declare that they have no competing interests.

Multimedia Appendix 1
Supplementary tables.

References


11. SNOMED International. URL: www.snomed.org/ [accessed 2022-03-16]


15. SNOMED International. URL: www.snomed.org [accessed 2022-03-16]

16. SIMPLIFIER.NET. Medizininformatik Initiative. URL: simplifier.net/organisation/koordinationsstellemit [accessed 2022-03-15]

17. SIMPLIFIER.NET. Kassenärztliche Bundesvereinigung (KBV). URL: simplifier.net/organisation/kassenarztlbundesvereinigungkvb [accessed 2022-03-15]

18. HL7 International. FHIR shorthand. URL: hl7.org/fhir/uv/shorthand/ [accessed 2022-04-25]


23. HL7 International. Implementation guide registry. URL: fhir.org/ig/registry/ [accessed 2022-11-3]


27. COVIM. COVIM – Collaborative & Immunity Platform of the NUM. URL: covim-netzwerk.de/ [accessed 2022-03-16]


30. NAPKON. NAPKON vaccination module. URL: bih-cei.github.io/napkon-vaccination/ [accessed 2022-03-15]


34. art-decor.org. NAPKON cardiology module. URL: art-decor.org/art-decor/decor-datasets-covid19?conceptId=2.16.840.1.113883.3.1937.777.53.1.2&effectiveDate=2020-08-12T00%3A00%3A00&conceptId=2.16.840.1.113883.3.1937.777.53.2.250&conceptEffectiveDate=2021-02-16T13%3A25%3A43&language=en-US [accessed 2022-03-16]

35. art-decor.org. NAPKON pediatrics module. URL: art-decor.org/art-decor/decor-datasets-covid19?conceptId=2.16.840.1.113883.3.1937.777.53.1.2&effectiveDate=2020-08-12T00%3A00%3A00&conceptId=2.16.840.1.113883.3.1937.777.53.2.12&conceptEffectiveDate=2020-09-18T09%3A20%3A12&language=en-US [accessed 2022-03-16]

36. art-decor.org. NAPKON vaccination module. URL: art-decor.org/art-decor/decor-datasets-covid19?conceptId=2.16.840.1.113883.3.1937.777.53.1.2&effectiveDate=2020-08-12T00%3A00%3A00&conceptId=2.16.840.1.113883.3.1937.777.53.2.45&conceptEffectiveDate=2021-08-25T12%3A45%3A26&language=en-US [accessed 2022-03-15]


Abbreviations

CDISC: Clinical Data Interchange Standards Consortium

CODEX+: Collaborative Data Exchange and Usage

COVIM: Collaborative Immunity Platform of the Netzwerk Universitätsmedizin

FAIR: Findable, Accessible, Interoperable, Reusable

FHIR: Fast Healthcare Interoperability Resources

FSH: Fast Healthcare Interoperability Resources Shorthand

GECCO: German Corona Consensus

HL7: Health Level Seven International

LOINC: Logical Observation Identifiers Names and Codes

NAPKON: Nationales Pandemie Kohorten Netz

ORCHESTRA: Connecting European Cohorts to Increase Common and Effective Response to SARS-CoV-2 Pandemic

SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms

© Gregor Lichtner, Thomas Haese, Sally Brose, Larissa Röhrig, Liudmila Lysyakova, Stefanie Rudolph, Maria Uebe, Julian Sass, Alexander Bartschke, David Hillus, Florian Kurth, Leif Erik Sander, Falk Eckart, Nicole Toepfner, Reinhard Berner, Anna Frey, Marcus Dörr, Jörg Janne Vehreschild, Christof von Kalle, Sylvia Thun. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 18.7.2023. 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.
Data Analysis of Physician Competence Research Trend: Social Network Analysis and Topic Modeling Approach

So Jung Yune†, PhD; Youngjon Kim‡, PhD; Jea Woog Lee§, PhD

†Department of Medical Education, Pusan National University, Busan, Republic of Korea
‡Department of Medical Education, Wonkwang University School of Medicine, Iksan, Republic of Korea
§Intelligence Informatics Processing Lab, Chung-Ang University, Seoul, Republic of Korea

†‡§all authors contributed equally

Corresponding Author:
Jea Woog Lee, PhD
Intelligence Informatics Processing Lab
Chung-Ang University
84, Heukseok-ro, Dongjak-gu,
Seoul, 06974
Republic of Korea
Phone: 82 10 5426 7318
Email: vyizeuks@cau.ac.kr

Abstract

Background: Studies on competency in medical education often explore the acquisition, performance, and evaluation of particular skills, knowledge, or behaviors that constitute physician competency. As physician competency reflects social demands according to changes in the medical environment, analyzing the research trends of physician competency by period is necessary to derive major research topics for future studies. Therefore, a more macroscopic method is required to analyze the core competencies of physicians in this era.

Objective: This study aimed to analyze research trends related to physicians’ competency in reflecting social needs according to changes in the medical environment.

Methods: We used topic modeling to identify potential research topics by analyzing data from studies related to physician competency published between 2011 and 2020. We preprocessed 1354 articles and extracted 272 keywords.

Results: The terms that appeared most frequently in the research related to physician competency since 2010 were knowledge, hospital, family, job, guidelines, management, and communication. The terms that appeared in most studies were education, model, knowledge, and hospital. Topic modeling revealed that the main topics about physician competency included Evidence-based clinical practice, Community-based healthcare, Patient care, Career and self-management, Continuous professional development, and Communication and cooperation. We divided the studies into 4 periods (2011-2013, 2014-2016, 2017-2019, and 2020-2021) and performed a linear regression analysis. The results showed a change in topics by period. The hot topics that have shown increased interest among scholars over time include Community-based healthcare, Career and self-management, and Continuous professional development.

Conclusions: On the basis of the analysis of research trends, it is predicted that physician professionalism and community-based medicine will continue to be studied in future studies on physician competency.

(JMIR Med Inform 2023;11:e47934) doi:10.2196/47934

KEYWORDS
physician competency; research trend; competency-based education; professionalism; topic modeling; latent Dirichlet allocation; LDA algorithm; data science; social network analysis
Introduction

Background

Medical publications began defining competencies in the 1970s [1-3]. Physician competency refers to the essential qualities that a physician should possess. The search for physician competency begins with the question of what it means to be a physician. Competency connotes various ideas, such as which physician traits are desired by society and what supports and promotes this transition of identity. Competency entails the concept of the physician as a professional, what the physician can do, and how the physician approaches their practice [4]. In summary, competency is considered a complex set of behaviors built on the components of knowledge, skills, attitudes, and competence as a personal ability [5]. Competency in this study is the core competency required to successfully perform a physician’s job and includes knowledge, skills, and attitude, regardless of the specific major.

In the medical profession, the competency theme began by pursuing ways for the medical circle to improve the performance of health care workers. The issue arose in response to the growing demands of medical consumers and society as a neoliberalistic ideology spread in the 1970s and consumers’ awareness of their rights increased. In 1972, the American Academy of Pediatrics discussed physician competency by publishing a foundation for evaluating pediatricians’ competency [3]. In 1978, the World Health Organization found the cause of declines in medical service quality to be inadequate education and attempted to improve health care providers’ competencies through Competency-Based Curriculum Development in Medical Education [6]. Since then, this approach has influenced medical education globally, starting with the Royal College of Physicians and Surgeons in Canada and the Accreditation Council for Graduate Medical Education in the United States. Similar programs have been established worldwide, influencing strategies for global human resources and international partnerships for medical training [7]. These include the Outcome Project of the US Accreditation Council for Graduate Medical Education, General Medical Council’s Tomorrow’s Doctors [8-10], Scottish Doctor [11], and Canadian CanMEDS framework [12].

For physicians, competency varies depending on the clinical, cultural, and geographical context [13]. In medical practice, the perception of a medical professional’s competent role changes continuously over time. In the early 19th century, physicians applied ointments and drew blood but did not deliver babies. In the 21st century, physicians are required to use advanced technologies and artificial intelligence in medical surgeries [14,15]. The ability to use technologies required by future health care systems is a challenge for physicians. However, at the same time, communication with patients and colleagues and interprofessional teamwork are essential human skills, and personal traits, such as empathy, humility, compassion, emotional intelligence, and a passion for continuous learning are also emphasized. Varying levels of health care infrastructure over time [16]; social awareness of minority groups [17]; and occasional health care challenges, such as global pandemics [18], emphasize certain specific physician competencies. Physicians’ competency, initially discussed within the scope of their social accountability, includes changes in their roles from the perspective of patients, health care, and self-management [19] of their health and wellness [20]. Thus, physicians’ abilities reflect their social situation and demands. In addition, physicians’ core abilities are expected to change over time. Therefore, this study, which analyzed changes in physician competency by period, will help understand the social demands expected of physicians in each period and identify research topics that should be important in medical education in the future.

To date, studies on physician competency have focused on literature review [21-23] or meta-analyses [24,25]. However, systematic literature reviews have limitations in deriving comprehensively synthesized results because their analyses focus on narrow areas such as subjects, variables, environment, and intervention. In medical education, studies on competency often explore the acquisition, performance, and evaluation of particular skills, knowledge, or behaviors that constitute physician competency. Some of these studies examined patient-physician communication [26], risk management for emergency physicians, technical skills in robotic surgery in urological practice [27], and the instruction of medical staff [28]. Nevertheless, such approaches fail to convey the trends in physician competency research because they explore the essential medical skills for a specific task in a certain context. In addition, people may have different thoughts about the core competencies that physicians should possess. For example, a patient’s expectations of a physician’s ability may differ from a senior physician’s expectation of a junior physician’s ability. Therefore, a more macroscopic method is required to analyze the core competencies of physicians in this era.

Recently, new big data analysis techniques such as social network analysis and topic modeling have been used. These approaches have the advantage of organizing the knowledge structure context by forecasting the trajectory of change in research and future issues and revealing the correlation between concepts. Social network analysis involves detecting influential core keywords in a vast amount of text and showing the relationship between keywords, allowing researchers to comprehend the context intuitively [29,30]. Topic modeling detects hidden topics in text data, analyzes the association and distribution of each topic, and provides integrated information [31]. From a microscopic perspective, it identifies core topics and their relationships. From a macroscopic perspective, it identifies the flow and context of core topics and the trend of topics by period [32]. Text network analysis and topic modeling are ideal approaches for analyzing trends in research on physicians’ competencies and accomplishing research objectives.

Objectives

The research problems based on the abovementioned research necessity are as follows:

1. Extract core keywords from physician competency studies and create a network
2. Examine the structure and characteristics of the network created based on physician competency studies.
3. Examine the main topics through topic modeling in physician competency studies.
4. Examine the trend of physician competency studies by topic based on time flow.

**Methods**

**Data Collection**

To examine the flow of academic research on physician competency, we selected related research articles as raw data, which were collected using NetMiner Biblio Data Collector (Cyram Inc). We selected 10 years (2011-2020) and collected 2164 articles on physician competency published in *Springer* using the following keywords: “(doctors or physicians) AND ((competence or competency or competencies) or (expertise or expert or proficiency) or (responsibility or accountability or liability or blameworthiness) or (profession or occupation or roles or duties or jobs) or (performance and practice) or professionalism)).” After eliminating 810 articles that overlapped or were unrelated, 1354 articles remained for our analysis. The research flow is illustrated in Figure 1.

*Figure 1.* Flow of research procedures. TF-IDF: term frequency–inverse document frequency.
Data Preprocess

Because we could not use the original text in the preselected articles in the analysis, we processed the sentences into separate words as units of analysis. The collected papers consisted of natural language sentences such as theories, knowledge, and opinions. However, the sentences cannot be used directly in the analysis. Therefore, steps must be taken to convert each sentence into an individual word that can be analyzed [33]. In this study, to extract nouns, adjectives, and verbs, we used the morpheme-refining function of the NetMiner program and extracted 5384 words from the titles, abstracts, and keywords of the research articles. Preprocessing was performed to convert these words into analyzable keyword data.

Nouns that were unsuitable for analysis were eliminated. First, we removed words that appeared 5 times or keywords that did not appear in more than 20 of the 1354 articles. In the final stage, we eliminated infrequently appearing words (<5 times) and extremely common words that appeared frequently in all papers (term frequency-inverse document frequency < 0.05) [34]. After preprocessing, 272 words were extracted.

Social Network Analysis

In this study, social network analysis was used to examine the knowledge structure and characteristics through keyword extraction and network generation from physician competency studies, designated as research problems 1 and 2. In addition, the roles of keywords in the network were determined by assessing their importance using social network analysis techniques such as degree centrality, closeness centrality, and betweenness centrality. Social network analysis determines core nodes based on degree centrality, closeness centrality, and betweenness centrality. The degree centrality of a node increases with the number of nodes directly connected to it. Thus, the degree centrality indicates the influence of a node (keyword) based on the number of connected nodes. The betweenness centrality indicates the centrality of a node (keyword) between 2 other nodes (keywords). Betweenness centrality increases when the number of times a node appears on the shortest path between 2 other nodes increases. Keywords with high betweenness centrality control the information flow and exert a substantial influence on the overall connectivity of the network. To visually understand the positions and relationships among keywords, we used spring mapping from NetMiner 4.0. Spring mapping maximizes the characteristics of branching out the graph by placing the connected nodes closer and the disconnected nodes farther, using a simulated annealing technique to balance the 2 forms of distance.

Topic Modeling

To address research problem 3, we conducted topic modeling to explore the topic areas of physician competency research. Topic modeling was performed using keywords extracted from social network analysis. NetMiner 4.0 was used as the analysis program. Topic modeling is used to predict latent topics based on the association of multiple words in a text. Topic modeling extracts topics from research papers through keyword exploration, which helps recognize knowledge structures and patterns [35]. Knowledge structures are defined as a visualization of keyword clustering and a network map of how concepts in domain knowledge are interrelated [36]. This pattern is defined as the process of change in the knowledge structure over a period [37]. Topic modeling, as a big data analysis technique, provides a quantitative approach for identifying previously undiscovered macrotopic areas in physician competency research. Topics were extracted by applying a latent Dirichlet allocation (LDA) algorithm. The LDA algorithm is a probabilistic model for predicting hidden topics by analyzing the distribution of words observed in a document. It is useful for reducing the data size and producing consistent topics.

Figure 2 shows a graphical model of LDA. The boxes in the figure are “plates,” which represent duplicates. The outer plate represents the document (M) and the inner plate represents repeatedly selected topics (z) and words (w) within the document (N). “|” indicates the distribution of topics in the document. Both “α” and “β” are hyperparameters that indicate Dirichlet distribution. LDA cannot be directly used to determine the number of topics, so the third hyperparameter is the “number of topics” the algorithm will discover. The probability calculation formula is as follows [38]:

\[
\alpha \quad \beta \\
\theta \\
z \\
w \\
N \\
M
\]

Figure 2. Illustration for the conceptual model of latent Dirichlet allocation. α: a parameter that represents the Dirichlet prior for the document topic distribution; β: a parameter that represents the Dirichlet for the word distribution; θ: a vector for topic distribution over a document d; z: a topic for a chosen word in a document; w: specific words in N; M: document length; N: number of words in the document.
To increase the accuracy of the results, appropriate Cronbach α values, β values, number of topics, and keywords should be determined [39]. To evaluate how well each keyword described each topic, we used one of the topic consistency metrics, the silhouette coefficient. The silhouette coefficient is an indicator that evaluates how well keywords, which are components of a topic, are classified [40]. A value closer to 1 indicates that the keywords within each cluster are well formed. Moreover, good clustering means that similarity to other topics is low and keywords within the same cluster describe a topic well. For topic \( t \) characterized by a higher-order word set \( W_t \) (any word whose probability exceeds a predefined threshold or a fixed number of high-order words), the consistency formula is defined as follows [41]:

\[
\text{silhouette coefficient} = \frac{b_t - a_t}{\max(a_t, b_t)}
\]

In this study, we set the term frequency–inverse document frequency threshold value at 0.5 and word length at 2. We used a silhouette coefficient to calculate the optimal values for Cronbach α, β, and the number of topics. A silhouette coefficient (or score) closer to 1 had higher explanatory power, validating Cronbach α, validating β, and the number of topics and descriptions of the object in the topic model were well matched. We also used a silhouette-clustering configuration. To determine the optimal number of topics, we conducted a comprehensive analysis by varying the number of topics from 5 to 30 and exploring Cronbach α values ranging from .01 to .99 as well as β values ranging from .01 to .99. The silhouette coefficient was used as the evaluation criterion. Our findings revealed that the highest silhouette coefficient of 0.782 was achieved using a Cronbach α value of .89, a β value of .97, and 6 topics. Subsequently, we proceeded with topic modeling using the identified parameters.

Analysis of Change in Topics by Period

To address research problem 4—the change in topics in physician competency research over time—the analysis was divided into 4 periods: 2011-2013, 2014-2016, 2017-2019, and 2020-2021. We divided the period into before and after the COVID-19 outbreak, and the researchers discussed and classified them. Subsequently, we analyzed how the percentages of each topic changed. To categorize the topics by checking the pattern of increased or decreased topics by period, we performed a linear regression analysis using SPSS (version 23.0; IBM Corp). We used the categorized periods as independent variables, and the percentage of each topic as the dependent variable. Following the linear regression analysis, we classified the keywords into 4 types based on the regression coefficient sign (+ or −) and the significance probability (\( P \) value): hot, warm, cool, and cold. If the coefficient is positive and the significance probability is ≤.05, it is classified as a “hot topic” with increasing research interest. Conversely, if the coefficient is negative and the significance probability is ≤.05, it is classified as a “cold topic” with decreasing research interest. Meanwhile, the clusters that were either positive or negative with no statistical significance and with a significance probability of ≥.05 were classified as “warm” and “cool” topics, respectively [42].

Ethics Approval

This study was conducted after obtaining approval from the Medical Research Ethics Review Committee of Chungnam National University Hospital (CNUH 2021-02-025).

Results

Core Keywords From Physician Competency Studies

Between 2011 and 2020, the words that appeared most frequently in physician competency studies were knowledge (604 times), hospital (598 times), family (597 times), job (573 times), guideline (491 times), management (482 times), and communication (443 times). The words that appeared in most studies were education (n=256), model (n=243), knowledge (n=238), and hospital (n=234). Table 1 presents the 25 words with the highest frequency and the number of articles in which they appeared.
Table 1. High-ranking keywords by frequency in research.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Keyword</th>
<th>Frequency, n</th>
<th>Articles in which it appears, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Knowledge</td>
<td>604</td>
<td>238</td>
</tr>
<tr>
<td>2</td>
<td>Hospital</td>
<td>598</td>
<td>234</td>
</tr>
<tr>
<td>3</td>
<td>Family</td>
<td>597</td>
<td>161</td>
</tr>
<tr>
<td>4</td>
<td>Job</td>
<td>573</td>
<td>111</td>
</tr>
<tr>
<td>5</td>
<td>Guideline</td>
<td>491</td>
<td>164</td>
</tr>
<tr>
<td>6</td>
<td>Management</td>
<td>482</td>
<td>206</td>
</tr>
<tr>
<td>7</td>
<td>Communication</td>
<td>443</td>
<td>154</td>
</tr>
<tr>
<td>8</td>
<td>Education</td>
<td>437</td>
<td>256</td>
</tr>
<tr>
<td>9</td>
<td>Model</td>
<td>425</td>
<td>243</td>
</tr>
<tr>
<td>10</td>
<td>Assessment</td>
<td>407</td>
<td>150</td>
</tr>
<tr>
<td>11</td>
<td>Attitude</td>
<td>388</td>
<td>151</td>
</tr>
<tr>
<td>12</td>
<td>Information</td>
<td>381</td>
<td>213</td>
</tr>
<tr>
<td>13</td>
<td>Health care</td>
<td>379</td>
<td>196</td>
</tr>
<tr>
<td>14</td>
<td>Experience</td>
<td>368</td>
<td>224</td>
</tr>
<tr>
<td>15</td>
<td>Medication</td>
<td>356</td>
<td>64</td>
</tr>
<tr>
<td>16</td>
<td>Intervention</td>
<td>332</td>
<td>152</td>
</tr>
<tr>
<td>17</td>
<td>Cancer</td>
<td>330</td>
<td>89</td>
</tr>
<tr>
<td>18</td>
<td>Disease</td>
<td>313</td>
<td>150</td>
</tr>
<tr>
<td>19</td>
<td>Change</td>
<td>276</td>
<td>147</td>
</tr>
<tr>
<td>20</td>
<td>Need</td>
<td>268</td>
<td>196</td>
</tr>
<tr>
<td>21</td>
<td>Development</td>
<td>255</td>
<td>170</td>
</tr>
<tr>
<td>22</td>
<td>Behavior</td>
<td>253</td>
<td>87</td>
</tr>
<tr>
<td>23</td>
<td>Barrier</td>
<td>238</td>
<td>97</td>
</tr>
<tr>
<td>24</td>
<td>Evidence</td>
<td>234</td>
<td>158</td>
</tr>
<tr>
<td>25</td>
<td>Practice</td>
<td>227</td>
<td>199</td>
</tr>
</tbody>
</table>

Social Network Analysis

Table 2 presents keyword degree, closeness, and betweenness centrality from physician competency studies. Each keyword and its degree are as follows. The higher the degree of the keyword, the stronger the influence in the network. The keywords with the highest degree and closeness centrality were, in order, model (0.988 and 0.988, respectively), education (0.981 and 0.982), experience (0.973 and 0.975), health care (0.973 and 0.975), hospital (0.973 and 0.975), and information (0.973 and 0.975). The keywords with the highest betweenness centrality were, in order, model (0.007), education (0.006), knowledge (0.006), management (0.006), and education (0.006).

In total, 29 words belonged to the top 30 keywords in all 3 centrality types. The keywords with a high degree centrality were also high in closeness centrality. However, the betweenness centrality of environment was higher than the degree and closeness centrality. Figure 3 shows the network map for centrality.
Table 2. High-ranking keywords by degree, closeness, and betweenness centrality.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Degree centrality</th>
<th>Closeness centrality</th>
<th>Betweenness centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Keyword</td>
<td>Degree</td>
<td>Keyword</td>
</tr>
<tr>
<td>1</td>
<td>Model</td>
<td>0.98881</td>
<td>Model</td>
</tr>
<tr>
<td>2</td>
<td>Education</td>
<td>0.98134</td>
<td>Education</td>
</tr>
<tr>
<td>3</td>
<td>Experience</td>
<td>0.97388</td>
<td>Experience</td>
</tr>
<tr>
<td>4</td>
<td>Health care</td>
<td>0.97388</td>
<td>Health care</td>
</tr>
<tr>
<td>5</td>
<td>Hospital</td>
<td>0.97388</td>
<td>Hospital</td>
</tr>
<tr>
<td>6</td>
<td>Information</td>
<td>0.97388</td>
<td>Information</td>
</tr>
<tr>
<td>7</td>
<td>Knowledge</td>
<td>0.97388</td>
<td>Knowledge</td>
</tr>
<tr>
<td>8</td>
<td>Management</td>
<td>0.97388</td>
<td>Management</td>
</tr>
<tr>
<td>9</td>
<td>Need</td>
<td>0.97388</td>
<td>Need</td>
</tr>
<tr>
<td>10</td>
<td>Development</td>
<td>0.96269</td>
<td>Development</td>
</tr>
<tr>
<td>11</td>
<td>Evidence</td>
<td>0.95522</td>
<td>Evidence</td>
</tr>
<tr>
<td>12</td>
<td>Change</td>
<td>0.95149</td>
<td>Change</td>
</tr>
<tr>
<td>13</td>
<td>Intervention</td>
<td>0.94776</td>
<td>Intervention</td>
</tr>
<tr>
<td>14</td>
<td>Family</td>
<td>0.93657</td>
<td>Family</td>
</tr>
<tr>
<td>15</td>
<td>Disease</td>
<td>0.93284</td>
<td>Disease</td>
</tr>
<tr>
<td>16</td>
<td>Guideline</td>
<td>0.92910</td>
<td>Guideline</td>
</tr>
<tr>
<td>17</td>
<td>Practice</td>
<td>0.92164</td>
<td>Practice</td>
</tr>
<tr>
<td>18</td>
<td>Assessment</td>
<td>0.91418</td>
<td>Assessment</td>
</tr>
<tr>
<td>19</td>
<td>Challenge</td>
<td>0.91418</td>
<td>Challenge</td>
</tr>
<tr>
<td>20</td>
<td>Attitude</td>
<td>0.91045</td>
<td>Attitude</td>
</tr>
<tr>
<td>21</td>
<td>Communication</td>
<td>0.90672</td>
<td>Communication</td>
</tr>
<tr>
<td>22</td>
<td>Evaluation</td>
<td>0.90299</td>
<td>Evaluation</td>
</tr>
<tr>
<td>23</td>
<td>Concern</td>
<td>0.88806</td>
<td>Concern</td>
</tr>
<tr>
<td>24</td>
<td>Influence</td>
<td>0.88060</td>
<td>Influence</td>
</tr>
<tr>
<td>25</td>
<td>Activity</td>
<td>0.86940</td>
<td>Activity</td>
</tr>
<tr>
<td>26</td>
<td>Improvement</td>
<td>0.86940</td>
<td>Improvement</td>
</tr>
<tr>
<td>27</td>
<td>Community</td>
<td>0.86194</td>
<td>Community</td>
</tr>
<tr>
<td>28</td>
<td>Barrier</td>
<td>0.85821</td>
<td>Barrier</td>
</tr>
<tr>
<td>29</td>
<td>Implementation</td>
<td>0.85821</td>
<td>Implementation</td>
</tr>
<tr>
<td>30</td>
<td>Cost</td>
<td>0.85448</td>
<td>Cost</td>
</tr>
</tbody>
</table>
Topic Modeling

Regarding the number of topics from physician competency studies, we decided on 6 topics (silhouette=0.782) by considering the silhouette coefficient and the validity of interpretation. The core keywords by topic are listed in Table 3. The top keywords in topic 1 were management, intervention, disease, cost, and medication. The top keywords in topic 2 were family, health care, information, community, and need. The high-ranking keywords in topic 3 were knowledge, attitude, cancer, guidelines, and barriers. The high-ranking keywords in topic 4 were, in order, hospital, job, burnout, model, gender, and intention. The top keywords in topic 5 were, in order, assessment, education, competency, development, and graduation. Finally, the high-ranking keywords in topic 6 were, in order, communication, consultation, experience, emergency, and feedback.

Topic groups were labeled based on high-ranking core keywords in terms of probability distribution: topic 1, Evidence-based clinical practice; topic 2, Community-based healthcare; topic 3, Patient care; topic 4, Career and self-management; topic 5, Continuous professional development; and topic 6, Communication and cooperation. On the basis of the nature of the topics, we divided the topic groups into 2 domains: those related to job competency and those related to personal competency. The job domain includes topics 1, 2, 3, and 6, and the personal domain includes topics 4 and 5. Figure 4 presents the results of visualizing the 7 networks of core keywords using a topic-keyword map.
Table 3. Core keywords by topic.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Domain and topic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Job</td>
</tr>
<tr>
<td></td>
<td>Evidence-based clinical practice</td>
</tr>
<tr>
<td>1</td>
<td>Management</td>
</tr>
<tr>
<td>2</td>
<td>Intervention</td>
</tr>
<tr>
<td>3</td>
<td>Disease</td>
</tr>
<tr>
<td>4</td>
<td>Cost</td>
</tr>
<tr>
<td>5</td>
<td>Medication</td>
</tr>
<tr>
<td>6</td>
<td>Diagnosis</td>
</tr>
<tr>
<td>7</td>
<td>Diabetes</td>
</tr>
<tr>
<td>8</td>
<td>Adherence</td>
</tr>
<tr>
<td>9</td>
<td>Guideline</td>
</tr>
<tr>
<td>10</td>
<td>Change</td>
</tr>
</tbody>
</table>

Note: PCP: primary care provider.

Figure 4. Topic modeling network: topic-keyword map.

Core Topics by Period

Table 4 presents the changes in topics over time and Table 5 lists the rate of each topic during this period. Over the past 10 years, each topic has been studied with similar weights. Communication and cooperation accounted for 18.61% (252/1354) of the articles, followed by Evidence-based clinical practice at 17.73% (240/1354) and Continuous professional development at 16.77% (227/1354). Regarding the domains, studies related to physicians’ job competencies accounted for 61.21% (910/1354), whereas those related to physicians’ personal competencies accounted for 32.79% (444/1354).
Table 4. Number of research on 6 topics by year (n=1354).

<table>
<thead>
<tr>
<th>Domain and topic</th>
<th>2011 (n=82)</th>
<th>2012 (n=107)</th>
<th>2013 (n=98)</th>
<th>2014 (n=152)</th>
<th>2015 (n=134)</th>
<th>2016 (n=120)</th>
<th>2017 (n=125)</th>
<th>2018 (n=141)</th>
<th>2019 (n=129)</th>
<th>2020 (n=167)</th>
<th>2021 (n=99)</th>
<th>Total (n=1354)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence-based clinical practice</td>
<td>15 (6.3%)</td>
<td>24 (10%)</td>
<td>24 (10%)</td>
<td>27 (11.3%)</td>
<td>22 (9.2%)</td>
<td>21 (8.8%)</td>
<td>25 (10.4%)</td>
<td>26 (10.8%)</td>
<td>25 (10.4%)</td>
<td>21 (8.8%)</td>
<td>10 (4.2%)</td>
<td>240 (100%)</td>
</tr>
<tr>
<td>Community-based healthcare</td>
<td>11 (5.2%)</td>
<td>13 (6.1%)</td>
<td>13 (6.1%)</td>
<td>22 (10.3%)</td>
<td>26 (12.2%)</td>
<td>22 (10.3%)</td>
<td>17 (8.8%)</td>
<td>18 (8.5%)</td>
<td>20 (9.4%)</td>
<td>29 (13.6%)</td>
<td>22 (10.3%)</td>
<td>213 (100%)</td>
</tr>
<tr>
<td>Patient care</td>
<td>13 (6.3%)</td>
<td>22 (10.7%)</td>
<td>6 (2.9%)</td>
<td>29 (14.2%)</td>
<td>12 (5.9%)</td>
<td>15 (7.3%)</td>
<td>18 (8.8%)</td>
<td>23 (11.2%)</td>
<td>14 (6.8%)</td>
<td>36 (17.6%)</td>
<td>17 (8.5%)</td>
<td>205 (100%)</td>
</tr>
<tr>
<td>Communication and cooperation</td>
<td>20 (7.9%)</td>
<td>21 (8.3%)</td>
<td>22 (8.7%)</td>
<td>25 (9.9%)</td>
<td>30 (12%)</td>
<td>21 (8.3%)</td>
<td>22 (8.7%)</td>
<td>21 (8.3%)</td>
<td>25 (9.9%)</td>
<td>24 (9.5%)</td>
<td>21 (8.3%)</td>
<td>252 (100%)</td>
</tr>
</tbody>
</table>

Table 5. The topic frequency and possession during the period (n=1354).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence-based clinical practice</td>
<td>63 (22%)</td>
<td>70 (17.2%)</td>
<td>76 (19.2%)</td>
<td>31 (11.6%)</td>
<td>240 (17.7%)</td>
</tr>
<tr>
<td>Community-based healthcare</td>
<td>37 (12.9%)</td>
<td>70 (17.2%)</td>
<td>55 (13.9%)</td>
<td>51 (19.2%)</td>
<td>213 (15.7%)</td>
</tr>
<tr>
<td>Patient care</td>
<td>41 (14.2%)</td>
<td>56 (13.8%)</td>
<td>55 (13.9%)</td>
<td>53 (19.9%)</td>
<td>205 (15.1%)</td>
</tr>
<tr>
<td>Communication and cooperation</td>
<td>63 (22.5%)</td>
<td>76 (18.7%)</td>
<td>68 (17.2%)</td>
<td>45 (16.9%)</td>
<td>252 (18.6%)</td>
</tr>
<tr>
<td>Career and self-management</td>
<td>40 (13.9%)</td>
<td>66 (16.3%)</td>
<td>70 (17.7%)</td>
<td>41 (15.4%)</td>
<td>217 (16%)</td>
</tr>
<tr>
<td>Continuous professional development</td>
<td>43 (15%)</td>
<td>68 (16.8%)</td>
<td>71 (18%)</td>
<td>45 (16.9%)</td>
<td>227 (16.8%)</td>
</tr>
</tbody>
</table>

Dividing the period into 3-year groups, Evidence-based clinical practice (63/240, 26.3%) and Communication and cooperation (37/213, 12.9%) were studied most often during the first period (2011-2013), whereas research on Community-based healthcare (41/205, 14.3%) and Career and self-management (40/217, 13.9%) was conducted on a small scale. Communication and cooperation was studied most often during the second period (2014-2016), but the weight decreased compared with the first period (from 22% to 18.7%).

The topics that increased in weight compared with the first period were Career and self-management (from 13.9% to 16.3%) and Continuous professional development (from 15% to 16.8%). During the third period (2017-2019), Evidence-based clinical practice (76/240, 19.2%) was studied the most, followed by Continuous professional development (71/227, 18%) and Career and self-management (70/217, 17.7%). During the fourth period, which covers the COVID-19 pandemic (2020-2021), Patient care (53/205, 19.9%) and Community-based healthcare (51/213, 19.2%) were the most studied. Compared with the third period, Patient care increased from 13.9% to 19.9%, and Community-based healthcare increased from 13.9% to 19.2%. Conversely, there were decreases in Evidence-based clinical practice (from 19.2% to 11.6%), Career and self-management (from 17.7% to 15.4%), and Continuous professional development (from 18% to 16.9%).

Regarding the overall possession during the first 3 periods before the COVID-19 pandemic, there was increased research interest in Career and self-management (from 13.9% to 16.3% to 17.7%) and Continuous professional development (from 15% to 16.8% to 18%) but decreased interest in Communication and cooperation (from 22% to 18.7% to 17.2%). In the fourth period, during the COVID-19 pandemic, there was increased research interest in Community-based healthcare (45/252, 16.9%) and Patient care (53/502, 19.9%) but decreased interest in Evidence-based clinical practice (31/240, 11.6%), Career and self-management (41/217, 15.4%), and Continuous professional development (45/227, 16.9%).

In terms of domains, studies on physicians’ personal competencies increased from the first to the third period (from 28.9% to 33% to 35.7%, respectively). However, it decreased...
to 32.3% between 2020 and 2021 after the COVID-19 outbreak. Studies on physicians’ job competency gradually decreased (from 71.2% to 67.8% to 64.3%), before increasing to 67.8% during the fourth period.

**Topic Characteristics by Period**

Table 6 presents the topic characteristics for each period. We performed linear regression analysis to examine the characteristics of the 6 topics. Three topics were classified as *hot topics* with a positive regression coefficient and statistical significance: topic 2 (B=1.315; \(t_9=2.621; P=.03\)), topic 4 (B=1.758; \(t_9=5.414; P=.001\)), and topic 5 (B=1.339; \(t_9=2.963; P=.02\)). We found no *cool* (negative regression coefficients and no statistical significance) or *cold* topics (negative regression coefficients and no statistical significance). Figure 5 shows the subject possession during this period.

<table>
<thead>
<tr>
<th>Domain and topic</th>
<th>B</th>
<th>β</th>
<th>t test (df)</th>
<th>P value</th>
<th>Durbin-Watson statistic</th>
<th>Topic type</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Job</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evidence-based clinical practice</td>
<td>0.388</td>
<td>.339</td>
<td>1.019 (9)</td>
<td>.34</td>
<td>1.535</td>
<td></td>
</tr>
<tr>
<td>Community-based healthcare</td>
<td>1.315</td>
<td>.680</td>
<td>2.621 (9)</td>
<td>.03</td>
<td>1.181</td>
<td>Hot</td>
</tr>
<tr>
<td>Patient care</td>
<td>1.248</td>
<td>.426</td>
<td>1.331 (9)</td>
<td>.22</td>
<td>2.960</td>
<td></td>
</tr>
<tr>
<td>Communication and cooperation</td>
<td>0.248</td>
<td>.251</td>
<td>0.733 (9)</td>
<td>.48</td>
<td>1.775</td>
<td></td>
</tr>
<tr>
<td><strong>Personal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Career and self-management</td>
<td>1.758</td>
<td>.886</td>
<td>5.414 (9)</td>
<td>.001</td>
<td>2.013</td>
<td>Hot</td>
</tr>
<tr>
<td>Continuous professional development</td>
<td>1.339</td>
<td>.723</td>
<td>2.963 (9)</td>
<td>.02</td>
<td>2.157</td>
<td>Hot</td>
</tr>
</tbody>
</table>

*Not available.*

**Figure 5.** Topic trend during the period.

**Discussion**

**Principal Findings**

In this study, we used social network analysis to examine keywords and their relationships in physician competency studies conducted over the past 10 years. Topic modeling identified the top 5 research topics, visualized the relationships among them, and described the research possession over time. Discussions on physicians’ competency arose in the 1990s because of the social atmosphere of consumerism, which demands accountability in all aspects of the profession. The medical and health care field emphasizes physicians’ roles and attitudes, reflecting the demands of medical consumers, such as citizens [43] and local communities [44]. Educational institutions have received demands to improve the curriculum considering educational outcomes [4,45]. Since the 1990s, some
countries have begun to define the competence of their physicians and specify their components [46]. After the “Project on the future global role of the physician in healthcare” of the World Federation for Medical Education in 2012, it has become more active in many countries [7]. As a goal of the medical community, a competency model for desirable physicians was constructed in the 2000s, centered on Canada [46], the United States [47], the United Kingdom [48], and Scotland [11]. Since then, scholars have actively studied the development of the curriculum and revision of content, reflecting this competency model [8,49,50]. The physician competency studies conducted since 2010, which we analyzed, are on the continuum of physician competency studies carried out over the past 20 years in a large framework.

The words that appeared most frequently in physician competency studies since 2010 are, in order, knowledge, hospital, family, job, guideline, management, and communication. The words that appeared in most studies were education, model, knowledge, and hospital.

Previously, a physician’s professional competence was defined as the habitual and careful use of communication, knowledge, skills, clinical reasoning, emotions, values, and reflection in everyday care for the benefit of individuals and communities [51]. The keywords frequently used in this study had a broad coverage, including the knowledge and context necessary for physicians to perform medical practice, a standardized framework necessary to meet social needs, and traits pertaining to physician groups or individuals.

In the network of top keywords, the keywords with high degree and closeness centrality values were model, education, experience, health care, and hospitals. Keywords with high betweenness centrality values were model, education, and knowledge. The keyword that emerged in the centrality analysis was change. Degree centrality indicates the number of times a node appears simultaneously with other nodes in a network. Closeness centrality connotes the distance between nodes within a network, and betweenness centrality connotes the role of a bridge between different nodes within a network. Network centrality analysis of keywords revealed that the 2 central nodes leading physician competency studies over the past 10 years were model and education. A model refers to a pattern, plan, or demonstration that illustrates the structure or work of an object, system, or concept. In competency studies, model began with a core competency model that explained what constitutes a physician. Later, scholars focused on different models essential for medical practice, such as specific clinical contexts, resident training, patient-physician communication, leadership [52], and health care management.

Competency-based education for medical students and residents refers to an educational method that uses the content or criteria derived from the previous competency model. In other words, a professional’s competency should be built gradually based on scientific knowledge, basic clinical skills, and moral development. Education and experience are essential for acquiring or maintaining expertise in a professional. Song et al [53] conducted a bibliometric analysis of medical expertise from 2010 to 2019. According to these studies, academic journals on medical education primarily include studies on expertise. Likewise, studies on physician competency over the past 10 years have examined methods for developing, implementing, and evaluating competency through education.

Through topic modeling, we identified 6 latent topics. Topic modeling is a researcher-centered content analysis method that identifies a specific pattern assumed to be latent in a document or text and derives a potentially meaningful topic. We also set topic names based on keywords derived from this study. On the basis of relevance of the topic, we divided it into 2 domains: job and personal competency. The job domain comprises Evidence-based clinical practice, Community-based healthcare, Patient care, and Communication and cooperation, whereas the personal domain comprises Career and self-management and Continuous professional development. The topics derived from topic modeling in this study cover the criteria for physician competency suggested in many countries. In Canada, the Canadian Medical Education Directions for Specialists [46] organized physician roles into 7 competencies: medical experts, communicators, collaborators, leaders, health advocates, scholars, and professionals. In the United Kingdom, “Good Medical Practice 2020” [54] describes physician competency in terms of 4 core competencies: knowledge, skill, and performance; safety and quality; communication, partnership, and teamwork; and trust maintenance. In the United States, the Accreditation Council for Graduate Medical Education laid out 6 core competencies [46] focusing on the areas of physician activities: patient care, professionalism, interpersonal and communication skills, medical knowledge, systems-based practice, and practice-based learning and improvement.

They also comprise the complex knowledge, skills, and attitudes that physicians must possess. Previous studies have emphasized the topics of physician competency research resulting from this study [5]. First, Evidence-based clinical practice is a core competency required for clinicians. It provides a framework for integrating research evidence into health care delivery [55], including patient history taking and analysis, physical examination, and diagnostic accuracy. The implementation of evidence-based practice principles has resulted in notable advances in improving the quality of delivered health care [56]. In addition, over the last 20 years, evidence-based practice has been increasingly integrated as a core component of undergraduate, postgraduate, and continuing education health programs worldwide [12]. Second, Community-based healthcare is increasingly emphasized in terms of the need for improvement from a broad cultural and institutional perspective to improve the quality of medical care [57]. Community-based healthcare has recently attracted more attention, as primary care is emphasized to solve social problems such as population aging and COVID-19 [58].

Third, the topic of Patient care means patient-centered care (PCC). PCC enhances health outcomes, such as improved patient satisfaction, behavior change, trust, patient adherence, providers' clinical accuracy, disease management plans, and active patient self-management [59]. Therefore, PCC is a crucial attribute of high-quality health care services [60]. Fourth, physicians’ interpersonal and communication skills have a significant impact on patient care. Furthermore, it is correlated with improved
health outcomes and quality [61]. Ineffective communication skills are associated with malpractice claims and suits [62] and medication errors [63]. Communication is a core clinical skill that can be taught and learned [64]. Moreover, interprofessional communication skills are essential competencies for medical students to become physicians [65].

Fifth, among the 2 topics belonging to the personal domain, Career and self-management is related to physicians’ burnout. Job satisfaction can affect physicians’ physical and mental illnesses, such as depression and burnout [66], and is related to patient safety and quality of care [67]. In addition, burnout syndrome is a major concern in occupational health [55]. Therefore, organizations should emphasize the importance of physicians’ self-care (rest, a healthy lifestyle, breaks, and sufficient sleep) and regular burnout screening [68]. Finally, Continuous professional development refers to the attitude of lifelong learning as a professional. Through professional development such as lifelong learning, medical specialists maintain their professional competence in addition to keeping track of and gaining advancing knowledge [69]. Furthermore, continuing professional development and lifelong learning are crucial for securing high-quality health care, patient safety, and societal trust in the health care system [70]. In other words, the subdomains of physician competency emphasized in many countries have been the main research topics over the past 10 years. Specifically, the main research topics include Evidence-based medicine (an explicit means for generating an important answerable question, interpreting new knowledge, and judging how to apply that knowledge in a clinical setting) [71]; Community-based healthcare (emphasizes social responsibility); Patient care (considering the patient’s condition and circumstances throughout the treatment process); Communication and cooperation with patients, families, and colleagues; Career and self-management; and Continuous professional development for maintaining competency as scholars and professionals.

An examination of the changes in research topics over the past 10 years revealed that more studies have been conducted on job competencies than on physicians’ personal competencies. Personal domain studies gradually increased from the first to the third period; however, after the COVID-19 outbreak (2020-2021), the number of job domain studies increased.

The most studied topic was Communication and cooperation. The topics that showed an increasing frequency and possession before the COVID-19 outbreak (first to third period) were Career and self-management and Continuous professional development; studies on Communication and cooperation showed a decreasing frequency and possession. Studies on Evidence-based clinical practice have also gradually decreased (first through second through fourth periods), except during the third period.

Most studies conducted before the COVID-19 outbreak covered physicians’ individual professionalism [10,72]. However, during the COVID-19 pandemic (fourth period), studies on Community-based healthcare and Patient care increased. This can be explained as follows: physicians’ social responsibility and community-centered care began to be emphasized during the COVID-19 pandemic, and the importance of care centered on patients and communities has resurfaced. Particularly during the COVID-19 pandemic, many problems threatened patients’ health because of the gap in health and medical care, despite individual physicians’ expertise and commitment. These challenging social situations highlight the importance of Community-based healthcare [73,74].

In times of crisis, the role of physicians can be broadened. For example, physicians have a duty not only to take care of their patients but also to protect them from infection, and thus take care of their families [18]. In addition, social interventions, such as school closures, affect the supply and demand for medical personnel. However, it is not easy to clarify whether a physician’s role in situations such as COVID-19 is regular duty. Nevertheless, COVID-19 has broadened the demand for physicians’ roles and competencies. This was also manifested in “hot topics” that have gradually increased over the past 10 years, such as Community-based healthcare, Career and self-management, and Continuous professional development. This indicates that topics related to Community-based healthcare are gradually becoming more important [73,74], as are topics related to the professionalism of individual physicians [10,51,53,72].

On the basis of the results of this study, the keywords that many researchers were interested in over the past 10 years were model and education. Therefore, they developed competency-based education and training systems at the hospital, university, and national levels. Consequently, many countries and training institutions, such as hospitals and universities, have developed and educated physicians with competency-based curricula. This was effective in cultivating physicians’ competency in responding to social needs.

The topics we should pay attention to are Community-based healthcare, Career and self-management, and Continuous professional development, which are research topics that have gradually increased with time. This indicates that the scope of physicians’ competencies has been studied more extensively and comprehensively than in the past. It is meaningful in that it defines physicians’ competency as the ability to develop into professional and social leaders, beyond just the ability necessary to perform a job.

The general public’s and patients’ expectations and consciousness of medical care are changing, and the medical system pursued by society is also changing. Consequently, perceptions of the roles of medicine and physicians are rapidly evolving. In line with these changes, research on the core competencies of physicians must be conducted using more detailed competencies and major fields.

Limitations
This study had some limitations. First, it was difficult to repeat the keyword refining process in the keyword network analysis. To eliminate researcher subjectivity, we have described the analytical procedure in this study. Second, the study period was not equally divided. The last period is short, covering only 6 months, between 2020 and 2021. Because we forecasted that the effect of the COVID-19 pandemic since December 2019
would be reflected between 2020 and the first half of 2021, we set these periods separately. Considering the gap between the time of undertaking and publishing this study, it is unreasonable to argue that the last period accurately reflected the pandemic. However, we believe that it is worth examining the effect of the pandemic because of its global nature, which surpasses its regional and cultural characteristics.

Third, similar to other studies, we explored articles published in English. Physician competency studies are influenced by cultural and social demands and changes in the context of medical services. Although we did not identify the regions in which the studies were published, similar studies on medical professionalism [52] and medical education [75] led us to believe that the English publications in this study were from North America, Europe, or parts of Asia. Future studies should analyze the differences in research trends in physician competency based on culture and region. Finally, this study investigated the core competencies for successfully performing a physician’s job, regardless of the specific major. Subsequently, we can research physician’s competencies for each major.

**Conclusions**

The top research topics on physician competency over the past 10 years are Evidence-based clinical practice, Community-based healthcare, Patient care, Career and self-management, Continuous professional development, and Communication and cooperation. The discussion of physician competency entails the establishment of a physician’s fundamental roles and competencies based on a constantly changing health care environment and the implementation of education from studying to competency acquisition. Studies on competency include discussions on the model physician desired by society, as well as the issue of wellness encompassing an individual physician’s job choice and quality of life.

The hot topics in physician competency studies conducted within the past 10 years are Community-based healthcare in the job domain and Career and self-management and Continuous professional development in the physician’s personal domain. These 2 areas are hot topics that have gradually gained interest over time.

**Conflicts of Interest**

None declared.

**References**


https://jmir.mir.org/2023/1/e47934

JMIR Med Inform 2023 | vol. 11 | e47934 | p.404

(page number not for citation purposes)


73. Lucey CR, Davis JA, Green MM. We have no choice but to transform: the future of medical education after the COVID-19 pandemic. Acad Med 2022 Mar 01;97(3S):S71-S81 [FREE Full text] [doi: 10.1097/ACM.0000000000004526] [Medline: 34789658]


Abbreviations

LDA: latent Dirichlet allocation
PCC: patient-centered care
Visual Analytics of Multidimensional Oral Health Surveys: Data Mining Study

Ting Xu¹, PhD; Yuming Ma², MSc; Tianya Pan², BSc; Yifei Chen², BSc; Yuhua Liu², PhD; Fudong Zhu³, PhD; Zhiguang Zhou², PhD; Qianming Chen³, PhD

¹Department of Stomatology, First Affiliated Hospital, Zhejiang University, Hangzhou, China
²School of Media and Design, Hangzhou Dianzi University, Hangzhou, China
³The Affiliated Stomatology Hospital Zhejiang University, Hangzhou, China

Corresponding Author:
Zhiguang Zhou, PhD
School of Media and Design, Hangzhou Dianzi University
Xueyuan Road #18
Hangzhou, 310018
China
Phone: 86 15957193211
Email: zhgzhou@hdu.edu.cn

Abstract

Background: Oral health surveys largely facilitate the prevention and treatment of oral diseases as well as the awareness of population health status. As oral health is always surveyed from a variety of perspectives, it is a difficult and complicated task to gain insights from multidimensional oral health surveys.

Objective: We aimed to develop a visualization framework for the visual analytics and deep mining of multidimensional oral health surveys.

Methods: First, diseases and groups were embedded into data portraits based on their multidimensional attributes. Subsequently, group classification and correlation pattern extraction were conducted to explore the correlation features among diseases, behaviors, symptoms, and cognitions. On the basis of the feature mining of diseases, groups, behaviors, and their attributes, a knowledge graph was constructed to reveal semantic information, integrate the graph query function, and describe the features of intrigue to users.

Results: A visualization framework was implemented for the exploration of multidimensional oral health surveys. A series of user-friendly interactions were integrated to propose a visual analysis system that can help users further achieve the regulations of oral health conditions.

Conclusions: A visualization framework is provided in this paper with a set of meaningful user interactions integrated, enabling users to intuitively understand the oral health situation and conduct in-depth data exploration and analysis. Case studies based on real-world data sets demonstrate the effectiveness of our system in the exploration of oral diseases.

(JMIR Med Inform 2023;11:e46275) doi:10.2196/46275

KEYWORDS
visual analytics; oral health data mining; knowledge graph; multidimensional data visualization

Introduction

Background

It is well known that oral health affects systemic health. Oral infections and inflammatory factors have been proven to be highly related to chronic diseases, such as cardiovascular and cerebrovascular diseases and diabetes mellitus [1]. In the field of clinical medicine, oral diseases can be prevented and treated by means of regular professional dental treatments and appropriate oral hygiene practices, which would do great favors for oral health, advance systemic well-being, and improve the quality of life.

As an effective way to investigate oral health status, oral health surveys can determine the frequency, intensity, and spread of oral diseases, such as oral behaviors, oral health cognition, and quality of life in a particular time frame [2]. On the basis of the analysis and mining of oral health surveys, we can obtain deeper
insights into the oral health status of individuals, understand oral diseases, and identify their impacting factors.

However, oral health surveys are always conducted from different perspectives and thus, are presented in the form of multiple dimensions. Traditional data mining methods always use simple statistical charts to visualize surveys, which are limited for the efficient and intuitive mining of deep-seated information. For example, it is difficult to observe the differences and similarities among different oral diseases. The oral health status across different areas and age groups lacks representative descriptions and intuitive comparisons. Thus, it is a difficult task to gain insights from multidimensional oral health surveys, especially for the exploration of relationships among diseases, behaviors, symptoms, and other attributes.

After a series of in-depth discussions with domain experts in the field of stomatology, it was concluded that a visualization system can deliver more comprehensive, interactive, and understandable information to users, which can further help them analyze oral health data rapidly and effectively. However, some challenges remain in implementing an oral health survey–oriented visualization system:

- **Challenge 1**: both oral diseases and related individuals have different general traits, which makes it difficult to present and compare the different characteristics of oral diseases in addition to their related populations (groups).
- **Challenge 2**: the occurrence and progression of oral diseases have their own rules, and unhealthy or careless behaviors will prompt the rise of oral diseases. Therefore, it is necessary to investigate the correlations between the diseases and behaviors.
- **Challenge 3**: oral health surveys include a rich set of data attributes, such as groups, diseases, and behaviors, which present various semantic relationships. It would be of great interest to explore the semantic relations from multidimensional attributes and provide an intelligent retrieval tool based on these relations.

To address the challenges, we developed a visualization framework for the visual analytics and deep mining of multidimensional oral health surveys. First, we designed a set of visualizations to depict the characteristics of diseases and groups combined with multidimensional attributes, such as the struct view, radar view, and cloud view, allowing the comparison of the different traits between various oral diseases and groups (challenge 1). We then designed a scatterplot matrix to analyze the correlation between diseases, behaviors, symptoms, and cognition based on group information, which can further help users discover the relationships among diseases, behaviors, and other attributes (challenge 2). Furthermore, a knowledge graph was designed to integrate diseases, groups, behaviors, and other information, allowing users to gain an overarching view of people with oral diseases. In addition, a query function was provided to conduct personalized retrieval, allowing users to obtain a more detailed understanding of human interests (challenge 3). A visualization framework was implemented to integrate a set of meaningful interactions, allowing users to obtain deeper insights into the patterns of oral diseases according to their requirements. Case studies based on real-world data sets were conducted to demonstrate the effectiveness of our system in visual analytics and deep mining of oral health surveys.

The major contributions of our work are summarized as follows:

- The characteristics of diseases and groups were depicted through portraits in light of multidimensional attributes, enabling users to intuitively and efficiently convey and disseminate information.
- A visualization framework was implemented to enable users to visually analyze and deeply mine the correlation features among oral diseases, behaviors, symptoms, and cognitions.
- A knowledge graph visualization was designed to generate structured knowledge containing semantics, supporting efficient queries on groups or attributes to grasp the semantic characteristics of multidimensional oral health surveys from macro and micro perspectives.

**Related Work**

This section covers 3 relevant topics: survey data visualization, multidimensional data visualization, and knowledge graph–based data mining.

**Survey Data Visualization**

Questionnaire survey is a key research tool to uncover and probe the existing states in many research domains [3]. Visualization provides analysts with deeper insights of information through visual recognition. Many researchers have applied data visualizations to realize hidden information capture and personalized exploration of questionnaire data. For example, Drapala et al [4] designed multidimensional data visualizations to explore surveys for the evaluation of information systems. Zhang et al [5] visualized the questionnaire data collected from patients and committed to predetermining orphan disease.

Surveys are widely used in medicine [6]. The World Health Organization provides guidelines for national oral surveys, enabling massive epidemiological studies and discussing survey principles. Powell et al [7] conducted a web-based questionnaire to determine the characteristics of health information seekers visiting a national health service website. O’Brien et al [8] conducted a web-based survey to investigate the use of disability and rehabilitation services among Canadian adults living with HIV. Aggarwal et al [9] piloted a large number of patient samples to explore the patients’ views on using their health data in artificial intelligence research. Nakamura et al [10] compared clinicians’ and patients’ perspectives on treating the symptoms of acute cerebral hemorrhage using survey data.

**Multidimensional Data Visualization**

Multidimensional data visualization [11] aims to express complex data in a visually intuitive format, using interactive elements to enable users’ comprehension of the correlation among various dimensions of the data. With the development of science and technology, multidimensional data have been reflected in a variety of fields. The oral surveys used in this study were multidimensional data with attributes, such as diseases, regions, ages, and behaviors. Currently, multidimensional data visualization includes spatial mapping, glyph [12], small multiples [13], and other methods. Examples
of spatial mapping include scatterplot matrices [14, 15], parallel coordinates [16, 17], table lenses [18], pixel charts [19], and dimensionality reduction [20, 21]. Scatterplot matrices and parallel coordinates are the 2 most widely used multidimensional data visualization strategies. The scatterplot matrix presents high-dimensional data using scatterplots, arranging them based on attributes. This mapping from multidimensional to 2D space helps identify correlations, clusters, outliers, and other notable characteristics. It is a valuable tool for exploring and analyzing complex data sets. Parallel coordinates use a series of parallel axes to represent each variable dimension of high-dimensional data, with the position along each axis corresponding to the variable’s value.

In addition to the conventional multidimensional visualizations mentioned above, users can use data portraits [22] to define and describe the various attributes of objects. Data portraits provide a more tangible representation of multidimensional data, allowing for a condensed and effective perception of information panoramas; for example, Xiong and Donath [23] proposed a novel graphical representation based on users’ past interactions, encoding people’s data with flower and garden metaphors. He et al [24] used accounting indexes to draw the data portrait of the value creation index of all 17 industries, by means of which the characteristics of various industries under COVID-19 can be captured. In this study, we applied data portraits to depict diseases and groups of different regions, genders, and ages.

Knowledge Graph–Based Data Mining

The knowledge graph, introduced by Google in 2012 to refine its search engine, is a typical multilateral relational graph comprising entities and relationships [25]. It serves as a semantic network that reveals the connections between various elements. Knowledge extraction [26], knowledge fusion [27], and knowledge reasoning [28] are the fundamental components involved in constructing a knowledge graph. Knowledge extraction is the process of extracting valuable structured information from large-scale text data, where entity extraction [29] refers to identifying specific entity objects in the text, whereas relation extraction involves extracting the associations and connections between entities. Knowledge fusion involves leveraging technologies such as information extraction, entity alignment, and relationship linking to integrate knowledge from multiple knowledge graphs. This integration results in a more comprehensive, consistent, and accurate knowledge system that enhances knowledge discovery, inference, and application. Knowledge reasoning can generate new factual conclusions by using entity and relation information, thereby expanding the knowledge graph. This process can be categorized into 3 types: logical rule–based reasoning [30], distributed feature representation–based reasoning [31], and deep learning–based reasoning [32].

The data derived from knowledge reasoning can be leveraged in a variety of downstream tasks related to knowledge graphs, such as recommendation systems [33], question answering [34], and information retrieval [35]. For instance, Li et al [36] introduced KG4Vis, a knowledge graph–based visual recommendation method that learns the embedding of knowledge graph entities and relations to capture ideal visual rules. Sousa and Couto [37] provided a new system, named K-BiOnt, by integrating knowledge graphs into biomedical relation extraction, improving the system’s ability to identify true relations. Tang et al [38] proposed an intelligent question-answering search system for electric power domain knowledge. The system uses knowledge reasoning to retrieve and analyze information accurately and presents the query results in a visual format. Latif et al [39] developed a visualization system, VisKonnect, to analyze the intertwined lives of historical figures according to the events they participated in through a knowledge graph.

Methods

Oral health data are introduced in this section. A series of analytical tasks are then defined following a thorough discussion with dental experts. Further presentation of the pipeline of our visualization system is encouraged, with the goal of completing the desired analysis tasks.

Data Description

In this study, the real-world data set was obtained from the Oral Health Status Survey and Prevention of Common Diseases in Zhejiang Province [40]. The survey covered several areas, including Jianggan, Hangzhou; Yuyao, Ningbo; Luqiao, Taizhou; Wenling, Taizhou; Wuyi, Jinhua; and Liandu, Lishui. The respondents were from 5 age groups: 3 to 5 years, 12 to 15 years, 35 to 44 years, 55 to 64 years, and 65 to 74 years, representing both urban and rural communities. The data set depicts the oral health status of individuals in 5 age groups in these 6 regions as well as behaviors, symptoms, and cognition associated with oral health. In total, 17 diseases, 14 behaviors, 10 symptoms, and 11 cognitions were considered as research qualities after sorting.

Ethical Considerations

As the data used in the study were deidentified, no ethical approval was sought.

Task Analysis

After detailed discussions with domain experts in the form of structured interviews, we developed a list of analytical tasks for the visual analysis of oral health based on oral health survey reports.

Task 1: How Can the Characteristics of Various Oral Diseases Be Described and Compared?

There are many types of oral diseases, including caries, periodontal disease, and oral mucosal disease. Unfortunately, some simple traditional statistical analyses struggle to uncover the underlying characteristics of the various diseases behind the data. Is a disease, for example, more likely to occur in men or women? At what age group may a malady be more likely to happen? Intuitive and efficient induction will play a vital role in medical research as well as in the formulation of preventive measures.
Task 2: How Can We Describe and Compare the Characteristics of Oral Diseases Among Different Groups?

Different age, region, and gender groups exhibit distinct overall characteristics in terms of prevalence and related attributes. Analyzing various groups based on disease, behavior, symptoms, cognition, and other dimensions can address the limitations of traditional summary evaluations. It enables us to grasp the specific requirements of different groups, aids in the efficient allocation of medical resources to enhance medical service quality, and provides a more comprehensive and precise depiction of the overall disease situation and characteristics of individuals.

Task 3: How Can the Association Among Oral Diseases, Behaviors, Symptoms, and Cognition Be Explored and Presented?

Each disease has its own set of rules regarding its occurrence and development, and it is often the case that bad behavior or carelessness can contribute to the likelihood of developing a disease. Is smoking associated with gum bleeding? Is tooth loss associated with food restriction symptoms? Grasping the relationship between oral diseases, behaviors, cognitions, and symptoms and how these factors interact with each other is a nontrivial task that requires the analysis of intricate and abstract data.

Task 4: How Can Semantic Information of Data Be Revealed in Interpretable Insight and Offer Assistance for Personalized Investigation?

How can we enhance and present the relationship between diseases, behavior, and other factors discovered after mining the correlation? How can users swiftly grasp the traits of specific diseases or groups at a detailed level? In addition, how can users gain a comprehensive understanding of the overall semantic context encompassing diseases, groups, behaviors, and other attributes from a macro perspective? Using effective visualizations can substantially aid users in comprehending and preventing oral diseases, promoting health awareness, and fostering healthy coping strategies.

System Overview

In this study, we developed a multidimensional survey visualization system for oral health that enables users to perceive the characteristics and patterns of oral diseases. Figure 1 shows the pipeline of the system to illustrate the design and implementation of the visualization framework.

Figure 1. The pipeline of our visualization system.
data, focusing on portrait analysis, correlation analysis, and semantic analysis. Various types of views, including the struct view, radar view, and cloud view, are used to depict the data portraits of diseases and groups. A scatterplot matrix view can be used to examine the correlation of attributes between groups and analyze the differences between groups. A graph view is used to link groups, diseases, behaviors, symptoms, and cognitions, thereby uncovering the semantic associations. Therefore, the system’s effectiveness in revealing multidimensional oral health surveys is evaluated through case studies, user studies, and expert interviews. In addition, a rich visual interface and user-friendly interactions are provided for users to explore the multidimensional oral health data in depth.

Visual Exploration

We developed a user-friendly visualization system to help users observe oral disease characteristics, disease-behavior correlations, and the semantic information of diseases and multidimensional attributes. This system includes a control panel (Figure 2A), group view (Figure 2B), scatterplot matrix view (Figure 2C), struct view (Figure 2D), graph view (Figure 2E), and radar view (Figure 2F), offering user-friendly interaction.

Figure 2. The visualization interface for a multidimensional oral health survey. (A) A control panel enabling users to load data sets and select attribute labels for different groups. (B) The group view to show the composition features and allow users to select groups. (C) The scatterplot matrix view to reveal the correlation features of attributes. (D) The struct view to present the characteristics of oral diseases. (E) The graph view illustrating the semantic knowledge of attributes such as population and disease. (F) The radar view with 2 cloud views to reveal the characteristics of different groups on attributes such as diseases and behaviors. AE: ashamed to eat; AL: attachment loss; BCGI: Bacteria can cause gum inflammation; BCTD: Bacteria can cause tooth decay; BGNBT: Bleeding gums are normal when brushing teeth; BT: brush the teeth; BTUPBG: Brushing teeth is useless in preventing bleeding gums; DC: difficulty chewing; CW: communication worry; DRoot: decayed teeth root due to caries; DROOT: decayed teeth root due to caries; DT: decayed tooth; DW: dietary worry; ESCTD: Eating sugar can cause tooth decay; FCPT: Fossa closure can protect the teeth; FDNPT: Fluoride does not protect teeth; FHE: father’s highest education; FL: food limitation; FRoot: fill teeth root due to caries; FROOT: fill teeth root due to caries; ft: fill deciduous tooth due to caries; FT: fill tooth due to caries; GB: bleeding gums index; HE: highest education; HS: hinder to speak; MHE: mother’s highest education; MT: missing tooth; OC: only child; OHIL: Oral health is important to life; PDD: periodontal pocket depth; PDL: periodontal pocket depth 4–6mm; PW: pronunciation worry; ROCF: Regular oral check-ups are important; SD: swallowing discomfort; ST: Time since the last dental visit; SW: sleep worry; TC: Tooth condition is innate, not acquired; TS: tooth sensitivity; TW: toothwash within 12 months; UT: Use the toothpick.

Data Portrait Analysis

Struct View

To visually and effectively depict different diseases, we designed the struct view. Disease profiles are presented as large rectangles containing smaller rectangles, categorized by area, gender, age, and city. Each small rectangle’s width represents the prevalence rate of the disease, whereas its color is randomly chosen. Clicking a small rectangle on the screen will display the region, gender, age, and urban and rural areas, along with the prevalence value. In total, 17 diseases were identified. To accommodate the limited screen space and enhance visual presentation, users can use sliding blocks.

Radar View

Radar view is a widely used metaphor in visualization. It allows for the presentation and comparison of group characteristics based on area, age, and gender. We developed a radar view specifically for this purpose. When the user selects a group, the radar view displays the disease attribute values in the specified region and age group for both men and women. Each axis maps the number of teeth with a type of oral disease to comprehend and compare the oral disease status of men and women in a fixed region and fixed age from a single radar view. By choosing different radar views, we can also compare different age groups and regional groups from the list.
These relations can be broken down into 4 categories: groups and diseases, behaviors, symptoms, and cognitions. Symptoms, and cognitions and established relations among consisting of entities such as groups, diseases, behaviors, and cognition labels. We provided different labels for different categories on the control panel: disease, behavior, symptom, and cognition categories. We constructed a large-scale knowledge graph (Figure 2E) consisting of entities such as groups, diseases, behaviors, symptoms, and cognitions and established relations among groups and diseases, behaviors, symptoms, and cognitions. These relations can be broken down into 4 categories: group has diseases, group holds behaviors, group shows symptoms, and group carries cognitions. There are 112 group entities in the graph; for example, “Jiangnan District, 12-15 years old, female.”

To characterize groups based on multidimensional attributes and establish the association between groups, we categorized the average tooth value for disease and the strength values of behavior, symptoms, and cognition under the guidance of experts. By considering the numerical distribution of all groups across each attribute and their respective strengths, we divided them into 4 to 5 categories. There are 66 disease entities, including the 4 categories of decayed tooth (DT) due to caries: almost no DT (DT1), mild DT (DT2), moderate DT (DT3), and severe DT (DT4). Similarly, behaviors, symptoms, and cognitions were classified into entities according to their strength value, with a total of 38 behavior entities, 30 symptom entities, and 60 cognition entities. This enables semantic associations among groups, diseases, behaviors, symptoms, and cognitions through the knowledge graph, providing users with a precise and comprehensive semantic expression. To efficiently extract group or attribute characteristics from large-scale entities, we set up a search function in our hub. Users can input keywords related to the node they wish to query, such as groups, diseases, behaviors, cognitions, and symptoms. When searching for a single group or multiple groups, associated disease, behavior, cognition, and symptom entities will be displayed below the search box, organized by category, and ranked to provide a visual description of various attributes of the group. Users can infer the risk of oral diseases through their own similar groups from group-related behaviors, cognitions, symptoms, and other attributes and further grant decision evidence for the prevention of oral diseases. Thus, the visualization tool can easily allow users to identify and intervene in potential oral disease risks and enable medical teams to formulate personalized prevention strategies.

Cloud View
To indicate distinct behaviors, symptoms, and cognitions of each group and explore their correlation with diseases, we offer cloud views that correspond to male and female groups in addition to the radar view. The original data provide the prevalence rate and average number of teeth for the disease attribute in each group. However, the attributes related to behaviors, cognition, and symptoms are usually represented by the proportions of individuals in different degrees. Drink sweat milk, for instance, divides individuals into 4 categories: seldom, 1 per month, 1 per week, and ≥1 per day. It is challenging to evaluate and compare the strengths of each group in these attributes. To address this, we calculated the weighted average and converted it to a value between 0 and 1, known as the strength value. Words are color-coded based on their attribute category, and their size indicates the attribute strength. In this way, users can gain a more intuitive understanding of the behaviors, cognition, and symptoms of different groups and compare them to disease features in the radar view.

Correlation Analysis
We used a scatterplot matrix (Figure 2C) to infer the correlations among complex data attributes. The scatterplot matrix, an extension of the scatterplot for multidimensional data, is crucial for visualizing binary relationships. Nevertheless, the number of matrix elements that can be displayed is constrained by screen size when there are too many dimensions. Here, we applied internet-based methods for users to select the disease, behavior, cognition, and symptom tags in the control panel. The selected tags served as dimensions in a scatterplot matrix view. As we need to use groups to model the associations between the attributes, we first selected several groups within the group view (Figure 2B). In the scatterplot matrix view located on the diagonal, the histogram is used to show and compare the performance of the selected groups in the corresponding attributes. For disease attributes, the height of the bar maps the prevalence, whereas for behaviors, cognitions, or symptoms, the height of the bar depicts the strength value. Scatterplots outside the diagonal are deployed to present 2-by-2 relationships between attributes, with each scatter representing a group. The scatterplot has 2 dimensions: the strength value of the attribute or the normalized number of teeth. Owing to their symmetrical nature, we designed a matrix above the diagonal that offers relationships for all groups, whereas the matrix below the diagonal presents the relationships of the currently selected group under the corresponding attributes. The distribution of scattered points serves as a visual representation of the correlation between the multidimensional attributes.

Semantic Analysis
How can acquired knowledge be logically and scientifically presented after obtaining relevant features and information? We provided a rich set of interactions to assist users in conducting an in-depth analysis of multidimensional oral health surveys. Groups, diseases, behaviors, cognitions, and symptoms can be selected by users, thus enabling personalized and targeted exploration. We offer operations for data loading in the control panel, as shown in Figure 2A. Users can select labels from 4 categories on the control panel: disease, behavior, symptom, and cognition labels. We provided different labels for different age groups because of substantial variations in survey data across age groups. We provided a nested pie chart to display the sample size composition of the groups in 6 regions, including urban and rural areas. The inner circles indicate gender, whereas the outer circles indicate age. Users can select a region, choose age and gender within the corresponding pie chart, and click the “Add” button to include the selected group. After adding the groups, they are displayed in a list below the button, distinguished by random colors. In this way, the chosen label determines the attributes of the scatter matrix, whereas the chosen groups facilitate attribute comparison in the diagonal. All attribute features of the chosen group are also visible in the cloud view and radar view. The graph view shows the semantic relationships of all groups and attributes and offers search capabilities to aid in investigating specific groups and attributes.
Results

Overview
As a web-based visual analysis system, this system was developed using the classic front-end–based frame of ES6+d3.js+csv. A Windows platform with a 2.3 GHz Intel Core i7 CPU and 16 GB of memory was used as the front-end page server. The evaluation experiments were performed using a Google Chrome web browser. Our system can facilitate the efficient and intuitive information mining of experts and users regarding oral diseases. Case studies, user studies, and expert interviews were conducted to demonstrate the usability and viability of our system.

Case Study

Case 1: Disease Analysis
Each disease has its own characteristics, and we introduced a structural view to reveal the characteristics of various diseases. As shown in Figure 3A, we recorded the patient composition of 5 diseases, fill deciduous tooth due to caries (ft), fill tooth due to caries (FT), decayed teeth root due to caries (DRoot), fill teeth root due to caries (FRoot), and attachment loss (AL)12+, in the 4 dimensions of age, city (urban or rural), area, and gender. We found that the width of block A is much larger than that of the 5 blocks on its right. Block A represents the proportion of people with ft in Jianggan. We examined the economic situation of 6 regions with this question in mind. Jianggan exhibits the best overall development among the 6 areas, which explains the higher prevalence of children with filling caries in Jianggan. Blocks B, C, D, and E represent the proportion of urban and female patients with FT and FRoot, respectively. These findings suggest that urban residents possess a better understanding of filling decayed teeth and roots compared with rural residents, and women demonstrate higher awareness than men. Blocks F, G, and H illustrate the proportions of individuals, aged 35 to 44 years, 55 to 64 years, and 65 to 74 years, with DRoot, which shows that the risk of DRoot will increase with age. Similarly, AL12+ is also more likely to occur in the older adult population. Block 1 represents the proportion of men with AL12+, which indicates that men are more likely to experience significant periodontal AL.

Figure 3. Characteristics and risk factors for diseases. (A) Portraits selected from the struct view. (B) Attribute correlations selected from scatterplot matrix. AL12+: attachment loss ≥12 mm; BT: brush the teeth; DC: difficulty chewing; DRoot: decayed teeth root due to caries; FDNPT: fluoride does not protect teeth; ft: fill deciduous tooth due to caries; FT: fill tooth due to caries; FRoot: fill teeth root due to caries; GBI: bleeding gums index; PDL: periodontal pocket depth 4~6mm; PDD: periodontal pocket depth ≥6 mm.

“You get what you grow, you get what you grow.” Oral diseases do not just appear out of nowhere; they are frequently tightly tied to certain actions and cognition. These diseases bring symptoms that affect our daily lives, and some can even spread and lead to other illnesses. As shown in Figure 3B, we intercepted 4 pairs of examples of correlation from the scatterplot matrix view: disease and behavior, disease and symptom, disease and cognition, and disease and disease. We can see that bleeding gums index (GBI) is negatively correlated with brush the teeth, DRoot is positively correlated with difficulty chewing, DRoot is negatively correlated with fluoride that does not protect teeth, and periodontal pocket depth (PPD) of ≥26 mm was positively correlated with the periodontal pocket length of 4-6 mm.

We have summarized some information after a thorough disease analysis. The periodontal health of men is significantly lower than that of women, and individuals lack caries awareness and treatment, and the rate of caries filling treatment is generally low. Middle-aged and older adult groups need to take more proactive measures to prevent and treat periodontal disease in rural areas, which is significantly lower than that in urban areas.

Case 2: Group Analysis
We proceeded with our investigation of the group after examining the characteristics of diseases. The radar view and cloud view in Figure 4A show the diseases, behaviors, symptoms, and cognitions of urban and rural groups of people aged 35 to 44 years, 55 to 64 years, and 65 to 74 years. The horizontal comparison reveals age characteristics, the vertical comparison reveals urban and rural characteristics, and the figure reveals gender characteristics. We learned that the overall disease conditions of the 35 to 44 years age group were less severe than those of the older group. With increasing age, AL, missing tooth (MT) due to caries, DRoot, and DT due to caries showed a deteriorating trend. In the cloud view, the number of yellow words representing symptoms of the group from 65 to 74 years was significantly higher than that of the age group from 35 to 34 years (data for the 55-64 years age group are unavailable and thus not included). When comparing urban and
rural regions, it is evident that the occurrence rates of calculus index, GBI, and DT due to caries were higher in rural areas, whereas FRoot and FT exhibited lower rates. Behaviors such as highest education, use fluoride toothpaste, and toothwash within 12 months, indicated in blue, and cognition, indicated in green, were generally stronger in cities than in villages. Analyzing men and women (55-64 years age group) in the city through the radar view and cloud view, it is apparent that men have greater severity of periodontal diseases, such as PPD ≥6mm, calculus index, GBI, and AL. On the other hand, women showed significantly higher occurrences of FRoot and FT than men, suggesting that women in this group were more conscious of filling teeth due to caries.

Figure 4. Characteristics and comparison of groups. (A) Radar views and cloud views of male and female groups of 3 ages in city and village. (B) The bar charts in the scatterplot matrix of boy and girls aged 12-15 years in each attribute in 6 regions. AE: ashamed to eat; AL: attachment loss; BCGI: Bacteria can cause gum inflammation; BCTD: Bacteria can cause tooth decay; BGNBT: Bleeding gums are normal when brushing teeth; BTUPBG: Brushing teeth is useless in preventing bleeding gums; DC: difficulty chewing; DRoot: decayed teeth root due to caries; DROOT: decayed teeth root due to caries; DT: decayed tooth; ESCTD: Eating sugar can cause tooth decay; FCPT: Fossa closure can protect the teeth; FDNPT: Fluoride does not protect teeth; FHE: father's highest education; FL: food limitation; FROOT: fill teeth root due to caries; ft: fill deciduous tooth due to caries; FT: fill tooth due to caries; GBI: bleeding gums index; HE: highest education; MHE: mother's highest education; MT: missing tooth; OC: only child; OHIL: Oral health is important to life; PDD: periodontal pocket depth; PDL: periodontal pocket depth 4~6mm; ROCI: Regular oral check-ups are important; ST: time since the last dental visit; TCI: Tooth condition is innate, not acquired; TS: tooth sensitivity.

The scatterplot matrix views facilitate intuitive comparison of attributes between different groups, as depicted in Figure 4B. This figure presents behavior comparisons between boys and girls aged 12-15 years in 6 regions: Jianggan, Yuyao, Luqiao, Wenling, Wuyi, and Liandu (from left to right). Each column represents a region, allowing for comparisons across different regions within each subpart. For example, the attributes of father’s highest education and mother’s highest education indicate that boys and girls in Jianggan outperform those in other regions. In addition, Liandu exhibits a longer time since the last dental visit (ST) compared with others. In the attributes smoke, only child, and ST, boys are significantly more numerous than girls, while in DT and FT, girls are more numerous than boys.

We summarized some information after a comprehensive group analysis. There are differences in oral health conditions. The higher quality of dental care and periodontal health in rural areas compared with urban areas may be because of the difference in economic prosperity and level of education and knowledge about oral health between the two. There are gender differences in oral health conditions. The mean and rate of caries in women were slightly higher than those in men, whereas the number of caries fillings and periodontal health in women were better than those in men. There are age differences in the oral health conditions. The prevalence of dental loss and periodontal diseases increased as individuals aged, especially among the older adult group with a weak awareness of the treatment of dental loss and caries. This information can aid medical teams in developing targeted and personalized prevention strategies to address these gender and age disparities in oral health.

Case 3: Semantic Analysis
We constructed a macro knowledge graph for groups, diseases, behaviors, symptoms, and cognitions, effectively converting complex and diverse objects into accessible and intuitive information. Figure 5A shows the semantic association between the different degrees of MT due to caries and the population.
MT1, MT2, MT3, and MT4 represent disease severity, with the population classified accordingly. Not only diseases but also behaviors, symptoms, and cognitive attributes can be categorized. Figure 5A displays the information of the group linked to the MT4 entity node, revealing that the severe dental disease group consisted entirely of older adults aged 65 to 74 years.

**Figure 5.** Discovery of query based on knowledge graph. (A) Similarity and difference attribute information of different degrees of MT due to caries and population. (B) Similarity and difference attribute information of different gender groups. (C) Similar and difference attribute information of different area groups. AE: ashamed to eat; AL: attachment loss; BCGI: Bacteria can cause gum inflammation; BCTD: Bacteria can cause tooth decay; BGNBT: Bleeding gums are normal when brushing teeth; BT: brush the teeth; BTUPBG: Brushing teeth is useless in preventing bleeding gums; CW: communication worry; DC: difficulty chewing; DRoot: decayed teeth root due to caries; DROOT: decayed teeth root due to caries; DT: decayed tooth; DW: dietary worry; ESCTD: Eating sugar can cause tooth decay; FCPT: Fossa closure can protect the teeth; FDNPT: Fluoride does not protect teeth; FHE: father’s highest education; FL: food limitation; FROOT: fill teeth root due to caries; ft: fill deciduous tooth due to caries; FHE: father’s highest education; HS: hinder to speak; MHE: mother’s highest education; MT: missing tooth; OC: only child; OHIL: Oral health is important to life; PDD: periodontal pocket depth; PDL: periodontal pocket depth 4~6mm; PW: pronunciation worry; ROCI: Regular oral check-ups are important; SD: swallowing discomfort; ST: Time since the last dental visit; SW: sleep worry; TCI: Tooth condition is innate, not acquired; TS: tooth sensitivity; TW: toothwash within 12 months; UT: Use the toothpick.

After exploring the groups corresponding to the attributes, we further explored the attributes of groups. Figure 5B illustrates the corresponding attribute association between men and women aged 35 to 44 years in Luqiao, and the commonalities and differences between groups are intuitively presented. For example, AL in both groups was mild, whereas men were more likely than women to have PDL.

Figure 5C illustrates the data for those aged 35 to 44 years in 6 regions. It provides a comprehensive overview of the semantic relationships between these groups, highlighting shared characteristics as well as unique attributes. Notably, we observed a consistent association between the groups facilitated by distinct attributes. Groups from the same region or gender exhibit stronger connections. Specifically, there was a significant overlap in attributes between Jianggan men (point L) and Jianggan women (point K), indicating a close relationship. In addition, a strong association exists between Jianggan women (point K) and Wuyi women (point I).

We summarized some information after a thorough semantic analysis. All age groups exhibited low performance in actions,
including using fluoride toothpaste, dental floss, and scheduling timely visits to an oral hospital. Thus, these actions should be appreciated and strengthened. Middle-aged individuals and older adults still have poor health knowledge and awareness of health care. It is necessary to disseminate and strengthen certain cognitions, such as the knowledge that fossa closure and fluoride can protect teeth.

User Study

To further evaluate the effectiveness of our system, we invited 20 undergraduate and graduate students (12 male students and 8 female students) in digital media technology to participate in a user study. We first introduce the purpose and features of this system and then teach students how to use it. Typically, users need only 10 to 15 minutes of training time to understand the meaning of each view and the function of our system. Afterward, they were asked to perform a series of tasks over a defined period, which were closely related to the analysis tasks in Methods section. The specific tasks were as follows:

- **Disease**
  - Task 1.1. Which disease is more prevalent in Jianggan than in other areas?
  - Task 1.2. Which gender has the highest prevalence of AL12+ disease?

- **Group**
  - Task 2.1. What are the 3 main behavioral characteristics of the male population aged 55 to 64 years in Jianggan?
  - Task 2.2. Which area has a higher prevalence of DRoot among girls aged 12 to 15 years?

- **Correlation**
  - Task 3.1. Is GBI positively or negatively correlated with use of the toothpick (UT)?
  - Task 3.2. Which disease is most likely to present with symptoms of difficulty chewing?

- **Semantics**
  - Task 4.1. What are the groups with severe DT (DT4)?
  - Task 4.2. What are the common characteristics of the women aged 55 to 64 years in Luqiao and Wenling?

To further demonstrate the effectiveness of the system, users will perform the task twice, once without the system and once with it. When the system was not applicable, we provided the users with a condensed version of the surveys. For each experiment, we set the maximum completion time to 90 seconds. Table 1 and Figure 6 record the percentage of the users able to complete the task correctly in a given time and the average and SD of the time completed.

**Figure 6.** User study results. (A) Comparison of completion time. (B) Completion of accuracy rate. DT: decayed tooth system for data mining; OUR: use our system for data mining; SUR: data mining without our system.

**Table 1** presents completion time, average completion time, and standard completion time for both cases. Figure 6 displays specific completion time and accuracy rate through a box chart and a bar chart. It is obvious that using this system would result in more accurate and efficient exploration of oral health surveys. In addition, we collected user feedback after performing the above tasks. They all agreed that the system was quite intriguing and may give them a more intuitive understanding of oral health.
Table 1. User study results.

<table>
<thead>
<tr>
<th>Category and user task</th>
<th>Accuracy rate (%)</th>
<th>Average completion times (s)</th>
<th>Standard completion times (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OUR(^a)</td>
<td>SUR(^b)</td>
<td>OUR</td>
</tr>
<tr>
<td>Disease</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1.1</td>
<td>100</td>
<td>80</td>
<td>10.47</td>
</tr>
<tr>
<td>T1.2</td>
<td>100</td>
<td>90</td>
<td>8.2</td>
</tr>
<tr>
<td>Group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2.1</td>
<td>95</td>
<td>85</td>
<td>19.05</td>
</tr>
<tr>
<td>T2.2</td>
<td>95</td>
<td>75</td>
<td>29.35</td>
</tr>
<tr>
<td>Correlation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3.1</td>
<td>100</td>
<td>85</td>
<td>14.54</td>
</tr>
<tr>
<td>T3.2</td>
<td>90</td>
<td>70</td>
<td>20.85</td>
</tr>
<tr>
<td>Semantics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T4.1</td>
<td>100</td>
<td>75</td>
<td>19.06</td>
</tr>
<tr>
<td>T4.2</td>
<td>90</td>
<td>45</td>
<td>34.73</td>
</tr>
</tbody>
</table>

\(^a\)OUR: use our system for data mining.

\(^b\)SUR: data mining without our system.

Expert Interview

After domain experts used this system to examine oral survey data, we conducted a semistructured interview to collect their opinions on system capability, visual design, and interaction.

System Capability and Effectiveness

The experts expressed their appreciation for the functions provided by this system. They concluded that the system makes it possible for both experts and regular users to quickly and intuitively perceive the initially complex and laborious large-scale oral survey data as well as to easily compare the characteristics of various diseases and groups. Moreover, they agreed that the system effectively revealed the correlation among diseases, behaviors, symptoms, and cognitions. In particular, the oral survey data were transformed into a knowledge graph, a novel approach that is not commonly used in daily survey data analysis. By leveraging the knowledge graph and its query function, this breakthrough enables researchers to go beyond traditional methods that focus on specific tasks and features. It allows them to comprehend large-scale data with complex semantic patterns, making it easier to understand. Ultimately, it greatly enhances their insights into the data.

Visual Design and Interactions

Domain experts praised the user-friendly interface and the well-designed system, aligning each view with its respective function and interaction. The layout facilitates rich interactions, correlation discovery, semantic analysis, and attribute feature exploration. Users can easily comprehend and use them without prior knowledge. One of the experts recognized the scatterplot matrix view’s value in conveniently analyzing group comparison and attribute correlation simultaneously. Another expert emphasized the search capability of the knowledge graph, which allowed him to independently examine valuable information that was difficult to find in regular visual graphs. He suggested that it would be better if the system could offer label options or prompts to explore the nodes. Overall, the experts evaluated the system’s integration of visualizations and interactions, offering a comprehensive range of intelligent explorations for oral health surveys.

Discussion

Principal Findings

A series of studies and experiments have demonstrated that our system can help users understand their oral health conditions and conduct in-depth data exploration and analysis. Furthermore, we conducted a thorough investigation of the visualization analysis tools to compare them with our system. We found that existing tools and libraries provide a rich set of plotting capabilities. However, the visualization analysis tools used in our study are primarily oriented toward specific tasks to visually present and obtain deep insights into oral health surveys. It is implemented using web-based technologies such as the D3.js visualization framework, which offers greater flexibility and customization options for analyzing oral health report data. Existing tools related to oral health analysis mostly include 3D digital dental model software and oral x-ray image processing software, which provide detailed visualization of dental structures. Nevertheless, these tools fail to capture the broader context of oral health such as group characteristics and disease patterns. Moreover, they often provide simple chart-based visualizations, such as pie charts and bar charts, lacking the personalized visualization design and interactive features essential for the comprehensive examination of intricate data. Therefore, it was concluded that our system allows for more customized visualizations based on specific requirements, facilitating a more detailed analysis of oral health surveys.

Overall, our system has specific advantages compared with other analysis tools; however, there are also some issues that
are not well solved, which will be addressed in future work. (1) Scalability is the major concern of this system. The current design in the scatterplot matrix view displays up to 8 attributes and 6 groups simultaneously, whereas the struct view shows up to 7 diseases simultaneously. Even if we set the interaction or scrolling function in it, it still imposes a heavy memory burden on users. Therefore, in the future, we intend to tackle the problem of how to show information more effectively in a limited screen space. (2) Despite its ability to analyze various factors, such as groups, diseases, behaviors, and other attributes in existing data, it currently lacks the capability to predict oral health for new groups or individuals. Combining the deep learning model and oral health professional knowledge, learning from existing multidimensional surveys, and predicting the prevalence of unknown groups will be the focus of future work. (3) In this study, the oral health sample was limited to Zhejiang Province, with a small-scale and narrow regional span, resulting in insufficient group differences. Future studies should consider the multidimensional feature of the disease to explore more robust results. We plan to expand our data collection by conducting oral surveys in additional regions, enabling a more comprehensive exploration of oral health from various dimensions.

Conclusions
In this study, we proposed a visualization framework for multidimensional oral health surveys. We drew data portraits for diseases and groups based on multidimensional attributes. Then, we built correlation patterns for diseases, behaviors, symptoms, and cognitions to reveal their correlation features. On the basis of the extricated knowledge of diseases, groups, behaviors, and other attributes, a knowledge graph is provided to reveal the semantic information. A series of user-friendly interactions are integrated to propose a visual analysis system that can help users further explore the regulations of oral health conditions. Case studies based on real-world data sets demonstrate the effectiveness of our system in the exploration of oral diseases, thereby offering enhanced data analysis capabilities and decision support for health care teams.

Acknowledgments
This work was supported in part by the National Natural Science Foundation of China (62277013 and 62177040), National Statistical Science Research Project (2022LY099), Zhejiang Provincial Natural Science Foundation of China (LTGG23H260003), the China Oral Health Foundation (A2021-008), and Zhejiang Statistical Science Research Project.

Conflicts of Interest
None declared.

References


Abbreviations

AL: attachment loss
DRoot: decayed teeth root due to caries
DT: decayed tooth
ft: fill deciduous tooth due to caries
FT: fill tooth due to caries
GBI: bleeding gums index
MT: missing tooth
PDD: periodontal pocket depth

© Ting Xu, Yuming Ma, Tianya Pan, Yifei Chen, Yuhua Liu, Fudong Zhu, Zhiguang Zhou, Qianming Chen. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 01.08.2023. 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.
A Linked Open Data–Based Terminology to Describe Libre/Free and Open-source Software: Incremental Development Study

Franziska Jahn¹, Dr rer med; Elske Ammenwerth², MET, Dr sc hum; Verena Dornauer², MSc, Dr med; Konrad Höfner¹, Dr rer nat; Michelle Bindel², MSc; Thomas Karopka³, Dipl-Ing; Alfred Winter¹, Dr sc hum

¹Institute of Medical Informatics, Statistics and Epidemiology, Faculty of Medicine, Leipzig University, Leipzig, Germany
²Institute of Medical Informatics, University for Health Sciences, Medical Informatics and Technology, Hall in Tirol, Austria
³GNU Solidario, Las Palmas de Gran Canaria, Spain

Corresponding Author:
Franziska Jahn, Dr rer med
Institute of Medical Informatics, Statistics and Epidemiology
Faculty of Medicine
Leipzig University
Härterstr. 16-18
Leipzig, 04107
Germany
Phone: 49 341 97 16194
Email: franziska.jahn@imise.uni-leipzig.de

Abstract

Background: There is a variety of libre/free and open-source software (LIF OSS) products for medicine and health care. To support health care and IT professionals select an appropriate software product for given tasks, several comparison studies and web platforms, such as Medfloss.org, are available. However, due to the lack of a uniform terminology for health informatics, ambiguous or imprecise terms are used to describe the functionalities of LIF OSS. This makes comparisons of LIF OSS difficult and may lead to inappropriate software selection decisions. Using Linked Open Data (LOD) promises to address these challenges.

Objective: We describe LIF OSS systematically with the help of the underlying Health Information Technology Ontology (HITO). We publish HITO and HITO-based software product descriptions using LOD to obtain the following benefits: (1) linking and reusing existing terminologies and (2) using Semantic Web tools for viewing and querying the LIF OSS data on the World Wide Web.

Methods: HITO was incrementally developed and implemented. First, classes for the description of software products in health IT evaluation studies were identified. Second, requirements for describing LIF OSS were elicited by interviewing domain experts. Third, to describe domain-specific functionalities of software products, existing catalogues of features and enterprise functions were analyzed and integrated into the HITO knowledge base. As a proof of concept, HITO was used to describe 25 LIF OSS products.

Results: HITO provides a defined set of classes and their relationships to describe LIF OSS in medicine and health care. With the help of linked or integrated catalogues for languages, programming languages, licenses, features, and enterprise functions, the functionalities of LIF OSS can be precisely described and compared. We publish HITO and the LIF OSS descriptions as LOD; they can be queried and viewed using different Semantic Web tools, such as Resource Description Framework (RDF) browsers, SPARQL Protocol and RDF Query Language (SPARQL) queries, and faceted searches. The advantages of providing HITO as LOD are demonstrated by practical examples.

Conclusions: HITO is a building block to achieving unambiguous communication among health IT professionals and researchers. Providing LIF OSS product information as LOD enables barrier-free and easy access to data that are often hidden in user manuals of software products or are not available at all. Efforts to establish a unique terminology of medical and health informatics should be further supported and continued.

(JMIR Med Inform 2023;11:e38861) doi:10.2196/38861

KEYWORDS
health informatics; ontology; free/libre open-source software; software applications; health IT; terminology
Introduction

Background

Libre/free and open-source software (LIFOSS) products are increasingly used to support various tasks in health care. LIFOSS generally refers to software products with openly available source code that users and developers can view, analyze, modify, and redistribute.

For example, there are free and open-source software products to implement radiology information systems, picture archiving and communication systems (PACS), patient administration systems, and electronic health record (EHR) systems. Especially in low-resource settings, using LIFOSS can help establish computer-based health information systems [1-3]. Along with their use in hospital and medical practice settings, LIFOSS products are available for mobile health, telemedicine, and public health (eg, [4,5]). Moreover, the COVID-19 pandemic has led to the development of numerous mobile applications for contact tracing, risk assessment, or appointment scheduling, which are often based on LIFOSS and used both in low-resource settings and industrial countries [6,7]. Since 2010, the Medfloss.org database has provided descriptions of LIFOSS used in health care and medicine [8]. As of October 2022, it lists 385 software products and describes them by characteristics like “license,” “application type,” “enterprise function,” “language,” “platform,” and “home page.” Due to the iterative and sometimes uncontrolled growth of the self-developed nomenclature over the last few years, there are several inconsistencies in the software descriptions on Medfloss.org. First, there are misleading assignments of descriptors to characteristics. For example, “laboratory,” “cellular networks,” and “virtual reality” are listed as enterprise functions supported by a software product. However, they describe the setting where the software might be used or special features of the software. Second, the lack of uniform terminology in health informatics has led to the use of synonyms or overlapping terms. For example, the borders between “electronic health systems,” “electronic medical record systems,” and “hospital management systems” are not clearly defined, sometimes leading to ambiguous descriptions of software products. The lack of uniform terminology for describing medical and health care software products and LIFOSS is also apparent when analyzing comparisons of LIFOSS for EHR systems. In several studies published during the last 15 years [1-3,9-11], each research group selected different criteria and descriptors for comparing LIFOSS for EHR systems. In several studies, the technical and functional characteristics of EHR systems are not clearly defined, sometimes leading to synonyms or homonyms. Linking these different terminologies requires semantic analyses based on a uniform set of concepts, along with easy-to-use methods and tools to link these data.

Linked Open Data (LOD) are regarded as the state-of-the-art principle for linking and structuring concepts from different terminologies. LOD are identifiable by a URI and provided using the Resource Description Framework (RDF) standard [15].

Benefits of Unified LIFOSS Terminology Using LOD

A unified LIFOSS terminology using LOD has several advantages. First, the use of predefined and open terminologies supports the search for and comparison of software products. Second, further knowledge, such as results of assessment and evaluation studies, can be linked easily to the software descriptions and thus support evidence-based health informatics. For example, linking systematic descriptions of software products with descriptors from other projects can support ontology-based approaches to software requirements engineering [16].

The Austrian-German research project “Health Information Technology Ontology (HITO)” aims to systematically describe software products and their installations in health care. It uses an underlying ontology named HITO. LOD methods and tools, and freely available catalogues to describe software characteristics. HITO is developed based on different use cases in which precise software descriptions are needed, such as when selecting LIFOSS or commercial software products, searching for evidence about the installation of software products, and communicating about software products among stakeholders in health care. In this paper, we focus on LIFOSS products and describe them with the help of openly available catalogues. Especially for LIFOSS products, information about software characteristics is freely available, and LIFOSS developers are likely to recognize the potential advantage of spreading knowledge about their products with the help of LOD.

Objectives

This study aims to (1) describe LIFOSS systematically with the help of precise descriptors that are captured in HITO and (2) publish HITO and HITO-based software product descriptions as LOD using Semantic Web tools for viewing and querying LIFOSS data on the World Wide Web.

Methods

Requirements Elicitation

Initial Steps

As the first step toward precise descriptions of software products, the classes and relationships that are useful to describe software products had to be identified. HITO was developed and refined iteratively by collecting requirements of several use
cases. Each use case focused on a situation in which a clear terminology for software products was considered essential. These use cases dealt with the description of LIFOSS or commercial software products for potential and current users in health care settings or the description of software product installations in evaluation studies on health IT interventions. We selected diverse use cases to identify the most relevant characteristics for these diverse situations. Based on these use cases, we incrementally built the ontology that contains a general pattern for describing software products in health care.

**Use Case 1: Evaluation of Digital Health Interventions**

The first HITO use case dealt with evaluation studies in health informatics. In evaluation studies, it is crucial to carefully describe the functionalities and application types with enough precision. Therefore, an opportunity to support Medfloss.org terminology may be to integrate existing terminologies such as WHO’s classification of digital health interventions [12] or the HL7 EHR System Functional Model [13]. Furthermore, the folksonomy (ie, the collection of users’ tags for certain objects) of the platform users is not handled by the search functions on platform. For each category, the list of descriptors has grown over the years as new LIFOSS descriptions were added.

<table>
<thead>
<tr>
<th>HITO class (characteristic of software products)</th>
<th>Description and examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software product</td>
<td>Piece of software that is sold as a commercial product or distributed under an open-source license</td>
</tr>
<tr>
<td>Feature</td>
<td>Functionalities offered by a software product that directly contribute to the fulfillment of 1 or more enterprise functions (eg, email notification of new results, user directory to control any access)</td>
</tr>
<tr>
<td>Application system type</td>
<td>Commonly used names for categories of software product installations in health care (eg, radiology information system, CPOE system)</td>
</tr>
<tr>
<td>Organizational unit</td>
<td>Health care setting in which the software product is used and an evaluation study was conducted (eg, laboratory, department of pediatrics)</td>
</tr>
<tr>
<td>User group</td>
<td>Health care staff who uses the software product installation (eg, nurse, radiologist)</td>
</tr>
</tbody>
</table>

*aHITO: Health Information Technology Ontology.  
bCPOE: computerized physician order entry.

**Use Case 2: Description of LIFOSS With Medfloss.org Project Database**

In the LIFOSS use case, we extended the list of HITO classes by carefully analyzing the Medfloss.org project database. Medfloss.org aims to offer an overview of LIFOSS projects related to medical informatics and health care [8]. Although Medfloss.org is not maintained anymore, it is still provided in cooperation with 3 LIFOSS-related working groups of the International Medical Informatics Association, the European Federation for Medical Informatics Association, and the International Society for Telemedicine and eHealth. Within Medfloss.org, the LIFOSS products are described by using a predefined set of categories.

We started this use case by surveying 2 operators of the Medfloss.org database. They were asked independently to answer a survey with 11 open questions and 1 closed question. The survey asked about the users of Medfloss.org, the relevance of the categories used to describe software products, and the positive and negative experiences with the categories used to describe LIFOSS.

The results of this survey show that Medfloss.org is intended to be used by physicians or other health care staff, IT administrators, information managers, and software developers to select appropriate LIFOSS for certain health care tasks. Each LIFOSS product is described by 11 categories, such as “enterprise function,” “application type,” and “license,” on the platform. For each category, the list of descriptors has grown over the years as new LIFOSS descriptions were added.

The answers of the platform operators dealing with the assessment of the current terminology and its representation on the platform were arranged according to a SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis (Textbox 1).

Both operators rated the “enterprise function,” “application type,” “status,” “license,” “standard,” “language,” “client type,” and “platform” categories of Medfloss.org terminology as “important” or “very important” for the description of LIFOSS. The categories “popularity,” “database,” and “programming language/toolkit” were rated less important by 1 of the website operators.

The strengths of the Medfloss.org terminology used include the rough categorization of software products by application system types and the faceted search on the website that is based on categories for describing the software. However, the survey showed that the categories currently lack the possibility to describe the functionalities and application types with enough precision. Therefore, an opportunity to support Medfloss.org terminology may be to integrate existing terminologies such as WHO’s classification of digital health interventions [12] or the HL7 EHR System Functional Model [13]. Furthermore, the folksonomy (ie, the collection of users’ tags for certain objects) of the platform users is not handled by the search functions on Medfloss.org [18].
Based on the SWOT analysis, we selected 9 (out of 10) of the Medfloss.org categories and added these to HITO (Table 2). Some classes used on Medfloss.org were renamed, such as “standard,” which was changed to “interoperability standard” to sum up interoperability standards, as well as frameworks describing how to use interoperability standards, such as Integrating the Healthcare Enterprise (IHE). A more fine-grained classification according to interoperability levels was examined for common interoperability standards but proved to be impractical. Many interoperability standards such as HL7 Fast Healthcare Interoperability Services (FHIR) can be assigned to multiple interoperability levels. The Medfloss.org category “programming language/toolkit” was split into 2 classes to distinguish between these different concepts.

Textbox 1. SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis for Medfloss.org.

<table>
<thead>
<tr>
<th>Strengths</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Current set of categories to describe libre/free and open-source software</td>
<td></td>
</tr>
<tr>
<td>Rough categorization by application system types</td>
<td></td>
</tr>
<tr>
<td>Usefulness of the categories to provide a faceted search on the platform</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Weaknesses</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual overlaps in categories (eg, electronic health record and electronic medical record)</td>
<td></td>
</tr>
<tr>
<td>Missing hierarchies for enterprise functions</td>
<td></td>
</tr>
<tr>
<td>Missing detailed functional descriptions</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Opportunities</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Enhancements of categories by existing terminologies seems possible</td>
<td></td>
</tr>
<tr>
<td>Modeling of user group–dependent categories may increase usefulness</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Threats</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No handling of synonyms within users’ search terms</td>
<td></td>
</tr>
</tbody>
</table>
### Further HITO Use Cases

We summarized 3 further use cases that have not elicited new HITO classes related to software product characteristics.

The third use case deals with the description of commercial software products used in health care. This use case confirmed the set of HITO classes that we already identified by describing LIFOSS products. However, for commercial software products, it is quite challenging to describe them based on these classes because meaningful descriptions of commercial software products are rarely publicly available.

In the fourth use case, the existing HITO classes were linked with competency levels for IT staff in health care organizations, which might be useful for creating job advertisements.

In the fifth use case, findings from the HITO project were discussed with practitioners like hospital chief information officers and industry representatives to discuss the applicability and broader use of the HITO project’s findings in practice.

### Software Product Descriptions as LOD

After building HITO from the described use cases, we published HITO and HITO-based software product descriptions using LOD.

LOD are web data with an open license. They allow the use of Semantic Web tools for viewing and querying LIFOSS data on the World Wide Web. LOD also allow for linking and reusing existing terminologies. To be considered “5-star LOD,” the data need to be machine-readable, presented in a nonproprietary format, use open standards of the World Wide Web Consortium (W3C), and be linked to other data [15]. We strived to achieve 5-star linked open HITO data.

Accordingly, we used RDF Schema (RDFS) and the Web Ontology Language (OWL). RDFS and OWL are W3C standards used to define the types of elements of discourse as classes. The class “software product,” for example, represents the set of all individual software products. Properties represent possible binary-typed relationships between individuals of certain classes, for example, between software products and features. Using RDF, facts (relationships between individuals) are expressed as subject-predicate-object triples, whose elements may be defined and stored in different places. This allows for reusing existing vocabularies and interlinking with existing knowledge bases, forming the LOD cloud. Each RDF resource (a class, an individual, or a relationship) has a URL where it is published both in human-readable (ie, HTML) format and in machine-processable RDF serialization format. To browse RDF data comfortably, tools like RickView [19] can be used and modified. SPARQL Protocol and RDF Query Language

---

**Table 2. HITO[a] to describe LIFOSS[b] products (HITO use case 2).**

<table>
<thead>
<tr>
<th>Medfloss.org class name</th>
<th>HITO class</th>
<th>Description and examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Client type</td>
<td>Client</td>
<td>The client type on which a software product can be run (mobile, native, or web).</td>
</tr>
<tr>
<td>Database</td>
<td>Database management system</td>
<td>Some examples are PostgreSQL[c] or MySQL.</td>
</tr>
<tr>
<td>Enterprise function</td>
<td>Enterprise function</td>
<td>Describes what action humans or machines must carry out in a certain enterprise to contribute to its mission or goals (eg, patient admission, order entry).</td>
</tr>
<tr>
<td>Home page</td>
<td>Home page</td>
<td>Home page of the software product or its development project.</td>
</tr>
<tr>
<td>Standard</td>
<td>Interoperability standard</td>
<td>Ability of 2 or more components to exchange information and to use the information that has been exchanged. Under this class name, interoperability standards (eg, HL7[d], FHIR[e] or DICOM[f]) or frameworks describing how to use standards (eg, IHE[g]) are summed up.</td>
</tr>
<tr>
<td>Language</td>
<td>Language</td>
<td>Languages in which the software product is available (eg, English, French, and German).</td>
</tr>
<tr>
<td>License</td>
<td>License</td>
<td>The license under which a software product is distributed.</td>
</tr>
<tr>
<td>Platform</td>
<td>Operating system</td>
<td>The operating system a software uses (eg, Windows). A software product might be able to run on a variety of operating systems.</td>
</tr>
<tr>
<td>Programming language/toolkit</td>
<td>Programming language</td>
<td>The programming language used to develop a software product (eg, Java or Python).</td>
</tr>
<tr>
<td>Programming language/toolkit</td>
<td>Programming library or toolkit</td>
<td>Programming toolkits are utility programs that are used to develop and maintain software. Programming libraries are a collection of prewritten functions that are ready to be used in coding. Both help programmers develop software in a fast and safe manner.</td>
</tr>
</tbody>
</table>

---

[a]HITO: Health Information Technology Ontology.
[b]LIFOSS: libre/free and open-source software.
[c]SQL: Structured Query Language.
[f]DICOM: Digital Imaging and Communications in Medicine.
[g]IHE: Integrating the Healthcare Enterprise.
SPARQL end points allow free access to structured read-only queries for humans and as application programming interfaces (APIs) for several tools.

Integration of Software Product and Health-Related Terminologies

One advantage of LOD is their easy integration of existing data sources that are already available in RDF format. Therefore, for the HITO classes (Tables 1 and 2) that characterize software products, we searched for “catalogues” (ie, lists of terms that can be used as instances for the respective class). For the selection of suitable catalogues, we defined the following criteria:

[The catalogue should be authored by an established scientific or standardization organization.] OR [The catalogue must be scientifically plausible with regard to reproducibility, having undergone peer-review or having been developed by more than 5 persons.] OR [The catalogue must be openly available and developed by a large community.]

We used a broad literature and web search and our knowledge of the field to identify related taxonomies and catalogues that can be useful for describing instances of HITO classes (Table 2). In the following paragraphs, we provide a brief overview of the catalogues we investigated.

We started with DBpedia, a popular and large knowledge base comprising billions of triples extracted from Wikipedia, texts, and other sources [20,21]. We analyzed DBpedia to identify possible instances or subclasses for the HITO classes. For the classes “language,” “operating system,” and “programming language,” we found DBpedia classes with suitable instances that we replaced the associated HITO classes with to increase interoperability. Other instances of DBpedia classes, such as the “license” class, were not suitable for integration because DBpedia does not semantically differentiate between licenses for software products and licenses for other purposes, like drivers’ licenses. For software product licenses, we integrated subclasses of the class “open-source software license” derived from the Software Ontology for biomedical software [22].

The most challenging task was the integration of catalogues for the classes “application system type,” “enterprise function,” “feature,” “organizational unit,” and “user group,” for which we needed instances or subclasses related to health care. We checked the following sources for the integration of catalogues into HITO:

- HL7 EHR System Functional Model [13]: Using the examples of 2 installations of commercial software products for EHR systems, we assessed whether the features of the software product could be described by this model. We found that the whole list of conformance criteria defined in this model would be too detailed for a HITO catalogue. However, the section labels that are provided by this model, such as “manage allergy, intolerance, and adverse reaction list,” provide an appropriate level of detail for feature descriptions in HITO.

- The textbook by Winter et al [14] describes a set of application system types and a set of enterprise functions in hospitals. An analysis of Medfloss.org revealed that the sets of enterprise functions or application system types could be used to tag 71% and 42%, respectively, of 356 Medfloss.org software products analyzed in use case 2 [23].

Describing Software Products Using HITO

With the help of HITO and the selected catalogues, we described 25 LIFOSS products were described. These were selected to represent different application system types and due to their comprehensive, openly available documentation. For the description of supported enterprise functions and features, we extracted terms from the software product documentation and linked as many as possible to catalogue entries of HITO. Information about the LIFOSS products was extracted by 1 team member (author MB) and checked by another team member (author FJ).
Results

HITO was modeled and published using Semantic Web technologies. The Unified Modeling Language class diagram in Figure 1 describes the structure of HITO. The classes shown to the left of “software product” describe the general characteristics of software products. For classes with domain-specific instances (ie, application system type, feature, enterprise function, organizational unit, and user group), we applied a scheme of 3 interrelated classes named <classname>Catalogue, <classname>Classified, and <classname>Citation. A <classname>Catalogue is a health IT–related collection, such as the HL7 EHR Functional Model or the WHO classifications for digital interventions. A <classname>Classified is a category that belongs to exactly 1 catalogue. A <classname>Citation is a textual label extracted from available software manuals, descriptions, and studies and thus represents a part of the folksonomy contained in HITO [28].

To illustrate how this scheme applies, we describe the domain-specific characteristics of the Orthanc software [29,30] in Table 3. The developers of Orthanc refer to it as “mini-PACS,” “DICOM (Digital Imaging and Communications in Medicine) server,” “VNA (vendor-neutral archive),” and “viewer of medical images.” These terms are assigned to the software product Orthanc as “application system type citations.” In turn, “DICOM server” and “mini-PACS” have links to the classified application system type “PACS” from the application system type catalogue in Winter et al [14]. The supported enterprise function citations extracted from the Orthanc website, such as “image archiving,” “image management,” and “research about the automated analysis of medical images,” were linked to the more general terms “laboratory and diagnostics imaging management” and “research and education” from enterprise function catalogues. For the 16 feature citations extracted from the Orthanc online documentation, we identified linkable classified features of the WHO classification of digital health interventions and the PACS feature list. Some classified features or enterprise function terms have direct links to the software product. These assignments had no match with citations and were done by domain experts.

Overall, with the help of HITO, we described 25 LIFOSS products in similar detail as Orthanc. All software product descriptions are available as LOD. We described HITO classes and relationships of the ontology and individual software products using RDF. The HITO SPARQL end point [31] allows queries using SPARQL (Figure 2).

The RickView application allows for browsing through the ontology and knowledge base [32]. Another suitable way to query is with a faceted search [33], whereby integrated terminologies can be used to find software products of a certain application system type or supporting combinations of features and functions (Figure 3).

The ontology is also publicly available under version control in a GitHub repository [34]. It can be downloaded in RDF Turtle format to be viewed in ontology editors like Protégé. HITO is dedicated to the public domain and uses a Creative Commons Zero v1.0 Universal license. There are a few exceptions for integrated terms from SNOMED [25] and the WHO classification of digital health interventions [12].
Table 3. Application system types, supported functions, features, user groups and organizational units of the Orthanc software.

<table>
<thead>
<tr>
<th>Application system type</th>
<th>PACS&lt;sup&gt;c&lt;/sup&gt; (application systems in [14])</th>
<th>PACS&lt;sup&gt;c&lt;/sup&gt; (application systems in [14])</th>
</tr>
</thead>
<tbody>
<tr>
<td>DICOM&lt;sup&gt;b&lt;/sup&gt; server</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>mini-PACS</td>
<td>Image archiving</td>
<td>Image archiving</td>
</tr>
<tr>
<td>VNA&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Image communication</td>
<td>Image communication</td>
</tr>
<tr>
<td>Viewer of medical images</td>
<td>Image distribution</td>
<td>Image management</td>
</tr>
<tr>
<td>Web viewer</td>
<td>Research about the automated analysis of medical images</td>
<td>Research and education (enterprise functions from [14])</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Enterprise function</th>
<th>Execution of radiological examinations (enterprise functions from [14])</th>
<th>Laboratory and diagnostics imaging management (enterprise functions from [12])</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>Image communication</td>
<td>Image distribution</td>
</tr>
<tr>
<td>Image archiving</td>
<td>Laboratory and diagnostics imaging management (enterprise functions from [12])</td>
<td>Image management</td>
</tr>
<tr>
<td>Image communication</td>
<td>Laboratory and diagnostics imaging management (enterprise functions from [12])</td>
<td>Research about the automated analysis of medical images</td>
</tr>
<tr>
<td>Image distribution</td>
<td>Research and education (enterprise functions from [14])</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>Capture diagnostic results from digital devices (features from [12])</th>
<th>Compatibility and integration with other systems and products (PACS feature list [26])</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>Data management for clinical routine and medical research</td>
<td>Compatibility and integration with other systems and products (PACS feature list [26])</td>
</tr>
<tr>
<td>DICOM identifiers</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>DICOM network protocol</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Evaluations Report</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Injury surveillance system registration report</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Listing available servers</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Plugin mechanism to add new modules</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Retrieve images</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Retrieving DICOM resources from WADO-RS&lt;sup&gt;f&lt;/sup&gt; server</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Search the content</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Send images</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Sending DICOM resources to a STOW-RS&lt;sup&gt;g&lt;/sup&gt; server</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Test the connection</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Top diseases report</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Research</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Software/hardware integrators in the medical field</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Network engineer</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>System engineer</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Organizational unit</th>
<th>Health center environment [25]</th>
<th>Hospital environment [25]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health centers</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Hospital environment</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Figure 2. A SPARQL Protocol and RDF Query Language (SPARQL) query (on the left) and its results (on the right). RDF: Resource Description Framework.
Discussion

Principal Results

In this project, we identified characteristics useful for describing software products and systematically captured them as classes in HITO. Accordingly, we exploited the properties of ontologies that enable semantic description and linking of data.

For a thorough functional description of software products in health care, we described the enterprise functions supported and features offered by the software products. For health-specific characteristics of software products, we analyzed and integrated existing terminologies for enterprise functions, features, application system types, organizational units, and user groups as catalogues. Because we expected the relevant sources for such catalogues to come not only from science but also from practice, a systematic review of the scientific literature for health IT–related terminologies would not have led to sufficient results. Accordingly, we based the selection of catalogues on statements by domain experts among the project team and project partners, supplemented by targeted PubMed and Google searches on specific application system types.

Thus, we used a case-based and agile approach to identify classes, relationships, and catalogues best suited for describing software products in health care. In the use cases considered so far, ontological reasoning had no relevance. Therefore, to date, only few axioms are used in HITO. As the next step toward more interoperability with other formal ontologies, HITO could use an upper-level ontology such as Basic Formal Ontology [35], General Formal Ontology [36], or gist [37]. A first feasibility check of these ontologies showed that the gist ontology, which defines typical upper-level business concepts, may be the most appropriate for the scope of HITO.

With the help of HITO, we described 25 LIFOS products in detail that, together with less detailed descriptions of single commercial software products and software products extracted from evaluation studies, form our knowledge base. The descriptions of these software products could be regarded as a proof of concept. However, we noticed the interpretative degrees of freedom in assigning correct enterprise functions and features to software products. To ensure the validity of further software descriptions, it would be helpful to calculate interrater reliability among 2 independent experts. For this, further software product entries of the Medfloss.org database that have not yet been considered in HITO’s knowledge base could be used.

As postulated by Berners-Lee [15], HITO’s availability as LOD facilitates its barrier-free access and use. In particular, the integrated catalogues for enterprise functions, features, and application system types provide HITO users with rich terminology for functionalities of software products. However, since there is more than 1 catalogue for each of these characteristics, new terminological problems arise. The
catalogue entries of different catalogues must be mapped to each other to achieve comparability of software products described with the help of different catalogues. Linking these catalogues is part of ongoing research. Together with the folksonomy terms that are already connected to catalogue entries, HITO users will be able to retrieve the most suitable software using a broad range of search terms. The catalogues currently integrated into HITO focus more on health care rather than on medical research tasks (ie, software products like research databases may not be sufficiently described by HITO). However, integrating further catalogues describing research-related enterprise functions or features could be possible.

Nevertheless, publishing HITO and its knowledge base as LOD implies that the contents of HITO are available under an open license. Thus, for the broadly accepted nomenclature SNOMED CT, we could only check the principal suitability of the SNOMED “environment” and “occupation” concepts based on a small set of examples that we included in HITO with permission from SNOMED International. Due to license requirements, SNOMED CT terms cannot be made available as LOD.

In summary, HITO provides an openly available framework for the description of health care–related software that can be used by researchers who publish studies on digital health interventions or by developers and users who need to describe software.

Conclusions

We recognize that health informatics continues to face a terminology problem. Establishing a uniform terminology for software products used in health care is currently unachievable due to several coexisting terminologies from both research and practice. Linking the terms from different terminologies by similarity relationships is the first step toward more transparency. This will also help identify misunderstandings that may be caused by synonyms, homonyms, or conceptual overlaps. Simply knowing that the term “EHR system” can stand for an institutional or cross-institutional application system or a collection of digital documents related to a person's health prevents problems related to misunderstanding. A researcher authoring a study on a digital health intervention by an EHR system knows that the term “EHR system” must be further specified, for example, by enterprise functions and features as listed in HITO.

Nevertheless, we should further strive for a consented, uniform terminology of health informatics. Taking different coexisting terminologies as a basis, methods of qualitative content analyses such as inductive category formation [38], supported by (semi)automatic text extraction, may lead the way toward an established language for health informatics.

Acknowledgments

The work presented here is part of the project “HITO: a Health IT Ontology,” funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - WI 1605/11-1 and the Austrian Science Fund (I 3726-N31). The publication of this article was partly funded by the Open Access Publishing Fund of Leipzig University, supported by the German Research Foundation for the Open Access Publication Funding program.

Conflicts of Interest

None declared.

References


27. SPARQL Query Editor. HITO. URL: http://hitontology.eu/sparql [accessed 2022-03-24]

28. HITO Health IT Ontology. HITO. URL: https://hitontology.eu/ontology/ [accessed 2022-03-22]


32. SNOMED CT browser. SNOMED International. URL: https://browser.ihtsdotools.org/? [accessed 2022-01-19]

Abbreviations

- API: application programming interface
- DICOM: Digital Imaging and Communications in Medicine
- EHR: electronic health record
- HITO: Health Information Technology Ontology
- HL7: Health Level 7
- LIF OSS: libre/free and open-source software
- LIS: laboratory information system
- LOD: Linked Open Data
- OWL: Web Ontology Language
- PACS: picture archiving and communication system
- RDF: Resource Description Framework
- RDFS: Resource Description Framework Schema
- SNOMED: Systematized Nomenclature of Medicine
- SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
- SPARQL: SPARQL Protocol and RDF Query Language
- SWOT: Strengths, Weaknesses, Opportunities, and Threats
- VNA: vendor-neutral archive
- WHO: World Health Organization

©Franziska Jahn, Elske Ammenwerth, Verena Dornauer, Konrad Höffner, Michelle Bindel, Thomas Karopka, Alfred Winter. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 20.01.2023. 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.
Systematized Nomenclature of Medicine–Clinical Terminology (SNOMED CT) Clinical Use Cases in the Context of Electronic Health Record Systems: Systematic Literature Review

Riikka Vuokko¹, PhD; Anne Vakkuri²*, MD, PhD; Sari Palojoki³*, PhD
¹Unit for Digitalization and Management, Ministry of Social Affairs and Health, Helsinki, Finland
²Perioperative, Intensive Care and Pain Medicine, Helsinki University Hospital, Vantaa, Finland
³Unit for Digital Transformation, European Centre for Disease Prevention and Control, Stockholm, Sweden
*these authors contributed equally

Corresponding Author:
Riikka Vuokko, PhD
Unit for Digitalization and Management
Ministry of Social Affairs and Health
PO Box 33
Helsinki, FI-00023 Government
Finland
Phone: 358 50 453 6377
Email: riikka.vuokko@gov.fi

Abstract

Background: The Systematized Medical Nomenclature for Medicine–Clinical Terminology (SNOMED CT) is a clinical terminology system that provides a standardized and scientifically validated way of representing clinical information captured by clinicians. It can be integrated into electronic health records (EHRs) to increase the possibilities for effective data use and ensure a better quality of documentation that supports continuity of care, thus enabling better quality in the care process. Even though SNOMED CT consists of extensively studied clinical terminology, previous research has repeatedly documented a lack of scientific evidence for SNOMED CT in the form of reported clinical use cases in electronic health record systems.

Objective: The aim of this study was to explore evidence in previous literature reviews of clinical use cases of SNOMED CT integrated into EHR systems or other clinical applications during the last 5 years of continued development. The study sought to identify the main clinical use purposes, use phases, and key clinical benefits documented in SNOMED CT use cases.

Methods: The Cochrane review protocol was applied for the study design. The application of the protocol was modified step-by-step to fit the research problem by first defining the search strategy, identifying the articles for the review by isolating the exclusion and inclusion criteria for assessing the search results, and lastly, evaluating and summarizing the review results.

Results: In total, 17 research articles illustrating SNOMED CT clinical use cases were reviewed. The use purpose of SNOMED CT was documented in all the articles, with the terminology as a standard in EHR being the most common (8/17). The clinical use phase was documented in all the articles. The most common category of use phases was SNOMED CT in development (6/17). Core benefits achieved by applying SNOMED CT in a clinical context were identified by the researchers. These were related to terminology use outcomes, that is, to data quality in general or to enabling a consistent way of indexing, storing, retrieving, and aggregating clinical data (8/17). Additional benefits were linked to the productivity of coding or to advances in the quality and continuity of care.

Conclusions: While the SNOMED CT use categories were well supported by previous research, this review demonstrates that further systematic research on clinical use cases is needed to promote the scalability of the review results. To achieve the best out-of-use case reports, more emphasis is suggested on describing the contextual factors, such as the electronic health care system and the use of previous frameworks to enable comparability of results. A lesson to be drawn from our study is that SNOMED CT is essential for structuring clinical data; however, research is needed to gather more evidence of how SNOMED CT benefits clinical care and patient safety.

(JMIR Med Inform 2023;11:e43750) doi:10.2196/43750
KEYWORDS
clinical; electronic health record; EHR; review method; literature review; SNOMED CT; Systematized Nomenclature for Medicine; use case; terminology; terminologies; SNOMED

Introduction

Background

The Systematized Medical Nomenclature for Medicine–Clinical Terminology (SNOMED CT) is an extensive, multi-hierarchical clinical terminology system. It provides a standardized and scientifically validated way of representing clinical information [1]. The application possibilities of SNOMED CT are well documented [1-5], and various guides describe the following types of implementation: clinical records, knowledge representation, data aggregation, and analysis. Specifically, the previous literature describes the various development goals of SNOMED CT. For example, SNOMED CT can be used as a standard for electronic health records (EHRs) for classifying or coding clinical information. Additionally, standardized terminology advances data indexing, storing, and retrieving. This supports sharing of patient information across medical domains and organizations in ways that promote continuity of care. As a large-scale terminology system, SNOMED CT also enables knowledge representations in clinical guidelines and care pathways, which can be used, for example, with decision support [2-5].

Data recorded in EHRs are primarily used to provide care to patients. The potential of SNOMED CT to improve data quality and facilitate interoperability, and thus improve patient safety, has long been noted in existing research. Studies have shown that structured and standardized EHRs also increase data reuse possibilities [6]. The European Union and the US Healthcare Information Technology Standards Panel have noted possibilities provided by SNOMED CT and taken steps toward increasing semantic interoperability, reuse, and the exchange of health data. Data recorded in local systems can also be used to support the achievement of broad health policy goals. The importance of SNOMED CT is expected to gradually grow, but at the same time, there is a need to tackle the complex implementation challenges that may arise [1,7,8].

When implemented in EHRs, SNOMED CT is used to represent clinical information consistently and comprehensively [1,2]. Despite the widespread adoption of EHRs that are certified to follow terminology standards, and although SNOMED CT is used in more than 50 countries, there are only a few published reviews about its clinical use. Most studies have focused on theory and predevelopment or design [1,2,8,9]. Moreover, studies in the past have analyzed general factors related to EHR adoption but have not explored the factors associated with less advanced EHR product implementations as compared to more advanced and mature EHR systems. In the context of SNOMED CT implementation, the maturity of an EHR system is a relevant factor to appropriately maximize the benefit from previous experiences [10-12].

In summary, even though SNOMED CT has been extensively studied as a clinical terminology system, previous research has repeatedly documented a lack of detailed evidence for SNOMED CT in clinical use cases [2,3]. Considering that implementing SNOMED CT is a challenging proposition [2], the identification of specific barriers and facilitators to implementing SNOMED CT in clinical use is of paramount importance to further promote its adoption. Evidence from use cases might support implementation and provide guidance on avoiding deployment pitfalls. Therefore, we aimed to explore the available evidence in previous literature reviews of clinical use cases of SNOMED CT integrated into EHR systems or clinical applications during the last 5 years of continued development [2,3,5,13].

Objectives

The aim of this review is to provide an overview of published studies on SNOMED CT clinical use cases in the context of EHRs. In this study, we apply categories from previous research for analyzing use purpose and use phase for the terminology. Moreover, we present core benefits by summarizing the observations from the EHR use cases [3,5,13].

Our research questions are as follows: (1) What are the main clinical use purposes of SNOMED CT during the last 5 years of development? (2) What kinds of use phases of SNOMED CT are identified in these studies? and (3) What are the summarized clinical benefits documented in each SNOMED CT use case?

Methods

To explore EHR-related SNOMED CT use cases in recent research, our research team set out to conduct a systematic literature review. Our team consisted of a medical expert with decades of experience in clinical care and two health and medical informatics experts. To analyze EHR use cases where SNOMED CT terminology was applied, we extended the concept of EHR systems to cover EHR-related applications and software in clinical use. The word “clinical” refers to “medical work or teaching that relates to the examination and treatment of ill people” [14]. In our review, a use case consisted of SNOMED CT integrated into an EHR system in various stages of use, either in preuse development or in design, piloting, testing, implementation, use, or postimplementation evaluation [15].

The team followed the Cochrane review protocol [16] to plan the necessary steps for this study design (Multimedia Appendix 1). Within the team, the application of the protocol was modified step-by-step to fit the research problem by first defining the search strategy, identifying the articles for the review by isolating the exclusion and inclusion criteria for assessing the search results, and lastly evaluating and summarizing the review results.

We defined our search strategy by adding variations of search terms and testing the suitability against the search results. A search with key words “EHR,” “EMR,” “electronic health record,” or “electronic health record system” produced a large number of search results. Combining these terms with
“SNOMED CT” produced relevant search results for our review purposes. Furthermore, adding filters (Textbox 1) did not cause a significant change in the search results. A search of PubMed using the systematic-review methods filter was undertaken in March 2022 and resulted in 162 original articles after removing duplicates (Figure 1). Our results cover the last 5 years of research; thus, our review forms a continuation of previous reviews [2-5].

Textbox 1. Search strategy and filters used.

<table>
<thead>
<tr>
<th>Search terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>(((((Ehr) OR (Emr)) OR (electronic health record)) OR (electronic health record system)) OR (electronic medical record) AND ((fha[Filter]) AND (fft[Filter]) AND (2016/1/1:2022[pdat]))) AND ((Snomed ct) OR (snomed CT) AND ((fha[Filter]) AND (fft[Filter]) AND (2016/1/1:2022[pdat]) AND (english[Filter])))</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Filters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract, Full text, English, Abstract, Full text, English, from 2016/1/1 – 2022</td>
</tr>
</tbody>
</table>

Figure 1. Application of the review protocol as a flowchart. EHR: electronic health record.

The exclusion and inclusion criteria were defined before conducting the search. We based our criteria on our research questions and previous research. We defined the exclusion criteria as follows: First, the original article had to document an EHR use case where SNOMED CT was being tested, piloted, implemented, or used in a clinical context. This excluded, for example, research concentrating on theoretical building, evaluation, or validation of SNOMED CT. Second, we excluded population studies and, for example, cohort studies where SNOMED CT was used to define, extract, and harmonize study data and where no EHR-related design or use goals were documented. Third, we excluded editorials, posters, and other
such sources to limit the review to original research articles. While reading the articles, we discussed how well the exclusion criteria corresponded to the delimitation made based on our research questions and the conclusions we arrived at while reading the research content as presented in the original articles.

After the first exclusion based on article headings and abstracts, we had excluded 87 articles. We included 31 articles that seemed to relate to our research question. Moreover, 44 borderline cases merited special consideration to determine if they should be excluded or included (Figure 1). The two researchers set out to read a total of 75 full-text articles. The information extraction and documentation template for final inclusion had been defined based on previous research [3,5,13] and research team agreement (Table 1). During the final reading, the documentation template for information extraction was concurrently refined.

Figure 1 illustrates how the search results were analyzed during exclusion and inclusion screening. In the end, we included 17 original articles in the review analysis. Our final inclusion was confirmed by the research documentation that illustrated the SNOMED CT use case in an EHR or EHR-related application and software in clinical use.

Table 1. Criteria to categorize Systematized Nomenclature of Medicine–Clinical Terminology use in the review.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical use context</td>
<td>Refers to clinical domain or specialty as documented in the study.</td>
</tr>
<tr>
<td>EHR&lt;sup&gt;a&lt;/sup&gt; system</td>
<td>Refers to EHR systems or other EHR-related applications or software as documented in the study.</td>
</tr>
<tr>
<td>Users</td>
<td>Refers to intended users of the SNOMED CT&lt;sup&gt;b&lt;/sup&gt; integrated into the EHR as documented in the study.</td>
</tr>
<tr>
<td>SNOMED CT use category</td>
<td>Refers to primary purpose for using SNOMED CT as documented in the study [3,5,13]. Based on research team agreement, the following categories were used: standard for EHR or for a clinical application, retrieval or analysis of patient data, data extraction (used to classify or code in a study), proving merit of SNOMED CT, and development of automated coding.</td>
</tr>
<tr>
<td>SNOMED CT use phase</td>
<td>Refers to the stage of the SNOMED CT use as documented in the study [3,5]. The stages used by the research team were “in development,” “in pilot,” “in implementation,” “in use,” and “after implementation [or in use] evaluation” (ie, proof of merit). Thus, for example, theoretical research was excluded.</td>
</tr>
<tr>
<td>SNOMED CT core benefits</td>
<td>Refers to research team’s summary of which areas of identified benefits the research added value to, if available [3,5]. The categories were “improving quality of care and patient safety”; “improving continuity of care”; “enabling a consistent way of indexing, storing, retrieving, and aggregating clinical data”; “improving data quality”; and “improving coding productivity.”</td>
</tr>
</tbody>
</table>

<sup>a</sup>EHR: electronic health record.

<sup>b</sup>SNOMED CT: Systematized Nomenclature of Medicine–Clinical Terminology.

Results

Characteristics of the Publications

In total, we analyzed 17 articles [17-33]. The earliest research selected for the final review was published in June 2017 and the last in February 2022. On an annual basis, the largest number of articles was published as late as 2021 (6/17). The country of publication was identified for all publications. The country that published the highest number of articles was the United Kingdom (5/17), followed by the United States (3/17), Australia (2/17), and Spain (2/17). Canada, Denmark, Switzerland, the Republic of Korea, and Germany each accounted for one use case. The selected articles were published in 13 different peer-reviewed journals. The characteristics of the publications are summarized in Multimedia Appendix 2.

Contextual Factors of the Clinical Use Cases

To verify the appropriateness of the use cases in regard to our research questions, the clinical context was described for 14 articles. In the remaining 3 articles (3/17), the clinical context was not made, but the research account otherwise described the clinical use case in relation to EHR use. With respect to specialties, 2 cases were from neurology, and 1 case each from pulmonology (asthma), cardiology, oncology, general medicine, pediatrics, and rare diseases. One of the cases did not describe an exact specialty but concerned the prehospital unit in emergency care, and one concerned an outpatient clinic. Four cases mentioned either primary care or tertiary care.

The EHRs were poorly described in most of the articles, and specific descriptions of the EHRs did not follow any uniform structure. Thus, the nature of the results is descriptive. Only one of the publications named the exact product. That study concerned a comprehensive hospital information system, a high-maturity EHR with tools for functions such as supporting care coordination and continuity of care. One of the systems was described as a prehospital patient record that was not integrated into the hospital EHR. One system used by general practitioners consisted of integrated software for clinical use, and one was described as a primary care EHR. Six systems (6/17) were hospital EHRs. Among the system types were also the following: an “outpatient and inpatient EHR,” a “centralized EHR with web-interface,” and a “local EHR.” One of the use cases described the system generally as an “ehr.”

To further verify the clinical orientation of the use cases, we analyzed the professional groups involved in each of them. Although users were described, it was not clearly stated which professional groups were the intended users of SNOMED CT (eg, nursing informatics, medical informatics, or multi-professional users). Seven of the use cases (7/17) described the users in an exact way. Four cases (4/17) were
applied by physicians, 2 (2/17) by nurses, and 1 by physicians and nurses. Two of the cases (2/17) were applied by a multi-professional team of clinicians, clinical and medical informatics professionals, clinical domain experts, terminologists, and clinical coders. Two of the articles specifically concerned clinical coders. Two of the remaining cases were generally described as having been applied by “clinicians.” Four of the cases (4/17) concerned researchers themselves, for whom specific clinical backgrounds were not reported. Contextual factors are presented in Multimedia Appendix 3.

**SNOMED CT Use Purpose**

All 17 articles described one or several use purposes for the terminology. Implementing SNOMED CT as a standard terminology in the EHR was typically grounded on clinical needs for standardizing patient information. Here, the SNOMED CT use purpose refers specifically to the primary use goal of the terminology as integrated or being implemented into an EHR. The most common category of use purpose (8/17) was SNOMED CT adopted as a common standard for EHRs. An additional 2 studies (2/17) described the goal of implementing SNOMED CT as a standard in a separate clinical application integrated in the EHR system. The use cases described communication and coordination needs, such as between hospital units or between inpatient and outpatient care, with the goal of promoting more reliable continuity of care and ultimately a higher quality of care. Accurate and timely diagnosis information with SNOMED CT deployment was reported as a crucial clinical need since it is major information in patient care. In 2 of the studies, SNOMED CT was implemented into the documentation of the problem list to increase the usefulness of the patient information and to organize the problem list content. Additionally, SNOMED CT was used to ensure effective data migration between systems.

The primary use purpose described in 2 articles was to retrieve or analyze patient data for clinical research. This enhanced data retrieval, analysis, and sharing for clinical research across multiple hospitals. Multisite data sharing and distributed analysis was supported by common terminology and by common data models. These 2 use cases utilized a medical annotation toolkit that included a web interface for extracting needed concepts. An additional 2 studies focused on data extraction, where SNOMED CT was used to classify and code patient data for research purposes. The use purpose in these 2 studies was building clinical pathways and patient selection criteria based on terminology coding. Natural language processing of clinical, pathology, and genomics data was used for further clinical research. Moreover, the use cases illustrated the challenges of data sharing between inpatient care and a virtual hospital visit.

Two of the original articles described, to a degree, the already established use of SNOMED CT with a focus on proving the merit of the terminology use in a clinical setting. The reasons for poor clinical coding of patient data after 2 decades of EHR use are manifold; two main reasons are lack of motivation and training. Support tools for the interoperable recording of diagnostic, treatment, and interventional patient information can be advanced, for example, with domain-specific development. One of the articles documented automated clinical coding as the driving purpose for SNOMED CT development. The development of computer-assisted coding may, through careful review and validation, improve the productivity of clinical coders. Different classification systems, such as the International Classification of Diseases–10, are typically linked and mapped to SNOMED CT for suitability in clinical use. An overview of the SNOMED CT use purposes is provided in Multimedia Appendix 3.

**SNOMED CT Use Phase**

The phase of use was identified in the literature with varying accuracy, which is why the research team discussed these categories during the analysis. Clinical use phase was documented in all 17 articles (100%), but in ambiguous ways. The most common category of use phases was SNOMED CT in development, which was documented in 6 articles (6/17). Development was described as an iterative process of analysis, validation, and standardization or building and mapping EHR-structured content that requires coordination and communication between stakeholders to improve the quality of care; as such, this was expected to be a process that could span several years.

SNOMED CT in use was identified as the use phase in 5 EHR-related use cases (5/17) and in implementation in 4 use cases (4/17). In the EHR-related use cases, SNOMED CT had been chosen as the base terminology system in the EHR or in a specific domain documented in the research to improve clinical information recording and coding, develop clinical pathways, and extract clinical data. The implementation cases addressed specifically improved the clinical recording of patient data by supporting clinicians’ language and semantic selection with SNOMED CT or with a combination of SNOMED CT and other classification or terminology systems. In addition, 1 article documented a pilot use of SNOMED CT in EHR use cases, and 2 articles described the merits of SNOMED CT use with a more proven merit approach or through after-implementation evaluation. The pilot case evaluated cases of missing or mislabeled clinical data with, for example, nonstandard concepts or use of abbreviations. The evaluation of SNOMED CT use aimed to determine what had been achieved with fully integrated EHR services in patient care and to evaluate the impact of using SNOMED CT to record clinical meanings. As additional benefits, the secondary use purpose for using patient information was mentioned. An overview of the SNOMED CT use phase is provided in Multimedia Appendix 3.

**Core Benefits of SNOMED CT**

The research team identified and summarized core benefits of SNOMED CT as documented in the 17 use cases. The team categorized the benefits based on the previous literature (Multimedia Appendix 3). The core benefits were related to terminology use outcomes. The most common category was increased data quality, with 8 articles (8/17). Semantic-level core benefits were built on the scope and comprehensiveness of the terminology. In the use cases, SNOMED CT supported not only clinical meaning standardization but also the language of choice. For clinical use purposes, custom concept dictionaries or language-specific subsets were built for a chosen language.
Further benefits of implementing SNOMED CT were 2-fold: the parallel development of EHR technology and standardization. One UK use case documented evidence of increased interface usability and user satisfaction by clinicians. However, clinicians reported that adopting a new approach for data recording was a gradual process requiring time.

Four articles (4/17) documented the benefits category of enabling a consistent way of indexing, storing, retrieving, and aggregating clinical data. One use case concentrated on the benefits for data retrieval. Documented benefits pointed to the documentation of clinical events with richer detail. The productivity of coding was the main benefit categorized in 1 use case, increased quality of care in 2 use cases, and increased continuity of care in 1 use case. These benefits depended on the possibility of accessing more complete and coherent patient information, regardless of where it was recorded, to support safe patient care. Additionally, in 1 use case, the core benefit was successful implementation of a new EHR through harmonizing data structures. An overview of the SNOMED CT core benefits is provided in Multimedia Appendix 3.

Discussion

Summary of Findings

This systematic review identified 17 articles in which SNOMED CT was implemented and used in a clinical context in EHRs or related clinical applications. We aimed to confirm whether research has developed to allow for a shift of focus from previously published reviews that described potential use toward studies documenting plausible benefits of SNOMED CT. We present findings related to clinical use purposes, use phases, and core benefits of SNOMED CT over the last 5 years of ongoing efforts. These review categories are based on previous research [3,5,13] that provided a strong starting point for this analysis.

The use purpose for SNOMED terminology based on previous research (Table 1) was identified in all the articles reviewed. As we evaluated the use cases, these categories served our research material well. The most applied use purpose category was SNOMED CT as the planned standard for EHRs or other related applications. Often frequently applied use categories were the goals of using SNOMED for retrieving and analyzing patient data or implementing the terminology to advance the coding of patient data. Only 2 of the articles in the review entailed proof of merit of EHR implementation as the use purpose category. Based on these results, the initial observation was that there might be a level of interconnectedness between the use purpose and use phase. To prove this, data on the maturity of EHR solutions are needed to research the possible interconnectedness of EHR use and SNOMED CT. Moreover, it might be relevant to analyze relationships between use purpose and use phase. This requires testing the categories and their possible relationships with different data sets.

Regarding the use phase results, all the reviewed articles included descriptions of the SNOMED CT use phase, although details of related contextual factors, such as clinical environment, varied. This hampered the assessment of the overall picture of the use phase. Considering the results, this may be a feature of this specific material, being typical of EHR-related use cases of SNOMED CT. Thus, in the future, it may prove fruitful to pay particular attention to the descriptions of these types of SNOMED CT use cases. We propose to describe the phase of use in a more structured and contextual manner. This kind of accuracy would increase the scalability of the use-case results. Lee et al [3] have already highlighted that only a few SNOMED CT implementation cases are being published in the scientific literature. Through the systematic investigation of previous theoretical work [3], and with time, more comparable scientific publications on SNOMED CT use cases in a clinical context could be published.

Based on our review, there is still little research evidence on the benefits for clinical use of SNOMED CT in EHR-related use cases. We identified the following frequently reported categories of core benefits: improvement of data quality and enabling a consistent way of indexing, storing, retrieving, and aggregating clinical data. Closely related to these benefits were improvement in quality of care—with the goal of achieving better patient safety—and, based on better data quality, enabling better continuity of care. Additionally, the review found individual remarks on improving the productivity of coding though automation or through terminological support for clinical users. Such tools had potential to increase user satisfaction, although there was evidence for a need to involve clinicians from different domains in development. Evidence of practical advancement may motivate various clinical specialties to become more involved in SNOMED CT development work from early on.

Although previous research has categorized the possible benefits of SNOMED CT [3,5], the core benefits in our review were summarized by the research team. By doing this, we aimed to describe the specific benefits of the EHR-related use of SNOMED CT. The objective is to highlight that the categorization applied in this study (Table 1) requires further testing with different data to assess its validity. The set of categories applied by our research team was not proven to be comprehensive or exhaustive. Therefore, in future research, it could prove relevant to carefully evaluate SNOMED CT use cases from the perspective of typical benefits. Additionally, it could be important to research what kinds of disadvantages, risks, and bottlenecks can be detected. By evaluating different use cases, it might be possible to extract the general success factors of clinical SNOMED CT implementation. Overall, the evaluation of SNOMED CT implementation requires more attention. As an example, the European large-scale implementation of SNOMED CT, which is being funded by the European Commission, could at the same time advance evaluation studies or require systematic evaluation as a part of the funding process.

Our review revealed that the EHR system or related software were poorly described in most of the articles, and specific descriptions of the EHRs were scattered due to a lack of uniform structure for such descriptions. This is clearly an issue that would require more attention in future use-case descriptions, given the fact that SNOMED CT is designed to support the use of EHRs. To describe the capabilities and overall maturity of
the EHR system is core information in use case descriptions if the research result aims to benefit the clinical implementation process by avoiding previous obstacles and possible mistakes. We recognize that addressing potential mistakes does not mean that specific implementation experiences are universally generalizable, but that such implementation is required to be adapted to other clinical contexts. To promote the scalability of previous experiences, we suggest, for example, the application of the electronic medical record maturity model (EMRAM) in future use cases. EMRAM is a widely used tool developed by the Healthcare Information and Management Systems Society to measure the rate of adoption of EHR functions in health care settings. Its stages match the technological progress of the overall digitalization of the health care setting. Moreover, one possible starting point for future studies is to recognize specific use cases for software applications in clinical specialties, in which SNOMED CT would be valuable to accelerate and facilitate specific types of clinical implementations.

Limitations
This systematic review has limitations that could affect the plausibility of the results. The methods and results of our systematic review are transparently reported in detail to allow readers to assess the trustworthiness and applicability of our findings [16]. However, even though the review’s methodological basis [16] is scientifically recommended, the varying levels of description in the research articles proved to be challenging; for example, the descriptions of background variables led to partial imprecision in the results. This especially affected the categories regarding the clinical context and type of EHR. The study’s risk of bias was carefully considered during the research process, and no assumptions were made about missing or unclear information from the studies.

Conclusions
This literature review demonstrates that systematic reviews are relevant to the development of an understanding of SNOMED CT use and its possible benefits in further facilitating multi-professional, clinically driven implementations by summarizing essential findings based on evidence-based results. Clinical use cases are needed to promote the scalability of review results. To achieve the best out-of-use case reports, more emphasis should be placed on describing the contextual factors, such as the electronic health care system currently in use and the use of previous frameworks, to allow the comparability of results. Regarding future research, although other systematic reviews have addressed similar questions to ours, this review is necessary to shift the focus onto more clinically grounded implementation outcomes and benefits of the use of SNOMED CT. Generally, further research evidence is still needed to determine how exactly SNOMED CT benefits clinical care and patient information quality.

Acknowledgments
The authors acknowledge Finnish governmental research funding (TYH2019244) provided for this study.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Outline of the review protocol applied in this research.
[DOCX File, 20 KB - medinform_v11i1e43750_app1.docx ]

Multimedia Appendix 2
Characteristics of the publications.
[DOCX File, 20 KB - medinform_v11i1e43750_app2.docx ]

Multimedia Appendix 3
Overview of the results.
[DOCX File, 22 KB - medinform_v11i1e43750_app3.docx ]

References


Abbreviations

EHR: electronic health record
EMRAM: Electronic Medical Record Adoption Model
SNOMED CT: Systematized Nomenclature of Medicine–Clinical Terminology

©Riiukka Vuokko, Anne Vakkuri, Sari Palojoki. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 06.02.2023. 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.
Structure of Health Information With Different Information Models: Evaluation Study With Competency Questions

Anna Rossander¹, MD; Daniel Karlsson², PhD

¹Department of Applied IT, University of Gothenburg, Gothenburg, Sweden
²Swedish eHealth Agency, Stockholm, Sweden

Corresponding Author:
Anna Rossander, MD
Department of Applied IT
University of Gothenburg
Department of Applied Information Technology, Division of Informatics
Box 100
Gothenburg, 405 30
Sweden
Phone: 46 735989141
Email: anna.rossander@gu.se

Abstract

Background: There is a flora of health care information models but no consensus on which to use. This leads to poor information sharing and duplicate modelling work. The amount and type of differences between models has, to our knowledge, not been evaluated.

Objective: This work aims to explore how information structured with various information models differ in practice. Our hypothesis is that differences between information models are overestimated. This work will also assess the usability of competency questions as a method for evaluation of information models within health care.

Methods: In this study, 4 information standards, 2 standards for secondary use, and 2 electronic health record systems were included as material. Competency questions were developed for a random selection of recommendations from a clinical guideline. The information needed to answer the competency questions was modelled according to each included information model, and the results were analyzed. Differences in structure and terminology were quantified for each combination of standards.

Results: In this study, 36 competency questions were developed and answered. In general, similarities between the included information models were larger than the differences. The demarcation between information model and terminology was overall similar; on average, 45% of the included structures were identical between models. Choices of terminology differed within and between models; on average, 11% was usable in interaction with each other. The information models included in this study were able to represent most information required for answering the competency questions.

Conclusions: Different but same same; in practice, different information models structure much information in a similar fashion. To increase interoperability within and between systems, it is more important to move toward structuring information with any information model rather than finding or developing a perfect information model. Competency questions are a feasible way of evaluating how information models perform in practice.

(JMIR Med Inform 2023;11:e46477) doi:10.2196/46477

KEYWORDS
informatics; health care; information model; terminology; terminologies; interoperability; competency question; interoperable; competency; EHR; electronic health record; guideline; standard; recommendation; information system

Introduction

Background

Increased use of standards is often suggested as part of the solution to the problem of siloed and unusable information in electronic health records (EHRs) [1], but there is no consensus yet on what standards to use [2]. Instead, there is a flora of standards and the same information is structured with different standards in different settings [3-5]. There are different types of information standards. Some standards primarily aim to
structure information within systems (intraoperability), whereas some are geared toward sharing information (interoperability) [6], but often, both types of standards may be used in both settings. The standards differ between and within themselves regarding the “boundary problem” [7,8], that is, the demarcation between what information is structured with the information model and what is structured with terminology or values. The standards also differ regarding if terminology is stated or not and if so which terminology. Additionally, the terminologies are sometimes standards in themselves (eg, Logical Observation Identifiers Names and Codes [LOINC] or Systematized Nomenclature of Medicine Clinical Terms [SNOMED CT]) but sometimes system-specific or information model–specific value sets.

This combination of possibilities leads to a flora of informatics components, seemingly nonreusable between settings, as noted in previous works [2,9]. However, if the information models structure information in similar ways and there is some agreement on terminologies, perhaps a way forward would be to continue using different standards. Information exchange would be facilitated but not plug and play, as the content would be similar, and the workload of structuring health care information could be shared between users of different information models. Previous work has compared system configurations in relation to a single standard and showed that different system configurations could be unified [10,11]. Works comparing different standards have shown discrepancies in coverage and lack of alignment, primarily regarding terminologies [12,13]. Our hypothesis is that the differences between information models are overestimated. This work contributes by evaluating both the amount and type of differences between models and by providing and testing a method for comparing structure and terminology choices of different standards.

Aim

This work aims to explore if a possible solution to the challenge of sharing information and burden of modelling work within health care would be to continue using different information models. This work also aims to assess the usability of competency questions (CQs) for the evaluation of information models within health care.

Research Questions

The objective of this study was to answer the following 2 research questions:

1. How does the content of health care information differ between information models?
2. Is the method of CQs a feasible way of comparing content in information models?

Methods

Choice of CQs

There are quantitative methods to evaluate information structures in use today. For example, the CAMMS (Common Assessment Method for Standards and Specifications) [14] is an established guide in Europe for assessing a wide range of aspects with primarily a quantitative outcome. The aim of this work was, however, to examine how a sample of clinically relevant information is structured with different information models and not how much of the information the models could structure. To expose how information was structured, a method with qualitative results, that is, including structure and terminology of content, was needed. CQs have been used to evaluate ontologies for a long time [15]. In brief, the ontology is tested by selecting a relevant scenario and then posing questions to the ontology to see if and how the information needed to describe the scenario is structured within the ontology. CQs yield both quantitative and qualitative data. To our knowledge, CQs have not yet been used to evaluate the combination of information model and terminology within health care.

Development of CQs

Domain knowledge has been used as the basis for CQs previously by, for example, Cui [16]. Guidelines are an established textual source for domain knowledge within health care, and we thus chose to use recommendations in a guideline as scenario for developing the CQs. The use of information needed to follow best practice as a starting point ensured that the study examined the tested information models regarding clinically relevant information as opposed to theoretical possibilities or boundaries. Any topic within health care could have been used as a starting point for this work. Central venous lines (also called central catheters) are one of the many domains where structured documentation could support adherence to best practice and facilitate research to develop best practice. To prevent misinterpretations due to translation during the work, we chose to work with Swedish guidelines [17]. In the chosen guideline, there were 104 recommendations that covered preparation, insertion, care, and removal of central venous catheters. Examples from the guidelines are as follows:

1. The tip of the central venous line should be placed distally in the superior caval vein or the right atrium, and the location should be controlled at the time of insertion.
2. Bandages with polyurethane film should be replaced every 3-5 days during inpatient care.
3. At emergency insertion of a central venous line, the advantages of a central venous line should be weighed against the risk of hemorrhage.

The recommendations in the guideline were graded as beneficial, equivocal, or harmful. All recommendations graded as beneficial were placed in a random order and CQs were developed iteratively from the top. Formulating the CQs is a semantic task. The recommendation was read, and if needed, divided into sections, and then questions were developed to cover all the information mentioned in the recommendation. An example is shown in Figure 1.
Recommendations that contained information not documented in the patient record, for example, “All departments using central venous lines should have access to blood- and catheter tip–culture techniques” were considered out of scope in this study and omitted. When the recommendations were not specific enough to develop CQs, the associated text in the guideline was used to interpret and operationalize the recommendation. For example, for the recommendation “Radiographic control after central venous line insertion should be performed if pneumothorax or hemothorax is suspected,” the text “Patients with pneumothorax who need treatment show new respiratory symptoms (dyspnea or cough) or oxygen saturation in blood lower than 90%” and “the risk increases with technical difficulties” was used to interpret the patients who had conditions indicative of pneumothorax or hemothorax.

The purpose of the study determines the number of CQs developed and used [18]. The focus of this work was on comparing how the different models structured clinically relevant information rather than the entire scope of each information model. During data collection, new CQs were iteratively developed and posed until further questions did not add additional types of clinical information. This is defined as the saturation point [19,20]. Despite this, a gap was discovered during data analysis regarding anatomical locations, which had not been covered by any of the initial CQs. Therefore, the next 2 recommendations in the randomized list containing information about the anatomical location were included as well. In total, 36 CQs based on 10 recommendations were developed. See Multimedia Appendix 1 for the list of included recommendations and developed CQs.

Materials

The information models tested and compared in this study were a purposive sample. In this study, the information models are the “participants” that were selected “based on the researchers’ judgment about what potential participants will be most informative” [18]. The intention was to compare some typical models that were in use already and some models that were often recommended. The included information models differ in nature in several aspects, but they are all aimed at structuring data and thus impact interoperability. See Table 1 [21-32] for the included information models.
Table 1. Information models included in this study.

<table>
<thead>
<tr>
<th>Information model</th>
<th>Description by information model provider</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Information standards</strong></td>
<td>FHIR is a standard for health care data exchange, published by HL7. OpenEHR is a nonprofit organization that publishes technical standards for an EHR platform along with domain-developed clinical models to define content. HCIMs are used to capture functional semantic (nontechnical) agreements for the standardization of information used in the care process. IPS CDA is the ISO reference model for electronic health record standards. The OMOP Common Data Model allows for the systematic analysis of disparate observational databases. SPOR is the Svenskt Perioperativt Register.</td>
</tr>
<tr>
<td>FHIR(^a) [21]</td>
<td>FHIR is a standard for health care data exchange, published by HL7. OpenEHR is a nonprofit organization that publishes technical standards for an EHR platform along with domain-developed clinical models to define content. HCIMs are used to capture functional semantic (nontechnical) agreements for the standardization of information used in the care process. IPS CDA is the ISO reference model for electronic health record standards. The OMOP Common Data Model allows for the systematic analysis of disparate observational databases. SPOR is the Svenskt Perioperativt Register.</td>
</tr>
<tr>
<td>openEHR [22]</td>
<td>OpenEHR is a nonprofit organization that publishes technical standards for an EHR platform along with domain-developed clinical models to define content. HCIMs are used to capture functional semantic (nontechnical) agreements for the standardization of information used in the care process. IPS CDA is the ISO reference model for electronic health record standards. The OMOP Common Data Model allows for the systematic analysis of disparate observational databases. SPOR is the Svenskt Perioperativt Register.</td>
</tr>
<tr>
<td>HCIM(^d) [24]</td>
<td>HCIMs are used to capture functional semantic (nontechnical) agreements for the standardization of information used in the care process. IPS CDA is the ISO reference model for electronic health record standards. The OMOP Common Data Model allows for the systematic analysis of disparate observational databases. SPOR is the Svenskt Perioperativt Register.</td>
</tr>
<tr>
<td>IPS CDA(^e) [26]</td>
<td>The goal of this project is to identify the required clinical data with associated vocabulary bindings and value sets for patient summary and to build an international document and associated templates based on HL7 CDA R2 with value sets to support data elements within those templates. IPS CDA is the ISO reference model for electronic health record standards. The OMOP Common Data Model allows for the systematic analysis of disparate observational databases. SPOR is the Svenskt Perioperativt Register.</td>
</tr>
<tr>
<td><strong>Standards for secondary use</strong></td>
<td>The OMOP Common Data Model allows for the systematic analysis of disparate observational databases. The concept behind this approach is to transform data contained within those databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes) and then perform systematic analyses by using a library of standard analytic routines that have been written based on the common format. SPOR is the Svenskt Perioperativt Register.</td>
</tr>
<tr>
<td>OMOP(^f) [28,29]</td>
<td>The OMOP Common Data Model allows for the systematic analysis of disparate observational databases. The concept behind this approach is to transform data contained within those databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes) and then perform systematic analyses by using a library of standard analytic routines that have been written based on the common format. SPOR is the Svenskt Perioperativt Register.</td>
</tr>
<tr>
<td><strong>SPOR(^g) [31]</strong></td>
<td>The purpose of SPOR is to, by means of integration with existing local operation planning systems, retrieve data from the perioperative process and thus offer a tool for local and national quality development (translation by authors). IPS CDA is the ISO reference model for electronic health record standards. The OMOP Common Data Model allows for the systematic analysis of disparate observational databases.</td>
</tr>
<tr>
<td><strong>System-specific formats</strong></td>
<td>A table with the recommendations and corresponding CQs was developed. For each model, the authors together modelled the information needed to answer the CQs based on information available on the internet about the models. Both the structure of the information model, that is, what archetype/profile/entry included models. Further, the 2 EHR systems offer a wide possibility for users to configure templates paired with limited reference information models, and thus, the results in this work provide examples of use in the selected EHR systems. The results might have been different if an application domain other than central venous lines had been used.</td>
</tr>
<tr>
<td>Electronic health record A</td>
<td>A health care information system used by approximately 70,000 health care staff. The OMOP Common Data Model allows for the systematic analysis of disparate observational databases. The concept behind this approach is to transform data contained within those databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes) and then perform systematic analyses by using a library of standard analytic routines that have been written based on the common format. SPOR is the Svenskt Perioperativt Register.</td>
</tr>
<tr>
<td>Electronic health record B</td>
<td>A health care information system developed and supplied by a global vendor. The OMOP Common Data Model allows for the systematic analysis of disparate observational databases. The concept behind this approach is to transform data contained within those databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes) and then perform systematic analyses by using a library of standard analytic routines that have been written based on the common format. SPOR is the Svenskt Perioperativt Register.</td>
</tr>
</tbody>
</table>

\(^a\)FHIR: Fast Healthcare Interoperability Resources.  
\(^b\)HL7: Health Level 7.  
\(^c\)HCIM: Health and Care Information Model.  
\(^e\)OMOP: Observational Medical Outcomes Partnership.  
\(^f\)SPOR: Svenskt Perioperativt Register.  

Information about the information standards and standards for secondary use were sought on publicly available sources online. Fast Healthcare Interoperability Resources (FHIR) and openEHR have national and local profiles in addition to the internationally published standards available, for example, on Simplifier [33] and in national or local clinical knowledge manager repositories [22]. An initial survey of these resources did not show profiles directly focused on the application domain, and hence, these resources were not included. The information standards are in continuous development; the latest available version was used and cited (see individual references). Draft versions were included when there was no published version of a relevant component. The Health and Care Information Models (HCIMs) are a precursor for International Organization for Standardization (ISO) 13972 [34], which were not yet published as a standard when work began, and they were therefore used as an example of that standard. The International Patient Summary (IPS) is published as both Clinical Document Architecture (CDA) and FHIR. Since FHIR was included separately, the IPS CDA format was chosen.

Templates from 2 EHR systems were included. The material consisted of locally configured user interfaces of the systems and not an information model or database model; thus, the types of results differ between the 2 EHR systems and the other included models. Further, the 2 EHR systems offer a wide possibility for users to configure templates paired with limited reference information models, and thus, the results in this work provide examples of use in the selected EHR systems. The results might have been different if an application domain other than central venous lines had been used.

SPOR (Svenskt Perioperativt Register) and the 2 EHR systems often structure the information according to the specific situation where the templates are used, as opposed to the information standards and OMOP (Observational Medical Outcomes Partnership), which are intended to be general purpose. Thus, for SPOR and the EHR systems, we have included examples of data elements where generic elements do not exist. For example, SPOR had a data element for “kind of venous access,” with access devices in the value set, where the information standards often had a generic device type data element, which could hold any device type.

**Answering the CQs**

A table with the recommendations and corresponding CQs was developed. For each model, the authors together modelled the information needed to answer the CQs based on information available on the internet about the models. Both the structure of the information model, that is, what archetype/profile/entry
and element was used, and terminological content, that is, what code/codesystem/unit, was documented. As an illustration, the answers for the CQ “Does the patient have new onset dyspnea?” are displayed in Table 2. For details, please see Multimedia Appendix 2. Note that the CQ does not specify what “new” means, that is, in terms of hours or days, but to determine if a symptom is new by any definition, the time of onset is needed.

Table 2. Example results for “Does the patient have new onset dyspnea?” for selected information models.

<table>
<thead>
<tr>
<th>Element and value</th>
<th>Value set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FHIR condition resource</strong>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>SNOMED CT&lt;sup&gt;b&lt;/sup&gt; descendents of 404684003 [Clinical finding (finding)] (Example)</td>
</tr>
<tr>
<td>condition.code = 267036007 [Dyspnea (finding)]</td>
<td>ISO&lt;sup&gt;c&lt;/sup&gt; 8601</td>
</tr>
<tr>
<td>condition.onsetTime</td>
<td></td>
</tr>
<tr>
<td><strong>IPS CDA&lt;sup&gt;d&lt;/sup&gt; IPS problem entry</strong></td>
<td>SNOMED CT CORE Problem List Disorders (preferred)</td>
</tr>
<tr>
<td>hl7:value = 267036007 [Dyspnea (finding)]</td>
<td>ISO 21090 → ISO 8601</td>
</tr>
<tr>
<td>hl7:effectiveTime</td>
<td></td>
</tr>
<tr>
<td><strong>OMOP&lt;sup&gt;e&lt;/sup&gt; condition occurrence</strong></td>
<td>SNOMED CT or ICDo3&lt;sup&gt;f&lt;/sup&gt;</td>
</tr>
<tr>
<td>condition.concept.ID = 267036007 [Dyspnea (finding)] (no code for dyspnea in ICDo3)</td>
<td>ISO 21090 → ISO 8601</td>
</tr>
<tr>
<td>condition_start_date or condition_start_datetime</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>FHIR: Fast Healthcare Interoperability Resources.
<sup>b</sup>SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.
<sup>c</sup>ISO: International Organization for Standardization.
<sup>d</sup>IPS CDA: International Patient Summary Clinical Document Architecture.
<sup>e</sup>OMOP: Observational Medical Outcomes Partnership.
<sup>f</sup>ICDo3: International Classification of Diseases for Oncology Third Edition.

For some standards, the same information could be structured in several ways. For example, with FHIR, the information needed to answer, “What day and time was the central venous line inserted?” could be structured with both a Procedure Resource and a DeviceUseStatement. With openEHR, both Evaluation Medical Device and Action Procedure could be used. In these cases, all options were documented as results. The answers to the questions were influenced by the knowledge of the modelers answering them. The background knowledge that the authors have together was estimated to be comparable to that of a system implementer. One of the authors (DK) is, by training, a computer scientist and health informatician with experience in, for example, European Committee for Standardization and ISO standards and EHR system configuration as well as SNOMED CT. The other author (AR) is a medical doctor and health informatician with experience in structuring quality registers, SNOMED CT, and EHR system configuration. Since the authors performed the modelling, they were both researchers and participants at the same time. This gave extra insights and understanding of the work performed but also introduced a risk of bias. However, none of the authors have either any background or held any position with any of the above organizations that biased the results in any way.

**Assessment of Coverage**

The models were graded for content coverage by type of clinical information. Coverage was graded into “structured” if the information needed to answer the CQs for that type of information was structured. It was graded “partially structured” when only parts of the information were structured. “Not structured” was used when there was no structure and “missing” when the information was not present in the information model.

**Assessment of Content Differences**

Tables were developed for structure and coding. For each combination of models, the way or different ways the information was structured was evaluated. The number of possible ways for each model was used as the denominator and the number of ways that were similar enough to be used in interaction with the other model was used as the numerator, and the ratios of the 2 compared models were multiplied. For example, procedure type could be structured in only 1 way in FHIR but in 2 separate ways in SPOR (see Table S1 in Multimedia Appendix 2)—all 3 had a similar distribution of information between element and value, and this thus gave the following result: 1/1 × 1/2 = 50%.

Another example is procedure status where HCIM had no status field but instead used a time stamp (see Table S2 in Multimedia Appendix 2). The other models, if anything, had a coded value for status, and the result for HCIM was thus 0 for all combinations for this type of information: 0/1 × x/y = 0%.

The value sets were assessed separately in the same fashion. Where a model had several possible terminologies, those that would have been used for the information needed to answer the CQs were used. The SNOMED CT Global Patient Set [35] and full SNOMED CT were considered usable in interaction, and LOINC terms that were in the same LOINC group were also considered sufficiently similar. For model-specific value sets, common in, for example, status elements, the included values were compared and if they were equivalent, this counted as
usable in interaction. All binding strengths have been assigned equal weight. When no value set was recommended, this was given the value 0.

Definitions

Information models, terminologies, and ontologies are all developed to structure information about things. In some subject areas, ontologies are in themselves sufficient to structure most information, whereas in health care, information models and terminologies or ontologies are usually used in conjunction. A division between information model and terminology is useful, as it makes the requirements on the terminology less complex. However, it leads to what is sometimes called the boundary problem, that is, the difficulty of deciding what information should be structured with the information model and what information should be structured with the terminology [7,8] (see example in Figure 2).

Figure 2. Examples of different ways of using information model and terminology in conjunction.

<table>
<thead>
<tr>
<th>allergy</th>
<th>allergy substance</th>
<th>allergy reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>anaphylaxis after nuts</td>
<td>nuts</td>
<td>anaphylaxis</td>
</tr>
</tbody>
</table>

reaction and allergen in the same term reaction and allergen in different elements in an information model

When not otherwise stated, we use the term information model for an information model, including its terminology bindings, that is, the terminologies or ontologies stated in it, when present. The term element is used for parts of the information model, sometimes also called attributes or headings. A value set is the stated terms or codes that are allowed for a certain element. Sometimes, a value set included entire terminology, for example, LOINC, International Classification of Diseases Tenth Revision (ICD-10), or SNOMED CT. Value set specifications may provide a binding strength to describe the flexibility with which members can be used while being compatible with the value set definition. For example, the FHIR framework provides 4 levels of binding strength: required (value set cannot be changed, eg, by extension), extensible (value set can be extended), preferred (the value set is recommended but not mandatory), and example (value set is an example only) [36]. For easier comparison of results, we have interpreted the binding strengths in the information models and described them using the FHIR definitions above.

Results

Research Question 1: How Does the Content of Health Care Information Differ Between Information Models?

The answers to the CQs included repeating types of information. The types of information are listed in Table 3, and the results below are presented per type. Note that the results depict the information needed to answer the CQs and not an overall evaluation of that information type.

<table>
<thead>
<tr>
<th>Table 3. Types of information.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanation</td>
</tr>
<tr>
<td>Time and period</td>
</tr>
<tr>
<td>Procedures</td>
</tr>
<tr>
<td>Conditions</td>
</tr>
<tr>
<td>Causalities</td>
</tr>
<tr>
<td>Medications</td>
</tr>
<tr>
<td>Device types</td>
</tr>
<tr>
<td>Results of examinations</td>
</tr>
<tr>
<td>Complex professional judgments</td>
</tr>
</tbody>
</table>

Coverage

Results regarding coverage, that is, what information the included models had capacity to structure in a way that allowed answering of CQs, is provided per information model and type of information in Figure 3. With some exceptions, the differences in coverage were small between the included information models. The information standards and OMOP had the broadest coverage, providing structure for most types of information. SPOR could only structure information that was
requested when the registry was constructed. The EHRs could hold more information than the table implies, but some of the information was locked into structures, making it difficult to find or use it in other situations. For example, in EHR B, information about radiographic control after insertion of a central venous catheter could be found under the heading “use,” where the options were “accepted for use,” “may be used before radiographic control,” “may not be used before radiographic control,” and “other.” Structures like this in EHR B were designed per instantiation and thus likely vary between settings.

**Figure 3.** Coverage per information model for competency questions. EHR: electronic health record; FHIR: Fast Healthcare Interoperability Resources; HCIM: Health and Care Information Model; IPS CDA: International Patient Summary Clinical Document Architecture; OMOP: Observational Medical Outcomes Partnership; SPOR: Svenskt Perioperativt Register.

<table>
<thead>
<tr>
<th></th>
<th>FHIR</th>
<th>openEHR</th>
<th>HCIM</th>
<th>IPS CDA</th>
<th>OMOP</th>
<th>SPOR</th>
<th>EHR A</th>
<th>EHR B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time and period</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>P</td>
</tr>
<tr>
<td>Procedure type</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>Procedure status</td>
<td>S</td>
<td>S</td>
<td>M</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>Procedure body location</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>N</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>Condition type</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>Condition status</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>Causality</td>
<td>S</td>
<td>S</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>Medications</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>M</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>Device types</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>Results of observations</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>Complex professional judgments</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

**Content Differences**

Overall, the differences regarding the structure of information between the included information models were small. On average, 45% of the included structures were identical between models, that is, had the same demarcation between information model and terminology (Figure 4). The choice of terminology, however, showed a greater variation with, on average, only 11% overlap between models (Figure 5). Differences regarding structure were smaller than those regarding terminology (see Figure 6 for results per information type). Qualitative data on the content, that is, how information was structured and terminology was used, are presented as per the type of information below. The full result tables are provided in Multimedia Appendix 2. In the value set columns in the tables of Multimedia Appendix 2, all value sets listed within the information model are provided, although not all of them were relevant to the CQs. However, in our analysis, only the relevant value sets were considered.

<table>
<thead>
<tr>
<th></th>
<th>Total average</th>
<th>Pairwise comparison average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>openEHR</td>
<td>HCIM</td>
</tr>
<tr>
<td>FHIR</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>openEHR</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>HCIM</td>
<td>40%</td>
<td>50%</td>
</tr>
<tr>
<td>IPS CDA</td>
<td>50%</td>
<td>70%</td>
</tr>
<tr>
<td>OMOP</td>
<td>50%</td>
<td>40%</td>
</tr>
<tr>
<td>SPOR</td>
<td>40%</td>
<td>50%</td>
</tr>
<tr>
<td>EHR A</td>
<td>30%</td>
<td>40%</td>
</tr>
<tr>
<td>EHR B</td>
<td>40%</td>
<td>70%</td>
</tr>
</tbody>
</table>

Figure 5. Percentage of terminologies usable in interaction with each other. EHR: electronic health record; FHIR: Fast Healthcare Interoperability Resources; HCIM: Health and Care Information Model; IPS CDA: International Patient Summary Clinical Document Architecture; OMOP: Observational Medical Outcomes Partnership; SPOR: Svenskt Perioperativt Register.

<table>
<thead>
<tr>
<th></th>
<th>Total average</th>
<th>Pairwise comparison average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>openEHR</td>
<td>HCIM</td>
</tr>
<tr>
<td>FHIR</td>
<td>20%</td>
<td>10%</td>
</tr>
<tr>
<td>openEHR</td>
<td>10%</td>
<td>40%</td>
</tr>
<tr>
<td>HCIM</td>
<td>20%</td>
<td>10%</td>
</tr>
<tr>
<td>IPS CDA</td>
<td>20%</td>
<td>40%</td>
</tr>
<tr>
<td>OMOP</td>
<td>20%</td>
<td>10%</td>
</tr>
<tr>
<td>SPOR</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>EHR A</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>EHR B</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>
Time and Period

Time is repeated in many different types of structures and thus not comparable in the same way as the other information types; hence, no table for time and period is provided in Multimedia Appendix 2. The information standards and OMOP used ISO 21090 [37] and ISO 8601 [38]. Since ISO 21090 is based on ISO 8601, they are equivalent in this setting. It was not possible to determine the exact format for the Swedish SPOR or the EHRs. Some modules in EHR B only handled time of documentation, as opposed to time of the actual event, procedure, or discovered condition. There were differences between the information models on how periods were represented. FHIR and IPS CDA used interval data types, while HCIM relied on having distinct data elements for start and end points of the period. openEHR had 3 different approaches. For action archetypes, periods could be deduced from the time difference between time-stamped events. For observation archetypes, time-related information was represented through the reference model, and for evaluation archetypes, distinct data elements were used for temporal information, similar to HCIMs.

Procedures

Information about procedures contain type of procedure (eg, insertion of central venous line), status of the procedure (eg, completed), and sometimes a location where the procedure was performed (eg, left subclavian vein). In FHIR, openEHR and OMOP procedures using a device could also be structured with the device as central information (for results regarding this, see “Device Types” below).

Procedure Type

All information standards and OMOP had a coded element within a dedicated procedure structure. SPOR and the 2 EHRs additionally had procedure-specific elements with a Boolean value. Of the information standards, none but HCIM strictly bound the procedure type to a terminology. HCIM had a binding strength of required and mandated the use of a code element in their ISO 21090–inspired CD datatype. The most commonly recommended terminologies in the included information models were SNOMED CT and the Swedish procedure classification KVÅ (Klassifikation av vårdätgårdar [39]; Swedish version and extension of Nordic Medico-Statistical Committee Classification of Surgical Procedures [40]) due to the Swedish context of some of the included information models. FHIR and IPS CDA used SNOMED CT, whereas HCIM and OMOP allowed several different value sets. The information models that used SNOMED CT pointed to different subsets. The Swedish registry SPOR and the 2 EHRs used KVÅ. They also, at times, used the term or code for the procedure as a question answered with a Boolean, for example, “C250 fluoroscopy during the procedure: yes/no.”

Procedure Status

All models that had a stated status used a separate element for this. HCIM had no explicit representation of procedure status; instead, time could be both in the past and future, indicating performed or planned procedures. It was unclear how planned but not performed procedures could be discerned from performed procedures as time passes. The standards for secondary use only represented performed procedures, and this was sufficient for answering the CQs in this work. The EHRs had multiple structures. FHIR, openEHR, and IPS CDA used coded text with native value sets, which were not always one-to-one mappable between each other.

Procedure Body Location

The body location of a procedure can be represented either within the value for the procedure type (see above) or in a separate element. All models except OMOP had one or many ways to separately structure body location. FHIR, openEHR, HCIM, and SPOR also had additional elements for laterality or location qualifiers. In FHIR and openEHR, these were placed in the reference model.
in an extension and cluster, respectively. Many procedures have multiple possible locations, for example, regarding placing a central venous line—relevant location includes place of insertion (e.g., left arm), the vessel in which the catheter is placed (e.g., upper caval vein), and catheter tip location (e.g., left atrium). None of the information models had a means to express the role of the body location in the procedure. All information standards, except openEHR, used SNOMED CT body structures. SPOR and the 2 EHRs used system-specific value sets.

Conditions

Condition Type
The demarcation between information model and terminology was identical for conditions in all the investigated models, except SPOR, which had an additional structure with separate Boolean elements for key conditions. Most of the compared information models used SNOMED CT as terminology, followed by ICD-10.

Condition Status
All information standards, OMOP, and EHR B, structured the status of the condition in a separate element. The information standards had at least 2 elements to capture both status (e.g., present, resolved, absent) and certainty of the status (e.g., unconfirmed, established, suspected). The difference between status of a condition and the certainty of the condition can lead to ambiguities; for example, in the FHIR Condition Resource, it was possible to have an active (clinicalStatus) and at the same time refuted (verificationStatus) condition. Two of the openEHR code sets and the code sets in HCIM contained codes from SNOMED CT but used different concepts, and the value sets were fully disjoint. All other codes were information model–specific. Most code sets had different granularity, that is, number of codes, making one-to-one mapping between them difficult.

Causality

Causality is a relation between entities where one is the cause of another, for example, that a deep vein thrombosis is a consequence of a central venous catheter. Of the included models, only openEHR had a separate element to document causality. openEHR also contained a LINK class, which would support this purpose, but there was no generic code set for the type of linking. In FHIR, there was a “dueTo” extension, which allowed linking conditions to their causes. The CDA standard had provisions for linking any CDA instance to any other instance, but this feature was not used for causality within the IPS CDA implementation guide. Both EHRs had structured lists for specific settings, for example, “reason for extraction” with local codes in the value set.

Medications

Only information regarding the type of medicinal product has been evaluated in this work. Information structures of timing, dosage, dose form, and substances were not included. The structure of information on medications varied depending on stage in the process of medication, that is, for example, prescribing, dispensing, administration, or consumption. There were 2 general patterns: one where there was a single coded element for the medicinal product and one where there was a complex structure of multiple elements, such as active ingredient, dose, and dose form. The information models that had terminology bindings pointed to multiple terminologies, except FHIR that stated SNOMED CT. EHR A used the Anatomic Therapeutic Chemical classification system and the Swedish national medicinal products terminology [41]; the configuration for EHR B was not finished at the time of data collection.

Device Types

Devices vary from short-time use artefacts as dressings to permanent implants as pacemakers. The results show how information about the type of devices was structured when the device was the central information. Several of the models had further elements for additional details, for example, batch number, size, or manufacturer. This was not included in this work. The information standards and OMOP had dedicated elements for devices with the name of the device in the value set, whereas SPOR and the EHRs used a terminology-bound element with a Boolean or a value set to further specify the device. FHIR, HCIM, IPS CDA, and OMOP all pointed to SNOMED CT as terminology. In general, it was also possible to document information about a device within the procedure where it was used as a distinct procedure type. This can be done with a term or concept where the device is included, for example, 1172566008 [Insertion of central venous catheter (procedure)], but FHIR also permitted a separate element within the procedure class holding the device (in this example, 52124006 [Central venous catheter, device (physical object)]).

Results of Observations

Common observations are bedside measurements, assessment scales, and laboratory results. Information regarding observations is often a combination of a question, a result value, and a unit. Sometimes these entities were structured in separate elements, and sometimes, a part of the information was structured by terminology binding the element. For example, a measurement of the oxygen saturation could be structured into “Measurement = oxygen saturation, value = 98, and unit = %” as well as “oxygen saturation in percent = 98.” There were 2 distinct approaches to representing the results of the observations. FHIR, IPS CDA, OMOP, and EHR A rely on external terminologies to express the type of observation, whereas openEHR and HCIM develop specific information models to express the type of observation where the name of the element was bound to a terminology. SPOR and EHR B had specific elements in the information model for observations but without any terminology binding.

Complex Professional Judgments

None of the included models had a structured way to document complex judgments such as “How big problems can be expected if the central venous line is replaced?” or “Is the patient’s life at risk?”

https://medinform.jmir.org/2023/1/e46477
Research Question 2: Are CQs a Feasible Way of Comparing Content in Information Models?

Development of CQs

In total, 36 CQs covering 10 recommendations were developed (Table 4). For 7 of the recommendations, the information in the recommendation was enough to develop the CQs—a task performed in a few minutes. For 2 recommendations, additional information from the guidelines was needed. One recommendation required information on what substances were included in “ADP (adenosine diphosphate) receptor antagonists” and “novel oral anticoagulants,” which was not present in the guidelines.

Table 4. Information needed to develop the competency questions.

<table>
<thead>
<tr>
<th>Recommendation only</th>
<th>Recommendation and textual guideline</th>
<th>Additional information needed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommendations (n=10), n</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Competency questions (n=36), n</td>
<td>19</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>

As described in the Methods section, the initially assumed saturation point was revised during analysis of results, and additional CQs were developed for 2 recommendations.

Answering the CQs

The most effort in data gathering was spent on searching information about the information models and modelling. CQs covering information frequently documented in a structured way, for example, “Does the patient have renal impairment?” were relatively straightforward to answer with all the included models. For information that is rarely structured, for example, “Was a micro puncture needle used?” or “What problems can be expected if the central venous line is replaced?” much time was spent on searching information about the different models to minimize risk that a possible solution was missed. The amount of work performed in modelling the information needed for the CQs is comparable to that performed in a real-life setting modelling clinical information. Time consumption thus varied widely both depending on complexity of the area and how well the chosen information model handled the area.

Assessment of Coverage and Content Differences

The results from the modelling work were complex, especially when information could be structured in several ways with the same information model or when terminology binding included multiple value sets. This was demanding to capture in a spreadsheet, but evaluation of tools was beyond the scope of this work.

Discussion

Research Question 1: How Does the Content of Health Care Information Differ Between Information Models?

When compared pairwise, the 8 included models had, on average, 45% identical structures and 11% terminologies that were sharable. Most overlaps regarding structure were present between information standards. Content that is not identical can still be similar, and our assessment is that the similarities were larger than the differences between the compared information models in general. The information models included in this study could represent most information required for answering the CQs.

Structure

Conditions and procedures have the highest overlap in structure. This information is thus readily sharable despite using different models if the used terminologies are the same or translatable. Representations of observations could be expected to be well-standardized due to its maturity but had only 40% overlap, mainly due to 2 different patterns of demarcation between information model and terminology. The same problem was present for medications and device types. Another common demarcation issue, present also for procedures and conditions, was that the EHRs and SPOR commonly used complex elements with a yes/no tick box, that is, with Boolean data type, whereas the information standards split the same information into several elements. Complex elements with tick boxes are tempting when developing structures for a specific use case but makes information sharing with structures developed for other use cases difficult. Conversion between these types of demarcation may be possible but includes risk of information loss and builds maintenance burden. In this material, the body locations for procedures and conditions could be represented either in a complex value for the procedure/condition itself (eg, insertion of central venous line in left subclavian vein or kidney failure) or separated in different ways (see examples in Figure 7). Similar issues can occur with other related information such as method.
Figure 7. Examples of using terminology or separate elements and classes. A. Using 1 element in 1 class instance. B. Using 2 elements in 2 class instances. C. Using 2 elements in 1 class instance. D. Using 3 elements in 2 class instances.

When sharing information structured with different demarcations, a compositional terminology for describing both the elements of the information model and the values in the model can be beneficial [42]. A compositional terminology allows for composition and decomposition of meaning, for example, splitting “kidney failure” into “organ failure” and “kidney” or vice versa. The information standards and OMOP all refer to SNOMED CT as a possible terminology for this type of information, and SNOMED CT logic representation may, in select cases, be used to transfer between different demarcations. For this to be possible, both the element name and the value must be terminology bound to concepts or postcoordinated expressions that are logically defined in relation to each other. Some use cases such as laterality are in that sense likely to be easier to coordinate, while others may introduce significant complexities regarding postcoordination or development of new concepts. Sometimes, different elements were mandatory in different models. In such cases, it might not be possible to share data even when the information is decomposable because obligatory information might be missing.

Terminologies

When the demarcation between information model and terminology is the same and the only difference between models is regarding terminology, information sharing possibilities depend on how easily those terminology-encoded values are converted into each other. All models used internal model-specific value sets for some elements. For example, in this material, the values for status for both procedures and conditions were different in all the included models, not only regarding terms but also by the number of values, making one-to-one mapping very difficult, not to say impossible, without information loss or distortion. Some values were, however, present in the code sets of all information models; for example, all models had a value to represent the status “the patient has this condition now” and that could thus be mapped between models. This confirms previous work, which showed that Apgar score representation had similar structures in HL7v3 DMIM (Health Level 7 version 3 Domain Message Information Model) and openEHR but were poorly bound to terminology [2] and that few value sets were aligned between the models when comparing openEHR and 3 HL7 formats for adverse sensitivity [12], although in the latter case, a joint openEHR-FHIR review has improved alignment [43].

The openEHR archetypes studied in this work were outliers compared to the other included information models in that few specific external terminologies were referred to. According to openEHR methodology, terminology binding is postponed to the templating phase, but while reviewing international templates, no additional terminology bindings were found. In this material, FHIR, HCIM, IPS CDA, and OMOP on the other hand often referred to external international terminologies, especially for larger value sets.

Where the models point to an existing terminology, there will be times when a suitable concept does not exist and therefore needs to be developed. In this material, there was, for example, no suitable concept within SNOMED CT to document the type of bandage used, despite SNOMED CT being the recommended terminology for several of the models. SNOMED CT provides the possibility to post coordinate concepts. However, postcoordination has drawbacks; for example, many health care information systems lack the capability to handle postcoordinated expressions, and postcoordinated expressions lack a human-readable term. Further, the concept model must permit the needed modelling, and the concepts needed for modelling must exist or be created [3]. Postcoordination has thus not been included as a possibility for value sets based on SNOMED CT.

Using the same terminology does not necessarily mean that the exact same code is used. For example, SNOMED CT contains 233527006 |Central venous cannula insertion (procedure)|, which has 16 more granular child concepts, and any of these could be used to document the insertion of a central venous cannula.
Internal Variability

The information standards aim to cover a wide range of information and offer complex structures to achieve this. They also sometimes have several different ways to structure the same information on varying levels of detail, leading to internal variability. This has been shown in evaluations of implementations of information models [13,44]. The standards for secondary use had a more rigid structure, only permitting 1 way to structure per type of information. The EHRs aim to capture all information and rely on free text to a higher degree than the other included types of models. Free text is, however, very hard to share unambiguously. Some information relevant in this work was structured very specifically in the EHRs, for example, “radiographic control before use of central venous catheter.” Other types of radiography procedures were not examined but it is unlikely that all radiography examinations are structured like this, and this is thus an example of the same procedure being structured in multiple ways also in the EHRs.

In FHIR and openEHR, structured information could be added in extensions or slots. In our material, this was found for causality (FHIR), procedure body location (FHIR and openEHR), and medication detail (openEHR). These might be tempting if the other option is free text; however, additions like this risk add to complexity and internal variability. Where information can be structured in several ways, there is a risk that instantiated information is erroneous, for example, an “upper arm fracture in the leg.” Having multiple elements to construct the meaning of a clinical statement increases the need for the sophisticated validation of information either during or after data entry to avoid mishaps.

Areas of Poor Coverage

Complex professional judgments were not possible to structure with any of the included models. Perhaps complex professional judgments are most easily documented as free text. Placing them in a terminology-bound element in an information model would facilitate identifying and sharing the information despite it being unstructured. openEHR and FHIR were the only included models that could structure causality, both by using extension and slots, thereby opening the potential for a higher degree of variability.

Research Question 2: Are CQs a Feasible Way of Comparing Content in Information Models?

The CQ method was a good way to probe deeper into information models from a clinically relevant perspective. The CQs revealed the types of information that were poorly structured or completely omitted—areas that are easily overseen when assessing the same information model from a theoretical perspective. One could argue that CQs leave the doorajar to bias from the evaluator, as opposed to a more formal method where the information models are described in the same format and then compared [45,46]. However, the information models relevant to compare are only available in different formalisms, often specific to the respective model, thereby restricting the use of such formal methods. Further, formal comparisons between models that differ in their demarcation between information model and terminology is not possible unless all elements are terminology bound to a machine-readable terminology, which they rarely are. Care should be taken when deciding how much effort to put into answering the CQs. Those covering information rarely structured are laborious to answer and perhaps more useful as a marker for where information is so complex that free text is the most suitable way to document it.

To cover all possible types of information, it might be necessary to push beyond the initially deemed saturation point. Perhaps, future work could use a 2-step saturation point by first developing CQs until no new types of information are uncovered and then modelling answers for the developed CQs, omitting those that duplicate already performed modelling work. This would avoid massive duplicative modelling.

Limitations

The CQs do not push the boundaries of what the information models can handle. For example, multiple body locations or status other than “performed” or “present” for procedures or conditions are not included nor are many of the intricacies about pharmaceutical information. There might be greater differences between models than this work has revealed. The information modelling done in this work is best effort but not best possible. The modelling was not discussed with additional parties external to this study; however, this might be in correspondence with results in a real-life setting, where the amount of effort is limited by existing resources, including access to domain and informatics experts.

Conclusions

Formal comparisons between information models show incompatibilities that are often merely theoretical [4,47], whereas practical work has shown that conversion between models for secondary use is doable [5,10,11]. This work shows that in practice, different information models structure much information in a similar fashion. To increase interoperability within and between systems, it is thus more important to move toward structuring information with any information model than finding or developing a single, perfect information model. When choosing an information model, one should consider that international standards have the best coverage and overlap between information models. They are also likely to be more widely adopted, decreasing the need for conversion before information exchange and have more users putting effort into developing them. As a final delimiter, assess the demarcation between information model and terminology and choose an information model that is similar to those of whom information is to be shared with. Put effort into decreasing internal variability and increasing terminology binding to external terminologies. The CQ method was successfully applied to the challenge of comparing health care information models. This method is a feasible way of evaluating how information models perform in practice, thereby adding valuable qualitative data on similarities and differences.
Acknowledgments
AR developed the competency questions. AR and DK performed modelling, analysis of results, and writing together. Thanks to Eva Blomqvist for contribution to initial ideas and anonymous reviewers for valuable comments.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Recommendations and corresponding competency questions.
[DOCX File, 17 KB - medinform_v11i1e46477_app1.docx ]

Multimedia Appendix 2
Structure of content per type of information.
[DOCX File, 55 KB - medinform_v11i1e46477_app2.docx ]

References
7. Markwell D, Sato L, Cheetham E. Representing clinical information using SNOMED clinical terms with different structural information models. 2008 Presented at: Proceedings of the Third International Conference on Knowledge Representation in Medicine; May 31-June 2; Phoenix, Arizona, USA.


21. HL7. FHIR v4.3.0. URL: http://hl7.org/fhir/ [accessed 2022-07-08]

22. openEHR. Clinical Knowledge Manager. URL: https://ckm.openehr.org/ckm/ [accessed 2022-07-08]

23. About us. openEHR. URL: https://www.openehr.org/about_us [accessed 2022-08-12]


25. HCIM Mainpage. URL: https://zibs.nl/wiki/HCIM_Mainpage [accessed 2022-08-12]


27. International Patient Summary - Project Information. URL: https://art-decor.org/art-decor/decor-project--hl7ips- [accessed 2022-08-12]


30. OMOP (Observational Medical Outcomes Partnership) common data model. Observational Health Data Sciences and Informatics. URL: https://www.ohdsi.org/data-standardization/the-common-data-model/ [accessed 2022-08-12]


33. SIMPLIFIER.NET. The FHIR Collaboration Platform. URL: https://simplifier.net/ [accessed 2022-07-11]


36. FHIR v4.3.0 Codesystem binding strength. HL7 FHIR. URL: http://hl7.org/fhir/R4B/codesystem-binding-strength.html [accessed 2022-10-31]


Abbreviations

CAMMS: Common Assessment Method for Standards and Specifications
CDA: Clinical Document Architecture
CQ: competency question
DMIM: Domain Message Information Model
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
HCIM: Health and Care Information Model
HL7: Health Level 7
ICD-10: International Classification of Diseases Tenth Revision
IPS: International Patient Summary
ISO: International Organization for Standardization
KVÅ: Klassifikation av vårdåtgärder
LOINC: Logical Observation Identifiers Names and Codes
OMOP: Observational Medical Outcomes Partnership
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
SPOR: Svenskt Perioperativt Register

Edited by C Lovis; submitted 19.02.23; peer-reviewed by C Gaudet-Blavignac, J Ehrsam; comments to author 17.03.23; revised version received 11.05.23; accepted 03.06.23; published 31.07.23.

Please cite as:
Rossander A, Karlsson D
Structure of Health Information With Different Information Models: Evaluation Study With Competency Questions
JMIR Med Inform 2023;11:e46477
URL: https://medinform.jmir.org/2023/1/e46477
doi:10.2196/46477
PMID:37523221

©Anna Rossander, Daniel Karlsson. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 31.07.2023. 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.
The Impact of an Electronic Portal on Patient Encounters in Primary Care: Interrupted Time-Series Analysis

Karen Ferguson¹,², MD, CCFP; Mark Fraser², MD, CCFP; Meltem Tuna³,⁴, MSc, PhD; Charles Bruntz⁵, MSc; Simone Dahrouge¹,⁶, MSc, PhD

¹Department of Family Medicine, University of Ottawa, Ottawa, ON, Canada
²West Carleton Family Health Team, Carp, ON, Canada
³ICES, Ottawa, ON, Canada
⁴Ottawa Hospital Research Institute, Ottawa, ON, Canada
⁵Champlain Family Health Teams, Ottawa, ON, Canada
⁶Bruyère Research Institute, Ottawa, ON, Canada

Corresponding Author:
Karen Ferguson, MD, CCFP
Department of Family Medicine
University of Ottawa
600 Peter Morand Crescent #201
Ottawa, ON, K1G 5Z3
Canada
Phone: 1 613 562 5800 ext 2982
Email: karen@wcfht.ca

Abstract

Background: Electronic patient portals are online applications that allow patients access to their own health information, a form of asynchronous virtual care. The long-term impact of portals on the use of traditional primary care services is unclear, but it is an important question at this juncture, when portals are being incorporated into many primary care practices.

Objective: We sought to investigate how an electronic patient portal affected the use of traditional, synchronous primary care services over a much longer time period than any existing studies and to assess the impact of portal messaging on clinicians’ workload.

Methods: We conducted a propensity-score–matched, open-cohort, interrupted time-series evaluation of a primary care portal from its implementation in 2010. We extracted information from the electronic medical record regarding age, sex, education, income, family health team enrollment, diagnoses at index date, and number of medications prescribed in the previous year. We also extracted the annual number of encounters for up to 8 years before and after the index date and provider time spent on secure messaging through the portal.

Results: A total of 7247 eligible portal patients and 7647 eligible potential controls were identified, with 3696 patients matched one to one. We found that portal registration was associated with an increase in the number of certain traditional encounters over the time period surrounding portal registration. Following the index year, there was a significant jump in annual number of visits to physicians in the portal arm (0.42 more visits/year vs control, \( P < .001 \)) but not for visits to nurse practitioners and physician assistants. The annual number of calls to the practice triage nurses also showed a greater increase in the portal arm compared to the control arm after the index year (an additional 0.10 calls, \( P = .006 \)). The average provider time spent on portal-related work was 5.7 minutes per patient per year.

Conclusions: We found that portal registration was associated with a subsequent increase in the number of some traditional encounters and an increase in clerical workload for providers. Portals have enormous potential to truly engage patients as partners in their own health care, but their impact on use of traditional health care services and clerical burden must also be considered when they are incorporated into primary care.

(JMIR Med Inform 2023;11:e43567)  doi:10.2196/43567
Introduction

Electronic patient portals are online applications that allow patients access to their own health information, a form of asynchronous virtual care. There has been a great deal of recent interest in patient portals, accompanied by increasing technology adoption by both clinicians and patients [1-3]. The COVID-19 pandemic has also highlighted the importance of virtual care, an area already identified as a national health care priority [4]. Although portal features vary, the safe communication channels in portals may provide alternative ways for patients to obtain services traditionally provided in person, such as renewing prescriptions, sending and receiving secure messages, obtaining test results, and booking appointments [5]. A recent survey indicated that approximately 20% of Canadians had accessed some of their own medical information electronically, and that almost 80% were interested in doing so [6]. However, that survey did not specifically address portals or patient access to their medical information in primary care practice settings, and we are not aware of any studies examining Canadian portal adoption in primary care. Our understanding of the potential value of patient portals is nascent, with portals expected to contribute to more authentic collaboration between clinicians and patients.

The long-term impact of portals on traditional primary care services is unclear, but it is an important question at this juncture, when portals are being incorporated into primary care practices. Many studies reporting on the impact of portals on the use of traditional services evaluated systems that only provided options for web messaging or booking appointments [7-13]. All existing studies that investigated portals with more diverse features were conducted in medical networks, such as health maintenance organizations, where the portals provided access across sectors, including primary care, specialty care, and hospital care; these studies may not be relevant to portals incorporated into exclusively primary care practices. Past studies also reported inconsistent findings regarding the impact of portals on traditional health care use. Some studies demonstrated an increase in visits [14-16] or telephone calls [17]. Others demonstrated no change in visits [18], a reduction in visits [19], or a reduction in hospital readmissions [20]. All these studies also had limited time frames, examining only the period 12 to 30 months after portal registration.

To our knowledge, no long-term evaluation of the impact of a primary care patient portal on traditional health care use has been conducted to date. Providers have expressed interest in patient portals but also concerns regarding medicolegal risk and clerical workload [21]. Some have described an increased clerical burden associated with portals as part of the electronic health environment [12,22]. For instance, a qualitative study examining online patient access to their own health records found that providers felt that their workload had increased as a result [23], while another found that some providers anticipated fewer administrative requests for information when patients had access to their own health records [24]. One study found that online patient access to encounter notes did not significantly affect physician workload [25], although others have described high volumes of portal messages sent by patients [26]. However, no studies have actually tracked the provider time spent specifically on portal-related work. There have also not been any large studies of the impact of electronic patient portals in a Canadian setting. We sought to investigate how an electronic patient portal affects traditional, synchronous, primary care health care use over a much longer time period than any existing studies, and to assess the impact of portal messaging on clinicians’ workload.

Methods

We conducted a propensity-score–matched, open-cohort, interrupted time-series (ITS) evaluation of a primary care portal from its implementation in 2010.

Setting and Study Participants

The practice was a semirural interprofessional clinic in southeastern Ontario, Canada, where 12 family physicians and other allied health providers provide comprehensive primary care under a single-payer model. Under this publicly funded model, physician compensation is primarily through capitation payments for rostered patients. The primary care patient portal initially offered access to laboratory results, the ability to receive secure messages (in 2012), send secure messages (in 2015), book appointments (in 2016), and renew prescriptions (in 2018.) All practice patients were invited to join the portal via email, posters, and telephone reminders and at in-person encounters.

We collected data for all practice patients except those seen exclusively for focused care (eg, obstetrical care). We retained data from all patients only for the period they were aged 18 years or older. Among patients who adopted the portal, we excluded those for whom we did not have at least one year of data prior to and following their portal registration (ie, index) date. For non–portal patients, we excluded those who did not have at least two consecutive years of data between 2009 and 2019.

Matching

We calculated propensity scores to estimate the probability of individuals registering for the portal using logistic regression [27,28]. Propensity scores were derived based on sex, age, whether the patient was rostered to the family health team, the presence of specific diagnoses on the index date, and the number of in-person and telephone encounters, as well as the number of unique medications prescribed in the 12 months prior to the index date. During the study time period, all appointments with medical doctors (MDs), nurse practitioners (NPs), and physician
assistants (PAs) were in person. Since education and income level were recorded for approximately a third of patients, these measures were not included in the propensity score matching. Control patients were entered into the equation for each year they were eligible (ie, for each year they had at least one year of data prior to and after the index date), with their corresponding profile for that year. July 1 of that year was considered the index date.

Variables, Data Sources, and Measurement

The study period was January 2002 through December 2019. We extracted electronic medical record information on patient age, sex, education, income, enrollment with the practice, and presence or absence of specific diagnoses on the index date. We also extracted the dates of in-person encounters with MDs, NPs, and PAs; dates of triage calls (TCs) to the practice triage nurses; and prescription dates and prescribed medications. Prescribed medications included only those that were identified as distinct medications using Anatomical Therapeutic Chemical codes. Diagnoses were defined based on diagnostic codes, using the earliest date when the diagnostic code was applied.

In order to study the clinician workload associated with the portal, two providers (KF and MF) time stamped their portal messages between February 20, 2020, and February 25, 2021. This allowed us to estimate the average provider time spent per message. We also collected the total number of portal messages sent by all providers to all portal patients between January 1, 2019, and December 31, 2019, in order to determine the average amount of time spent per patient on portal-related work.

Analyses

We described the profile of eligible patients prior to matching on their index date for portal patients on July 1 of the median year for which they were eligible to be matched for non–portal patients and again for the matched patients on their index date in both arms. The main study outcome was the frequency of in-person encounters with MDs, NPs, or PAs, as well as frequency of TCs. We used an ITS design to evaluate the impact of portal registration on use of these traditional health care services over time and compared use by portal users to their frequency of TCs. We used an ITS design to evaluate the impact of the intervention as a dummy variable. Intervention and time interaction is also included in the model to identify the effect of the intervention on both arms (ie, portal and non–portal) over time.

We plotted the annual number of in-person encounters with MDs, in-person encounters with NPs or PAs, and TCs across time and overlaid the estimates derived from the ITS equations. During the study time period, all appointments with MDs, NPs, and PAs were in person. Because the year-0 results showed a spike in service use in both study groups, likely related to the attribution of the index date, we excluded that year from the ITS model. Also, although the spike in service use at year 1 in the portal arm may represent a transient change in behavior associated with the initial adoption, we also excluded this from the ITS to obtain a more reliable estimate of the impact of portal adoption over time, recognizing that this approach omits significant use; this should be considered in result interpretation.

We also depicted the number of visits per calendar year for patients who adopted the portal grouped by year of portal registration to demonstrate the pattern of changes in these visits over time for the intervention arm.

Ethics Approval

Ethics approval was received from the Bruyère Research Ethics Board (M16-20-012).

Results

Matching

Of the 14,894 patients who met the study criteria, 7247 (48.7%) were portal participants. Of these, 3696 were matched one to one with a control patient (Figure 1). The profile of all eligible patients before and after propensity matching is shown in Table 1. Before matching, portal users differed from non–portal users, but after matching, the mean propensity scores of the 2 groups and their index years, the prevalence of chronic conditions, sex, rostering status, and total visits and medications in the previous years showed good agreement. Income and education levels, which could not be included in the propensity score derivation because of poor data completeness, remained higher in the portal group.

We used a caliper of 0.2 for matching and limited the potential matching pool for each portal patient to non–portal patients with an index date that was within 1 year of the portal patient’s index date. We identified all potential controls for each portal patient and assigned matches prioritizing first portal patients who had a unique match, then non–portal patients who had a unique portal match. We repeated this after each match to minimize loss of controls. When more than one match was possible, we attributed the control patient whose propensity score was closest to the portal patient’s score. The balance of baseline covariates between the matched portal users and non–portal users was assessed using standardized differences, with values <0.1 representing negligible differences.
Figure 1. Study patient selection.
Table 1. Portal and control patients before and after matching.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Portal, index date (n=7247)</th>
<th>Non-portal, median index date (n=7647)</th>
<th>Total (n=14,894)</th>
<th>Portal, index date (n=3696)</th>
<th>Non-portal, median index date (n=3696)</th>
<th>Total (n=7392)</th>
<th>P value</th>
<th>Standard difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>On index date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propensity score, mean (SD)</td>
<td>0.21 (0.10)</td>
<td>0.10 (0.08)</td>
<td>0.15 (0.11)</td>
<td></td>
<td></td>
<td></td>
<td>&lt;.001</td>
<td>1.28</td>
</tr>
<tr>
<td>Propensity score groups (participants), n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.001</td>
<td>.10</td>
</tr>
<tr>
<td>0</td>
<td>291 (4)</td>
<td>2687 (35.1)</td>
<td>2978 (20)</td>
<td>0.85</td>
<td>738 (20)</td>
<td>740 (20)</td>
<td>1.478</td>
<td>0.82</td>
</tr>
<tr>
<td>1</td>
<td>947 (13.1)</td>
<td>2032 (26.6)</td>
<td>2979 (20)</td>
<td>0.34</td>
<td>735 (19.9)</td>
<td>743 (20.1)</td>
<td>1.478</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>1349 (18.6)</td>
<td>1631 (21.3)</td>
<td>2980 (20)</td>
<td>0.07</td>
<td>743 (20.1)</td>
<td>736 (19.9)</td>
<td>1.478</td>
<td>0.01</td>
</tr>
<tr>
<td>3</td>
<td>2010 (27.7)</td>
<td>968 (12.7)</td>
<td>2978 (20)</td>
<td>0.38</td>
<td>740 (20)</td>
<td>742 (20.1)</td>
<td>1.478</td>
<td>0.01</td>
</tr>
<tr>
<td>4</td>
<td>2650 (36.6)</td>
<td>329 (4.3)</td>
<td>2979 (20)</td>
<td>0.87</td>
<td>740 (20)</td>
<td>735 (19.9)</td>
<td>1.478</td>
<td>0.01</td>
</tr>
<tr>
<td>Index year (participants), n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.001</td>
<td>.99</td>
</tr>
<tr>
<td>2010</td>
<td>1-5</td>
<td>145-152</td>
<td>153 (1)</td>
<td>0.2</td>
<td>1-5</td>
<td>1-5</td>
<td>1-5</td>
<td>0</td>
</tr>
<tr>
<td>2011</td>
<td>1335 (18.4)</td>
<td>522 (6.8)</td>
<td>1857 (12.5)</td>
<td>0.35</td>
<td>654 (17.7)</td>
<td>654 (17.7)</td>
<td>1.478</td>
<td>0</td>
</tr>
<tr>
<td>2012</td>
<td>1240 (17.1)</td>
<td>390 (5.1)</td>
<td>1630 (10.9)</td>
<td>0.39</td>
<td>579 (15.7)</td>
<td>596 (16.1)</td>
<td>1.478</td>
<td>0.01</td>
</tr>
<tr>
<td>2013</td>
<td>948 (13.1)</td>
<td>349 (4.6)</td>
<td>1297 (8.7)</td>
<td>0.3</td>
<td>461 (12.5)</td>
<td>452 (12.2)</td>
<td>1.478</td>
<td>0.01</td>
</tr>
<tr>
<td>2014</td>
<td>531 (7.3)</td>
<td>3717 (48.6)</td>
<td>4248 (28.5)</td>
<td>1.04</td>
<td>296-300</td>
<td>275-279</td>
<td>1.478</td>
<td>0.02</td>
</tr>
<tr>
<td>2015</td>
<td>664 (9.2)</td>
<td>838 (11)</td>
<td>1502 (10.1)</td>
<td>0.06</td>
<td>344 (9.3)</td>
<td>353 (9.6)</td>
<td>1.478</td>
<td>0.01</td>
</tr>
<tr>
<td>2016</td>
<td>970 (13.4)</td>
<td>670 (8.8)</td>
<td>1640 (11)</td>
<td>0.15</td>
<td>515 (13.9)</td>
<td>518 (14)</td>
<td>1.478</td>
<td>0.01</td>
</tr>
<tr>
<td>2017</td>
<td>793 (10.9)</td>
<td>681 (8.9)</td>
<td>1474 (9.9)</td>
<td>0.07</td>
<td>464 (12.6)</td>
<td>455 (12.3)</td>
<td>1.478</td>
<td>0.01</td>
</tr>
<tr>
<td>2018</td>
<td>761-765b</td>
<td>328-332b</td>
<td>1093 (7.3)</td>
<td>0.24</td>
<td>378 (10.2)</td>
<td>388 (10.5)</td>
<td>1.478</td>
<td>0.01</td>
</tr>
<tr>
<td>Age at index date (years), mean (SD)</td>
<td>48.9 (14.9)</td>
<td>45.2 (19.3)</td>
<td>47.0 (17.4)</td>
<td>&lt;.001</td>
<td>46.56 (15.17)</td>
<td>46.18 (15.93)</td>
<td>1.478</td>
<td>0.29</td>
</tr>
<tr>
<td>Sex (participants), n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.001</td>
<td>.74</td>
</tr>
<tr>
<td>Female</td>
<td>4334 (59.8)</td>
<td>3748 (49)</td>
<td>8082 (54.3)</td>
<td></td>
<td>2104 (56.9)</td>
<td>2090 (56.5)</td>
<td>4194 (56.7)</td>
<td>0.01</td>
</tr>
<tr>
<td>Male</td>
<td>2913 (40.2)</td>
<td>3899 (50.1)</td>
<td>6812 (45.7)</td>
<td></td>
<td>1592 (43.1)</td>
<td>1606 (43.5)</td>
<td>3198 (43.3)</td>
<td>0.01</td>
</tr>
<tr>
<td>Rostered, n (%)</td>
<td>7162 (98.8)</td>
<td>7224 (94.5)</td>
<td>14386 (96.6)</td>
<td>&lt;.001</td>
<td>3665 (99.2)</td>
<td>3652 (98.8)</td>
<td>7317 (99)</td>
<td>0.13</td>
</tr>
<tr>
<td>Coronary artery disease, n (%)</td>
<td>243 (3.4)</td>
<td>268 (3.5)</td>
<td>511 (3.4)</td>
<td>.61</td>
<td>68 (1.8)</td>
<td>79 (2.1)</td>
<td>1.478</td>
<td>0.36</td>
</tr>
<tr>
<td>Congestive heart failure, n (%)</td>
<td>79 (1.1)</td>
<td>112 (1.5)</td>
<td>191 (1.3)</td>
<td>.04</td>
<td>24 (0.6)</td>
<td>31 (0.8)</td>
<td>1.478</td>
<td>0.34</td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disease, n (%)</td>
<td>128 (1.8)</td>
<td>235 (3.1)</td>
<td>363 (2.4)</td>
<td>&lt;.001</td>
<td>43 (1.2)</td>
<td>54 (1.5)</td>
<td>1.478</td>
<td>0.26</td>
</tr>
<tr>
<td>Diabetes mellitus, n (%)</td>
<td>440 (6.1)</td>
<td>482 (6.3)</td>
<td>922 (6.2)</td>
<td>.56</td>
<td>149 (4)</td>
<td>154 (4.2)</td>
<td>303 (4.1)</td>
<td>0.77</td>
</tr>
<tr>
<td>Hypertension, n (%)</td>
<td>1343 (18.5)</td>
<td>1190 (15.6)</td>
<td>2533 (17)</td>
<td>&lt;.001</td>
<td>495 (13.4)</td>
<td>474 (12.8)</td>
<td>969 (13.1)</td>
<td>0.47</td>
</tr>
<tr>
<td>Income level (CAD $) (participants), n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>&lt;40,000</td>
<td>233 (8.7)</td>
<td>432 (15.8)</td>
<td>665 (12.3)</td>
<td>0.22</td>
<td>121 (8.7)</td>
<td>150 (12.2)</td>
<td>271 (10.4)</td>
<td>0.11</td>
</tr>
<tr>
<td>40,000-60,000</td>
<td>357 (13.3)</td>
<td>484 (17.7)</td>
<td>841 (15.5)</td>
<td>0.12</td>
<td>173 (12.5)</td>
<td>218 (17.7)</td>
<td>391 (14.9)</td>
<td>0.15</td>
</tr>
<tr>
<td>60,000-100,000</td>
<td>846 (31.6)</td>
<td>836 (30.5)</td>
<td>1682 (31)</td>
<td>0.02</td>
<td>437 (31.5)</td>
<td>397 (32.3)</td>
<td>834 (31.9)</td>
<td>0.02</td>
</tr>
<tr>
<td>&gt;100,000</td>
<td>1242 (46.4)</td>
<td>989 (36.1)</td>
<td>2231 (41.2)</td>
<td>0.21</td>
<td>657 (47.3)</td>
<td>465 (37.8)</td>
<td>1122 (42.9)</td>
<td>0.19</td>
</tr>
<tr>
<td>Education level (participants), n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Note: SD = standard deviation; CAD = Canadian dollars; <.001 is considered statistically significant.
### In the 12 months prior to index date

#### Medical doctor visits\(^e\), n (%)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Portal, index date (n=7247)</th>
<th>Non–portal, median index date (n=7647)</th>
<th>Total (n=14,894)</th>
<th>(P) value</th>
<th>Portal, index date (n=3696)</th>
<th>Control, index date (n=3696)</th>
<th>Total (n=7392)</th>
<th>(P) value</th>
<th>Standard difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>High school or less</td>
<td>526 (21)</td>
<td>1037 (39.7)</td>
<td>1563 (30.5)</td>
<td>0.41</td>
<td>284 (21.7)</td>
<td>380 (32)</td>
<td>664 (26.6)</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>812 (32.4)</td>
<td>745 (28.5)</td>
<td>1557 (30.4)</td>
<td>0.08</td>
<td>397 (30.4)</td>
<td>381 (32.1)</td>
<td>778 (31.2)</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>University or more</td>
<td>1363 (50.5)</td>
<td>979 (35.5)</td>
<td>2342 (42.9)</td>
<td>0.31</td>
<td>626 (47.9)</td>
<td>427 (35.9)</td>
<td>1053 (42.2)</td>
<td>0.24</td>
<td></td>
</tr>
</tbody>
</table>

#### Nurse practitioner or physician assistant visits\(^e\), n (%)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Portal, index date (n=7247)</th>
<th>Non–portal, median index date (n=7647)</th>
<th>Total (n=14,894)</th>
<th>(P) value</th>
<th>Portal, index date (n=3696)</th>
<th>Control, index date (n=3696)</th>
<th>Total (n=7392)</th>
<th>(P) value</th>
<th>Standard difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>High school or less</td>
<td>526 (21)</td>
<td>1037 (39.7)</td>
<td>1563 (30.5)</td>
<td>0.41</td>
<td>284 (21.7)</td>
<td>380 (32)</td>
<td>664 (26.6)</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>812 (32.4)</td>
<td>745 (28.5)</td>
<td>1557 (30.4)</td>
<td>0.08</td>
<td>397 (30.4)</td>
<td>381 (32.1)</td>
<td>778 (31.2)</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>University or more</td>
<td>1363 (50.5)</td>
<td>979 (35.5)</td>
<td>2342 (42.9)</td>
<td>0.31</td>
<td>626 (47.9)</td>
<td>427 (35.9)</td>
<td>1053 (42.2)</td>
<td>0.24</td>
<td></td>
</tr>
</tbody>
</table>

#### Calls to triage nurses\(^e\), n (%)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Portal, index date (n=7247)</th>
<th>Non–portal, median index date (n=7647)</th>
<th>Total (n=14,894)</th>
<th>(P) value</th>
<th>Portal, index date (n=3696)</th>
<th>Control, index date (n=3696)</th>
<th>Total (n=7392)</th>
<th>(P) value</th>
<th>Standard difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>High school or less</td>
<td>526 (21)</td>
<td>1037 (39.7)</td>
<td>1563 (30.5)</td>
<td>0.41</td>
<td>284 (21.7)</td>
<td>380 (32)</td>
<td>664 (26.6)</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>812 (32.4)</td>
<td>745 (28.5)</td>
<td>1557 (30.4)</td>
<td>0.08</td>
<td>397 (30.4)</td>
<td>381 (32.1)</td>
<td>778 (31.2)</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>University or more</td>
<td>1363 (50.5)</td>
<td>979 (35.5)</td>
<td>2342 (42.9)</td>
<td>0.31</td>
<td>626 (47.9)</td>
<td>427 (35.9)</td>
<td>1053 (42.2)</td>
<td>0.24</td>
<td></td>
</tr>
</tbody>
</table>

#### Medications prescribed\(^e\), n (%)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Portal, index date (n=7247)</th>
<th>Non–portal, median index date (n=7647)</th>
<th>Total (n=14,894)</th>
<th>(P) value</th>
<th>Portal, index date (n=3696)</th>
<th>Control, index date (n=3696)</th>
<th>Total (n=7392)</th>
<th>(P) value</th>
<th>Standard difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>High school or less</td>
<td>526 (21)</td>
<td>1037 (39.7)</td>
<td>1563 (30.5)</td>
<td>0.41</td>
<td>284 (21.7)</td>
<td>380 (32)</td>
<td>664 (26.6)</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>812 (32.4)</td>
<td>745 (28.5)</td>
<td>1557 (30.4)</td>
<td>0.08</td>
<td>397 (30.4)</td>
<td>381 (32.1)</td>
<td>778 (31.2)</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>University or more</td>
<td>1363 (50.5)</td>
<td>979 (35.5)</td>
<td>2342 (42.9)</td>
<td>0.31</td>
<td>626 (47.9)</td>
<td>427 (35.9)</td>
<td>1053 (42.2)</td>
<td>0.24</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)Index date for unmatched control patients: July 1 of the median year of their eligible time period.

\(^b\)The value of n for 2010 was smaller than 6. The cells for 2010, 2014, and 2018 therefore do not have precise values due to ethics agreements.

\(^c\)Income and education level were not available for all patients.

\(^d\)CAD $1.00=US $0.75 on January 12, 2023.

\(^e\)Number of medications prescribed and number of encounters refer to the 12-month period prior to the index date. The medication records were mapped to the Drug Product Database to assign Anatomical Therapeutic Chemical codes and schedules based on drug identification numbers. We excluded 12.5% of medication records, including records for medications classified as “over-the-counter” or “ethical” in the schedules from the Drug Product Database (4.9%) and medications reclassified manually as “over-the-counter,” “other,” or “N/A” (4.7%). Medications with no drug identification number were classified manually. Those which could not be attributed a drug identification number were also excluded (2.9%). Variable categorization for number of visits, telephone calls and medications was based on clinical judgement and the number of participants in each category.

### Analyses

We plotted the number of visits in relation to the index year for portal and control patients, the estimated slopes for the years before and after the transition, and the shift in visits at the index date derived from the ITS equation (Figures 2-4). The information for the years prior to the index date for portal patients demonstrates their annual visits for the years prior to their registration on the portal. The index date for the control patients was assigned to be within 1 year of the portal’s patient index date in order to control for temporal factors such as health care use trends. The outputs of the ITS analyses are provided in Table 2. The intercepts and slopes prior to the index year were similar in the control and portal arms for MDs, NPs/PAs, and TCs (\(P > .05\)). After the index year, there was a significant jump in MD visits in the portal arm (0.42 more visits/year vs control, \(P < .001\)) but not for NP or PA visits. The TCs also showed a greater increase in visits in the portal arm compared to the control arm after the index year (0.102 more visits/year vs control, \(P = .006\)).
Figure 2. Interrupted time series for MD face-to-face visits for portal patients versus controls. The intercepts ($P=.86$) and slopes ($P=.15$) prior to the index year were similar in the control and portal arms. After the index year, there was no significant change in the number of MD visits in the control arm. However, in the portal arm, there was a significant jump in number of visits and a new intercept (0.42 more visits/year vs control, $P<.001$). The slope for MD visits increased after the index date in the control arm but became negative in the portal arm, representing an annual reduction of 0.054 visits per year for the portal arm compared to the control arm ($P=.01$). The two slopes would be expected to cross after 10 years. CLI: confidence limit interval; MD: medical doctor.

<table>
<thead>
<tr>
<th>Portal Patients</th>
<th>Control Patients</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preintervention slope</td>
<td>0.023</td>
<td>0.0089</td>
</tr>
<tr>
<td>Postintervention slope</td>
<td>0.010</td>
<td>0.045</td>
</tr>
<tr>
<td>Intervention absolute increase in annual # of visits</td>
<td>0.42</td>
<td>0.0022 &lt;.001</td>
</tr>
</tbody>
</table>

Figure 3. Interrupted time series for nurse practitioner or physician assistant face-to-face visits for portal patients versus controls. The intercepts ($P=.59$) and slopes ($P=.12$) prior to the index year were similar in the control and portal arms. After the index year, there was not a significant change in the number of nurse practitioner or physician assistant visits in the portal arm compared to the control arm ($P=.21$). The slope flattened after the index date in the control arm, but it was relatively unchanged in the portal arm, demonstrating an annual increase of 0.028 visits per year in the portal arm compared to the control arm ($P=.01$). CLI: confidence limit interval; PA: physician assistant; NP: nurse practitioner.

<table>
<thead>
<tr>
<th>Portal Patients</th>
<th>Control Patients</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preintervention slope</td>
<td>0.039</td>
<td>0.049</td>
</tr>
<tr>
<td>Postintervention slope</td>
<td>0.038</td>
<td>0.0071</td>
</tr>
<tr>
<td>Intervention absolute increase in annual # of visits</td>
<td>0.058</td>
<td>-0.011 .21</td>
</tr>
</tbody>
</table>
Figure 4. Interrupted time series for triage calls for portal patients versus controls. The intercepts ($P=0.10$) and slopes ($P=0.26$) prior to the index year were similar in the control and portal arms. The number of triage calls following the index year showed a higher value than anticipated based on the preindex slope in the control arm (0.062 more calls annually, $P=0.02$), but a significantly greater jump after the index year in the portal arm (0.10 more calls annually, $P=0.006$). The slopes for annual triage calls were similar in the pre- and postindex periods for both the control arm and portal arm. CLI: confidence limit interval.

Table 2. Outputs of the interrupted time series. “Annual visits” indicates slope; “period” indicates the pre- or postindex period.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Medical doctor visits</th>
<th>Nurse practitioner or physician assistant visits</th>
<th>Triage calls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>$P$ value</td>
<td>Estimate</td>
</tr>
<tr>
<td>Intercept$^a$</td>
<td>1.572</td>
<td>&lt;.001</td>
<td>0.618</td>
</tr>
<tr>
<td>Annual visits (slope)$^b$</td>
<td>0.009</td>
<td>.19</td>
<td>0.049</td>
</tr>
<tr>
<td>Period (before or after index)$^c$</td>
<td>0.002</td>
<td>.97</td>
<td>-0.011</td>
</tr>
<tr>
<td>Annual visits × period$^d$</td>
<td>0.036</td>
<td>.002</td>
<td>-0.042</td>
</tr>
<tr>
<td>Portal$^e$</td>
<td>0.008</td>
<td>.86</td>
<td>-0.018</td>
</tr>
<tr>
<td>Portal × annual visits$^f$</td>
<td>0.014</td>
<td>.15</td>
<td>-0.010</td>
</tr>
<tr>
<td>Portal × period$^g$</td>
<td>0.417</td>
<td>&lt;.001</td>
<td>0.069</td>
</tr>
<tr>
<td>Portal × annual visits × period$^h$</td>
<td>-0.054</td>
<td>.001</td>
<td>0.028</td>
</tr>
</tbody>
</table>

$^a$Control arm intercept.
$^b$Pre–index date slope of annual visits for the control arm.
$^c$Change in number of visits in year 2 post–index date relative to that anticipated from preindex slope for the control arm.
$^d$Change in the slope of annual visits in the postindex period relative to the preindex period for the control arm.
$^e$Difference between the portal arm and control arm in the intercept.
$^f$Difference between the portal arm and control arm in the pre–index date slope of annual visits.
$^g$Difference between the portal arm and control arm in the change in number of visits in year 2 post–index date relative to that anticipated from preindex slope.
$^h$Difference between the portal arm and control arm in the change of the slope of annual visits in the postindex period relative to the preindex period.
We also plotted the visit rates for each year for patients having enrolled in the portal, grouped by year of portal registration (Figures 5-7).

The 2 physicians who time stamped 2061 portal messages spent an average of 3.83 minutes on each message. We also extracted the total number of portal messages sent by all providers between January 1 and December 31, 2019, and found that an average of 1.49 messages were sent to each portal patient in the practice. Thus, the average amount of provider time devoted to portal messages was estimated to be 5.7 minutes per portal patient per year.

**Figure 5.** Number of visits to medical doctors per calendar year for each patient group (registered on the portal in 2011-2012, 2013-2014, 2015-2016, and 2017-2018). To reduce noise, the number of visits represents the running average of that year, the previous year, and the following year. The MD visits showed a slight increase in the number of visits in the years immediately following portal registration, followed by an apparent drop in annual rate of visits.

![Figure 5](image)

**Figure 6.** Number of visits to nurse practitioners or physician assistants for each patient group (registered on the portal in 2011-2012, 2013-2014, 2015-2016, and 2017-2018). To reduce noise, the number of visits represents the running average of that year, the previous year, and the following year. The NP and PA visits began in 2006 and show a rapid rise in the number of visits until 2010, then a considerable flattening of that slope afterwards with a potential small spike following the year of registration.

![Figure 6](image)
Figure 7. Number of triage calls per calendar year for each patient group (registered on the portal in 2011-2012, 2013-2014, 2015-2016, and 2017-2018). To reduce noise, the number of visits represents the running average of that year, the previous year, and the following year. The nurse triage calls were introduced in 2010 and show a consistent rise in frequency over time with a small increase in calls associated with the year of portal registration.

Discussion

Main Findings

Our findings suggest that portal registration is associated with an increase in service use, but that some reductions may be expected over subsequent years. Compared to matched controls, portal registration was associated with a significant initial increase in the number of in-person MD encounters and telephone calls, but a subsequent drop in the rate of MD visits and increase in NP visits over time. MDs spent an estimated 5.7 minutes per patient annually to respond to portal messages.

Limitations and Comparison With Prior Work

We believe that ours is the first study to examine the trend in encounters after portal registration over an extended time span and the first study to examine the impact of an exclusively primary care portal on traditional health care usage. It is possible that the observed increase in encounters was due to differences between the two groups that were not captured in the propensity matching. For instance, patients might have registered on the portal when they developed a new health concern, anticipating an increased requirement for health services. The reason for the gradual decrease in MD visits but increase in NP visits that took place after the initial jump in MD visits associated with portal registration is difficult to determine without further study. It is possible that patients initially presented to their own physician after sending them a portal message or viewing results, but the physician shared follow-up care with the nurse practitioner or physician assistant.

There may have been differences in areas such as electronic literacy or internet access that were not identified. It is also possible that the higher frequency of in-person encounters after portal registration was due to increased engagement by patients in their health. For instance, access to laboratory results may have generated questions from patients [29]. Increased awareness of being due for cancer screening or diabetes or blood pressure monitoring may have resulted in a higher number of encounters but improved quality of care or patient satisfaction. We did not examine these areas as they were beyond the scope of this study, but they would benefit from future research. While some past studies demonstrated improvements in certain health outcomes associated with electronic patient portals [30-33], only a few were based in primary care [31,34]. Several systematic reviews that evaluated a variety of portals in different practice settings suggested that portals or similar digital health services may result in improved patient satisfaction, but they did not demonstrate a meaningful impact on health outcomes, cost, or use [35-40].

We found that providers spent less than 6 minutes per year on clerical work for each patient registered on the portal. This is a small amount of time per patient but is significant when considering the context of an entire primary care practice. We note that the time-stamping of messages was performed during the COVID-19 pandemic, while the number of messages sent by all providers was collected prior to 2020. We consider that even if the COVID-19 pandemic resulted in an increased number of messages, the provider time per message would not have changed significantly. Therefore, our estimate of portal-related clerical work reflects prepandemic time requirements, and these may have increased since 2020 due to increased patient interest in asynchronous virtual care. This would also be an area for further study. Portals that do not allow incoming messages or any secure messaging would reduce or eliminate this time requirement but might also limit patient engagement and other potential benefits of the portal. Since the clerical burden associated with electronic environments in health care has been associated with professional burnout, [22,26] it is important to consider the provider time requirement associated with patient portals. The time and cost associated with incorporating a patient portal are currently not specifically addressed in either fee-for-service or capitated Canadian primary care funding models.

There are other limitations to this study. We examined the long-term impact of an electronic patient portal in a single
primary care practice, which may not be reflective of the impact in other primary care practices. However, portal adoption has not been widespread for long enough to allow study of the long-term impact of portals across multiple sites. Additionally, the impact of patient portals in other settings, such as hospitals, laboratories, or specialist practices, may be quite different. Further research is needed into electronic patient portals in different settings to determine their impact on various health outcomes.

Conclusions

Electronic patient portals are increasingly being adopted by providers and sought after by patients. We found that portal registration was associated with a subsequent increase in the number of some traditional encounters and a small increase in clerical workload for providers. Portals have enormous potential to truly engage patients as partners in their own health care, but their impact on use of traditional health care services and clerical burden must also be considered when they are incorporated into primary care.

Acknowledgments

This study was funded by a CAD $20,000 (US $15,000) research grant from the University of Ottawa Department of Family Medicine Program for Research and Innovation in Primary Care and Medical Education.

Conflicts of Interest

None declared.

References


Abbreviations
- ITS: interrupted time-series
- MD: medical doctor
- NP: nurse practitioner
- PA: physician assistant
- TC: triage call

Please cite as: Ferguson K, Fraser M, Tuna M, Bruntz C, Dahrouge S
The Impact of an Electronic Portal on Patient Encounters in Primary Care: Interrupted Time-Series Analysis
JMIR Med Inform 2023;11:e43567
URL: https://medinform.jmir.org/2023/1/e43567
doi:10.2196/43567
PMID:36745495

©Karen Ferguson, Mark Fraser, Meltem Tuna, Charles Bruntz, Simone Dahrouge. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 06.02.2023. 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.
An Electronic Dashboard to Improve Dosing of Hydroxychloroquine Within the Veterans Health Care System: Time Series Analysis

Anna Montgomery¹, MPH; Gary Tarasovsky¹,², BSc; Zara Izadi³, PharmD, PhD; Stephen Shiboski², PhD; Mary A Whooley¹,²,³, MSc, MD; Jo Dana¹, NP; Iziegbe Ehiorobo², MD; Jennifer Barton⁴, MD; Lori Bennett⁵, PharmD; Lorinda Chung⁶,⁷, MD; Kimberly Reiter⁸,⁹, MD; Elizabeth Wahl¹⁰, MD; Meera Subash¹¹, MD; Gabriela Schmajuk¹,²,³, MSc, MD

¹San Francisco VA Medical Center, San Francisco, CA, United States
²University of California San Francisco, San Francisco, CA, United States
³UCSF Philip R Lee Institute for Health Policy Studies, San Francisco, CA, United States
⁴VA Portland Health Care System, Portland, OR, United States
⁵Ralph H Johnson VA Medical Center, Charleston, SC, United States
⁶Palo Alto VA Medical Center, Palo Alto, CA, United States
⁷Stanford University, Palo Alto, CA, United States
⁸Raymond G Murphy VA Medical Center, Albuquerque, AZ, United States
⁹University of New Mexico School of Medicine, Albuquerque, AZ, United States
¹⁰Seattle/Puget Sound VA Healthcare System, Seattle, WA, United States
¹¹UT Physicians Center for Autoimmunity, Houston, TX, United States

Corresponding Author:
Gabriela Schmajuk, MSc, MD
University of California San Francisco
4150 Clement St
San Francisco, CA, 94121
United States
Phone: 1 415 221 4810
Email: Gabriela.schmajuk@ucsf.edu

Abstract

Background: Hydroxychloroquine (HCQ) is commonly used for patients with autoimmune conditions. Long-term use of HCQ can cause retinal toxicity, but this risk can be reduced if high doses are avoided.

Objective: We developed and piloted an electronic health record–based dashboard to improve the safe prescribing of HCQ within the Veterans Health Administration (VHA). We observed pilot facilities over a 1-year period to determine whether they were able to improve the proportion of patients receiving inappropriate doses of HCQ.

Methods: Patients receiving HCQ were identified from the VHA corporate data warehouse. Using PowerBI (Microsoft Corp), we constructed a dashboard to display patient identifiers and the most recent HCQ dose and weight (flagged if ≥5.2 mg/kg/day). Six VHA pilot facilities were enlisted to test the dashboard and invited to participate in monthly webinars. We performed an interrupted time series analysis using synthetic controls to assess changes in the proportion of patients receiving HCQ ≥5.2 mg/kg/day between October 2020 and November 2021.

Results: At the start of the study period, we identified 18,525 total users of HCQ nationwide at 128 facilities in the VHA, including 1365 patients at the 6 pilot facilities. Nationwide, at baseline, 19.8% (3671/18,525) of patients were receiving high doses of HCQ. We observed significant improvements in the proportion of HCQ prescribed at doses ≥5.2 mg/kg/day among pilot facilities after the dashboard was deployed (–0.06; 95% CI –0.08 to –0.04). The difference in the postintervention linear trend for pilot versus synthetic controls was also significant (–0.06; 95% CI –0.08 to –0.05).

Conclusions: The use of an electronic health record–based dashboard reduced the proportion of patients receiving higher than recommended doses of HCQ and significantly improved performance at 6 VHA facilities. National roll-out of the dashboard will enable further improvements in the safe prescribing of HCQ.
Introduction

Hydroxychloroquine (HCQ) is among the most commonly used medications for patients with autoimmune conditions and received special attention in 2020 as a potential treatment for COVID-19, resulting in drug shortages for chronic users [1]. These drug shortages, combined with recent guidelines emphasizing toxicities associated with long-term use, highlighted the issue of prescribing HCQ in appropriate doses. Long-term use of HCQ, especially at higher doses, can cause severe retinal toxicity in some patients. The risk of this toxicity is reduced if the average daily dose of HCQ is $\leq 5$ mg/kg/day [2,3]. However, recent studies have revealed that 30%-40% of patients prescribed HCQ receive doses $>5$ mg/kg/day [4,5].

Previous studies have shown that enterprise-wide national dashboards are capable of improving care, but they have not been developed quickly enough, or disseminated widely enough, to make meaningful, population-level impacts on process or outcome measures [6-10]. Local, electronic health record (EHR)–based medication safety dashboards have been used to support medication safety but have not been scaled to date [11-14].

In this study, we sought to develop and deploy a national EHR-based medication safety dashboard within the Veterans Health Administration (VHA) to reduce inappropriate HCQ dosing. The VHA is the largest integrated health care delivery system in the United States, serving over 9 million veterans nationwide. Six VHA pilot facilities were enlisted to test the dashboard and invited to participate in monthly webinars. We followed pilot facilities over a 1-year period to determine whether they were able to improve the proportion of patients receiving inappropriate doses of HCQ.

Methods

Dashboard Development

The dashboard was developed as part of an ongoing project to improve the safe prescribing of high-risk disease-modifying antirheumatic drugs among VHA patients by the San Francisco VA’s Measurement Science Quality Enhancement Research Initiative. It was created using PowerBI (Microsoft Corp), a data management software package available within the VHA for approved users with secure access to EHR data. The VHA’s corporate data warehouse (CDW), which contains national VHA EHR data, served as the data source for the dashboard (Multimedia Appendix 1). PowerBI allows developers to extract, analyze, and display data from a variety of sources and features interactive tables and graphs that can be filtered or expanded using a graphical user interface [15]. Notably, PowerBI dashboards are “read-only,” that is, users can see and filter data elements, but in order to change data (eg, update the dose of HCQ), they must do so within the EHR.

Facilities, Patients, and Data Elements

All 130 VHA facilities were eligible to be included in the study. We excluded 1 facility that had transitioned to the Cerner EHR and did not have patient data available in VistA, and 1 facility with fewer than 10 patients on HCQ, leaving 128 facilities for the analysis. Patients from included facilities in the VHA with a current, active prescription for HCQ were included in the data captured by the dashboard. Patients were excluded if the patient was deceased, or if the HCQ prescription indicated that it was a placebo or study drug. We extracted values for each patient’s most recently prescribed HCQ dose (in mg per day), derived from the “quantity dispensed” and “days-supply” fields in the medication order. We also extracted the most recently captured body weight (in kg) to calculate the HCQ dose in mg/kg/day. These data were then linked to Microsoft PowerBI Gateway servers, which are automatically updated every 24 hours to reflect new information from CDW (Multimedia Appendix 1).

Dashboard Features

Figure 1 illustrates the dashboard using fictitious patient data. The dashboard displayed patient identifiers (first and last name, last 4 digits of their social security number, and VHA facility), HCQ in mg/kg/day based on weight and the number of pills prescribed per day. Rows were marked with a red x-mark if the HCQ daily dose was calculated to be $\geq 5.2$ mg/kg/day (vs a green check mark if $<5.2$ mg/kg/day) in the column immediately to the right of the dose. Patients without a recorded weight within the past 3 years were flagged with a yellow circle, indicating missing data. Rows could be filtered by facility location, provider, OCT exam date, or by HCQ dose. The dashboard also displayed national-, facility-, and prescriber-level performance (proportion of patients with HCQ doses $\geq 5.2$ mg/kg/day out of the total number of patients receiving HCQ) shown as pie charts for benchmarking. A user guide and video tutorial for the dashboard were available via a web-based link on the dashboard landing page. User interactions (number of times the dashboard is accessed per authorized user) were tracked using the PowerBI Activity Log feature [16].

https://medinform.jmir.org/2023/1/e44455
Study Period

Initial queries using CDW and PowerBI began in June 2020. The beginning of the study period—when baseline data collection on HCQ dosing across all VHA facilities started—began on August 11, 2020, prior to sharing the dashboard with any pilot testing facilities (see Multimedia Appendix 2). Of note, the final dashboard was developed over a period of under 5 months (June 2020 to October 2020).

Pilot Testing Facilities

We enlisted rheumatology providers, pharmacists, and dermatologists from 6 VHA pilot facilities to test the dashboard between October 26, 2020, and December 6, 2021. Pilot facilities were selected based on their willingness to participate in a related study involving screening for infections prior to immunosuppression. The 6 VHA pilot facilities included Ralph H Johnson VA Medical Center, Charleston, SC; Palo Alto VA Health Care System, Palo Alto, CA; VA Portland Health Care System, Portland, OR; Raymond G Murphy VA Medical Center, Albuquerque, NM; San Francisco VA Medical Center, San Francisco, CA; and Puget Sound/Seattle VA Healthcare System, Puget Sound, WA.

Pilot facilities were invited to use the dashboard via email. Once they agreed, they were granted secure access along with any additional staff at that facility. All facilities were trained in the use of the dashboard via a web-based webinar. Site personnel were invited to participate in web-based meetings of a Rheumatology Quality and Safety Workgroup to share feedback, address any barriers, and update information on the use of the dashboard, every other month. Each facility leader was also sent a quarterly facility-specific report via email with run charts depicting the proportion of patients on HCQ at doses ≥5.2 mg/kg/day and the number of times their facility accessed the dashboard during that quarter.

Each pilot facility was encouraged to develop an individualized workflow for use of the dashboard. For example, some facilities would check the dashboard weekly or monthly, while others used the downloadable report feature to distribute flagged patients to individual providers or trainees. All facility workflows included review of the dashboard, review of EHR charts of flagged patients, and HCQ dose adjustment if appropriate.

Control Facilities

Facilities in the control group did not have access to the dashboard and were not contacted as part of this study. Data on patients receiving HCQ were collected from CDW using the same process as was used for pilot facilities.

Complex Medication Instructions and Policy Change

On the dashboard, HCQ dose in mg/kg/day was calculated based on the number of pills prescribed and the patient’s most recent weight. However, occasionally, HCQ orders had complex instructions (eg, “Take 2 pills Monday through Friday, and 1 pill on Saturdays and Sundays”), which resulted in miscalculations of the daily dose based on these fields. Two authors (AM and GS) reviewed 939 randomly selected charts and found 3% (28/939) of HCQ orders contained complex instructions. In order to reduce the chances of misclassifying patients as having an inappropriate HCQ dose due to complex instructions or fluctuating patient weights, on November 30, 2020, we made a policy change to designate doses of ≥5.2 mg/kg/day as “high dose” (as opposed to doses ≥5.0 mg/kg/day).

Covariates and Descriptive Variables

We assessed facility characteristic variables that might be important in relation to medication safety practices in general and HCQ dosing specifically: facility region (Midwest, North Atlantic, Continental, Southeast, and Pacific); facility
complexity (high, medium, and low); and the total number of patients prescribed HCQ at the facility [17].

In addition, we reported facility-level HCQ patient characteristics including the proportion of patients who were ≥55 years; self-identified non-Hispanic White, self-identified Hispanic or Latinx; with at least 1 VHA rheumatology clinic visit within 12 months of the beginning of the study period; and with a rural residence. Facility-level HCQ patient clinical factors included the proportion of patients with rheumatic diseases (rheumatoid arthritis, systemic lupus erythematosus [SLE], or other); with OCT exam documented; and the proportion with inappropriate HCQ dosing at baseline (August 11, 2020). A patient was considered to have a diagnosis if they had at least 2 codes (at least 30 days apart) for a specific condition listed here: rheumatoid arthritis, SLE, polymyalgia rheumatica, discoid lupus, nongout crystal arthropathy, undifferentiated connective tissue disease, sarcoidosis, antiphospholipid antibody syndrome, mixed connective tissue disease, systemic sclerosis, osteoarthritis, inflammatory myopathies (including polymyositis and dermatomyositis), psoriatic arthritis, ankylosing spondylitis, antineutrophil cytoplasmic antibody-associated vasculitis, other vasculitis (including Kawasaki disease), dermatitis, or giant cell arteritis.

**Statistical Analysis**

Descriptive statistics were used to summarize facility characteristics and facility-level patient characteristics. We used interrupted time series (ITS) analysis to assess the effects of the dashboard on observed changes in the proportion of patients with HCQ doses ≥5.2 mg/kg/day. ITS is a strong quasiexperimental study design that can be used for single- and multiple group comparisons. In an ITS analysis, the outcome variable of interest (eg, the average proportion of patients with HCQ doses ≥5.2 mg/kg/day) is observed over multiple time periods before and after an intervention that is expected to “interrupt” the trend over time. ITS has been previously found useful when evaluating health care interventions for its ability to evaluate the causal impact of policy changes and health care interventions without random assignment [18,19]. We used the itsa command, which is available in the official Stata packages newey and prais [19].

Due to large variability in key facility characteristics observed at baseline between pilot and control facilities (proportion of patients on HCQ at doses ≥5.2 mg/kg/day, facility complexity, and mean number of patients prescribed HCQ at the facility; ITS regression output are displayed in Multimedia Appendix 3), we opted to implement a robust matching method using synthetic controls to measure the impact of the dashboard on HCQ dosing at the pilot facilities. Using this approach, pilot facility performance was compared to matched synthetic controls using the synth package in Stata [20]. Synthetic controls were constructed from a weighted combination of control units not exposed to the dashboard but with preintervention outcome dynamics and covariate levels similar to the pilot facilities prior to any interventions [21]. Matching was based on observed changes in the proportion of patients with HCQ doses ≥5.2 mg/kg/day, facility complexity, and the mean number of patients prescribed HCQ at the facility. To assess the balance of the pilot facilities and their synthetic controls, we used the absolute standardized mean difference (ASMD). As a rule of thumb, ASMD < 0.10 is an indicator of a good balance between synthetic control unit and a treated unit [22].

As part of the multiple group ITS analysis, pilot facilities were compared to synthetic controls in weekly increments of the proportion of patients with HCQ doses ≥5.2 mg/kg/day. We estimated the coefficients using segmented ordinary least square (OLS) linear regression models in which the errors were assumed to follow a first-order autoregressive process [19]. The model was specified to base the pooled autocorrelation estimate on the autocorrelation of the residuals. We expressed the effect of the dashboard on the proportion of patients with HCQ doses ≥5.2 mg/kg/day as intercept and slope changes. The intervention date was set as October 26, 2020 (the date the pilot facilities were granted access to the dashboard). We incorporated the policy shift (shift from recording the proportion of patients receiving ≥5.0 mg/kg/day to those receiving ≥5.2 mg/kg/day on November 30, 2020) using established methods [18].

Statistical analyses were performed using Stata 15 (StataCorp LLC). A P value <.05 was used as the criterion for statistical significance.

**Secondary Analyses**

As secondary analyses, we compared the 6 pilot facilities to other facilities using modified Xbar-R charts. We used Microsoft QI Macros, a statistical process control software package plugin for Microsoft Excel, to generate modified Xbar-R charts to analyze the overall trends and stability in the proportion of patients with HCQ doses ≥5.2 mg/kg/day over time. Upper and lower control limits varied based on the average proportion of patients with HCQ doses ≥5.2 mg/kg/day. A continuous change of 6 or more points in a row or 8 or more points on the same side of the centerline is considered a significant trend [23].

We performed 2 separate comparisons: (1) pilot facilities versus all other facilities nationally and (2) pilot facilities versus matched control facilities. Matched control facilities were selected based on (1) the slope of proportion of patients prescribed HCQ at doses ≥5.2 mg/kg/day during the baseline period (August 11, 2020, to October 25, 2020); (2) the total number of patients prescribed HCQ; and (3) high facility complexity. Since pilot facilities had a mean) of 228 (SD 69) patients prescribed HCQ, we required matched control facilities to have at least 75 patients prescribed HCQ. Application of these criteria resulted in 8 matched control facilities, which were all included in the matched control sensitivity analysis.

**Feedback From Pilot Facilities**

At the end of the study period, clinicians at pilot facilities were sent a confidential survey to solicit quantitative and qualitative feedback about the dashboard. The 14-item survey included questions about the capacity in which sites used the dashboard, usability of the dashboard, suggestions for improvement, and the likelihood of recommending the dashboard to a colleague or trainee.
Ethics Approval
All VHA authors of this manuscript attest that the activities that resulted in producing this manuscript were conducted as part of a nonresearch evaluation under the authority of the National Rheumatology Field Advisory Committee and Center for Medication Safety. This work was approved by the VA Quality Enhancement Research Initiative (QUERI; IRB 15-18358).

Results

Pilot Facilities and Workflows
We identified 18,525 total users of HCQ nationwide in the VHA, including 1365 patients at the 6 pilot facilities. Across the 6 pilot facilities, 36 providers were granted access to the dashboard including 14 rheumatologists, 12 physician residents, 3 rheumatology fellows, 2 nurse practitioners specializing in rheumatology, 2 clinical pharmacists, 1 dermatologist, 1 registered nurse coordinator, and 1 primary care physician. Different pilot facilities developed different workflows around dashboard use to suit their needs. Some facilities requested access for all their clinicians (attendings and trainees) and had each one review their own patients. Others had a designated reviewer who checked the dashboard once a month or once a quarter. Another was able to download a spreadsheet containing the dashboard data and distribute it securely for clinician review (an example of typical dashboard clinic workflow for users is available in Multimedia Appendix 4). All pilot facilities had at least 20 interactions with the dashboard starting in October 2020; the median weekly number of dashboard interactions over the course of the study period was 8 (IQR 4-15).

Baseline Facility Characteristics
Table 1 shows the characteristics of pilot facilities at all facilities nationally. Nationwide, at the start of the study period, 19.8% (3671/18,525; range 4.26% to 44%) of patients prescribed HCQ were receiving HCQ ≥5 mg/kg/day versus 16.1% (220/1365) among pilot facilities.
Table 1. Facility characteristics and practice-level patient characteristics for the pilot versus all facilities at baseline (November 8, 2020).

<table>
<thead>
<tr>
<th>Facility characteristics</th>
<th>Pilot facilities (n=6)</th>
<th>All facilities (N=128)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity\textsuperscript{a}, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High complexity</td>
<td>6 (100)</td>
<td>84 (66)</td>
</tr>
<tr>
<td>Medium complexity</td>
<td>0 (0)</td>
<td>18 (14)</td>
</tr>
<tr>
<td>Low complexity</td>
<td>0 (0)</td>
<td>26 (20)</td>
</tr>
<tr>
<td>Geographic location, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continental</td>
<td>0 (0)</td>
<td>24 (19)</td>
</tr>
<tr>
<td>Midwest</td>
<td>0 (0)</td>
<td>26 (20)</td>
</tr>
<tr>
<td>North Atlantic</td>
<td>0 (0)</td>
<td>36 (28)</td>
</tr>
<tr>
<td>Pacific</td>
<td>5 (83)</td>
<td>22 (17)</td>
</tr>
<tr>
<td>Southeast</td>
<td>1 (17)</td>
<td>20 (16)</td>
</tr>
<tr>
<td>Total patients prescribed HCQ\textsuperscript{b}, mean (SD)</td>
<td>228 (69)</td>
<td>146 (107)</td>
</tr>
</tbody>
</table>

**Facility-level HCQ patient characteristics, mean (SD)**

<table>
<thead>
<tr>
<th>Facility-level HCQ patient characteristics, mean (SD)</th>
<th>Pilot facilities (n=6)</th>
<th>All facilities (N=128)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion of male patients</td>
<td>0.72 (0.05)</td>
<td>0.71 (0.10)</td>
</tr>
<tr>
<td>Proportion of patients aged &gt;55 years</td>
<td>0.79 (0.10)</td>
<td>0.76 (0.09)</td>
</tr>
<tr>
<td>Proportion of non-White patients</td>
<td>0.33 (0.10)</td>
<td>0.32 (0.18)</td>
</tr>
<tr>
<td>Proportion of Hispanic/Latinx patients</td>
<td>0.05 (0.03)</td>
<td>0.05 (0.04)</td>
</tr>
<tr>
<td>Proportion of patients who visited a VA rheumatology clinic within 1 year of baseline</td>
<td>0.44 (0.08)</td>
<td>0.59 (0.19)</td>
</tr>
<tr>
<td>Proportion of patients with a rural residence</td>
<td>0.33 (0.02)</td>
<td>0.32 (0.21)</td>
</tr>
</tbody>
</table>

**Facility-level HCQ patient clinical factors, mean (SD)**

<table>
<thead>
<tr>
<th>Facility-level HCQ patient clinical factors, mean (SD)</th>
<th>Pilot facilities (n=6)</th>
<th>All facilities (N=128)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion of patients with rheumatoid arthritis\textsuperscript{c}</td>
<td>0.43 (0.15)</td>
<td>0.45 (0.10)</td>
</tr>
<tr>
<td>Proportion of patients with systemic lupus erythematosus\textsuperscript{c}</td>
<td>0.14 (0.02)</td>
<td>0.16 (0.04)</td>
</tr>
<tr>
<td>Proportion of patients with other rheumatic disease\textsuperscript{c}</td>
<td>0.20 (0.12)</td>
<td>0.20 (0.12)</td>
</tr>
<tr>
<td>Proportion of patients with HCQ dose ≥5 mg/kg/day at baseline (August 11, 2020)</td>
<td>0.16 (0.03)</td>
<td>0.20 (0.07)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Station complexity: high complexity facilities have large levels of patient volume, patient risk, teaching and research, and contain level 4 to 5 intensive care units; medium complexity facilities have medium levels of patient volume, medium patient risk, some teaching and/or research, and contain level 3 and 4 intensive care units; low complexity facilities have the smallest level of patient volume, little or no teaching/research, the lowest number of physician specialists per patient, and contain level 1 and 2 intensive care units.

\textsuperscript{b}HCQ: hydroxychloroquine.

\textsuperscript{c}Rheumatic diseases were identified as veterans with 2 or more ICD-10 codes within the same disease category, separated by 30 or more days. Other autoimmune rheumatic diseases included: polymyalgia rheumatica, discoid lupus erythematosus, nongout crystal arthropathy, undifferentiated connective tissue disease, sarcoidosis, antiphospholipid syndrome, mixed connective tissue disease, systemic sclerosis, osteoarthritis, inflammatory myopathies (including polymyositis and dermatomyositis), psoriatic arthritis, ankylosing spondylitis, antineutrophil cytoplasmic antibody-associated vasculitis, and other vasculitides (including Kawasaki disease), lymphocytic infiltrates of the skin, or giant cell arteritis.

**ITS Analysis With Synthetic Controls**

Pilot facilities and synthetic controls were well matched in their predictor balance (ASMD=0.05). The postintervention linear trend showed pilot facilities’ proportion of patients with HCQ doses ≥5.2 mg/kg/day changed by −0.06 (95% CI −0.08 to −0.04) after the policy change, while the synthetic controls remained stable (0.006; 95% CI −0.00 to 0.01), with a statistically significant difference between the 2 groups by the end of the study period (−0.06; 95% CI −0.08 to −0.05; Multimedia Appendix 5 and Figure 2).
Secondary Analyses
As seen in Multimedia Appendix 6, the modified Xbar-R control chart showed meaningful improvements in the proportion of patients receiving HCQ doses $\geq 5.2$ mg/kg among pilot facilities over the course of the study period. There was a downward trend of 21 points outside of the upper and lower control limits, indicating a significant overall average process change. In contrast, the 8 matched control facilities’ proportion remained stable (ie, within the control limits). A comparison of pilot facilities to all other facilities nationally revealed similar results (Multimedia Appendix 7).

Feedback From Pilot Sites
Six clinicians, 1 from each pilot facility, responded to the web-based survey. Of these 6 clinicians, 5 reported that the dashboard was extremely easy to use, 5 answered they were extremely likely to use the dashboard in the future, and 5 responded they were extremely likely to recommend the dashboard to a colleague or trainee.

Discussion
In an era where the complexity of care and the number of evidence-based practices are ever expanding, the cognitive load required to address these practices during a short office visit can be overwhelming for clinicians. EHR-based dashboards are 1 method to support clinicians in evidence-based care of their patients. In this study, we developed an EHR-based medication safety dashboard to improve the safe prescribing of HCQ within the VHA. As part of a multipronged intervention, we found that audit and feedback via the dashboard resulted in a clinically meaningful and statistically significant reduction in the proportion of patients receiving high doses of HCQ among pilot facilities. Based on our linear postintervention trends, on the assumption that all facilities will behave similarly to the pilot sites, it would take approximately 4 years to reduce the proportion of patients receiving high HCQ doses from 16% to less than 5%.

Several features of the infrastructure available through the VHA made this a successful pilot. First, enterprise-wide PowerBI software was easily accessible as a pre-existing software suite available within the VHA for internal users. A new workspace was requested and granted within 48 hours; no new software installation was required. Second, it was straightforward to query VHA CDW data and then link these data to PowerBI servers. Construction of a first prototype of the dashboard took only a few months, and the final version (after several iterations) was available in 5 months. Beyond the VHA infrastructure, this pilot was feasible because of its limited scope to a single medication—HCQ comes in a single pill size, and most prescriptions have the number of pills dispensed corresponding to the daily dose, which facilitated calculating dose in mg/kg/day. Finally, because of its intuitive user interface, training required to use the dashboard by pilot sites was minimal.

There are few descriptions of EHR-integrated medication safety dashboards in the literature, and those that have been reported...
have also been successful [24]. For example, with the use of the UK SMASH dashboard, the prevalence of potentially unsafe prescribing of nonsteroidal anti-inflammatory drugs and other medications was reduced by 41% at intervention facilities [10,11]. Another, US-based, local, pharmacist-led medication safety program, which included a dashboard and educational outreach, reduced errors by 27%-49% after 6 and 12 months of use [25]. Several US patient registries have also developed clinician-facing dashboards to improve quality and medication safety and demonstrated significant improvements over time [26].

Although our pilot project was successful, we do note some limitations for the dashboard. First, although the development and validation of data in the dashboard were smooth, there were a small fraction of patients whose calculated doses remained inaccurate due to complex instructions that did not match the number of pills supplied. We attempted to mitigate these inaccuracies by only flagging doses ≥5.2 mg/kg/day instead of 5 mg/kg/day. We made this choice to avoid falsely labeling HCQ orders as high at the expense of missing some patients receiving doses above those recommended in the latest guidelines. Further work is needed to explore whether this tradeoff is worthwhile, especially since many clinicians use complex dosing in order to avoid average daily doses of ≥5 mg/kg/day, so the use of complex instructions may be correlated with appropriate dosing. In the future, a potential solution could be to develop an algorithm that captures information from the complex instructions using natural language processing techniques. Second, many pilot users requested additional features that are not available through PowerBI. Most importantly, users wished to be able to annotate dashboard tables directly or unflag patients who might be receiving higher than recommended doses of HCQ deliberately due to severe disease. Unfortunately, these features were not available in the VHA implementation of PowerBI at the time of this study.

One critical question for the future is whether the improvements observed in this pilot study will be sustainable. Clinician buy-in and ongoing utilization are crucial to the effectiveness of this dashboard as a sustainable audit-and-feedback tool [25]. Several of our pilot facilities started using this dashboard as a component of their routine quality improvement activities and reported dashboard use as part of a pay-for-performance program. Other facilities incorporated its use into trainee quality improvement activities. These additional use cases make sustainability more likely.

Another important question for future studies is about the clinical effects of reducing HCQ doses for some patients. Some recent observational studies have suggested that patients with SLE who decrease their HCQ dose may be at increased risk for disease flares [27,28]. It seems unlikely that small changes in dosing would have a large effect, but nevertheless, this is an important question to investigate. Unfortunately, since this is a national study limited by using structured EHR data, it is impossible to ascertain the condition of any specific patient before or after introducing the dashboard.

Moving forward, we will test the effects of the dashboard in a national roll-out across all VHA facilities. Additional mixed methods research will aid our understanding of provider adoption and sustained use of the dashboard and whether other interventions are needed to support safe prescribing of HCQ (eg, clinical decision support for weight-based dosing, or other pharmacy-based alerts or workflows) [29]. We also plan to roll-out additional dashboards focused on other important rheumatology safety issues, including pretreatment screening for latent infections in patients receiving biologic and targeted small molecule medications and HLA B:5801 testing for eligible patients receiving allopurinol. Our hope is that with a suite of dashboards and associated toolkits, quality improvement activities will be more feasible for all clinicians.

In summary, we successfully developed and deployed an EHR-based medication safety dashboard to improve the safe prescribing of HCQ within the VHA. The use of the dashboard significantly reduced the proportion of patients receiving higher than recommended doses of HCQ at 6 VHA facilities. National roll-out of the dashboard will enable further improvements in the safe prescribing of HCQ.

Acknowledgments
This work was funded by the VA Quality Enhancement Research Initiative (QUERI) grant QIS 19-459.

Data Availability
The data that support the findings of this study are available from the Department of Veterans Affairs, but restrictions apply to the availability of these data. Data are, however, available from the authors upon reasonable request and with permission of the Department of Veterans Affairs.

Authors’ Contributions
AM made substantial contributions to the conception, design, data analysis and interpretation, and the creation of the dashboards that were used in this work. GT contributed to the study design, the creation of the dashboard, and assisted with the interpretation of the data used in this work. SS and ZI made substantial contributions to the data analysis and interpretation of the data along with assistance with the drafting of the manuscript. MAW, JD, IE, JB, LB, LC, KR, EW, and MS made substantial contributions to the data acquisition and the revisions to the manuscript. GS made substantial contributions to the conception, study design, data interpretation, and drafting of the manuscript for submission. All authors read and approved the final manuscript.
Conflicts of Interest

ZI is an employee at BMS.

Multimedia Appendix 1
Architecture diagram of the hydroxychloroquine patient safety dashboard. The electronic health record data were pulled from the corporate data warehouse SQL Servers. The data were then directly linked from the SQL Servers to the PowerBI gateway and presented to the end user via a web interface.

Multimedia Appendix 2
Study timeline.

Multimedia Appendix 3

Multimedia Appendix 4
Hydroxychloroquine patient safety dashboard pilot facility Swimlane diagram.

Multimedia Appendix 5
Comparison of a linear postintervention trend for pilot versus synthetic control pilot facilities after the policy change (November 30, 2020).

Multimedia Appendix 6
Modified Xbar control chart showing mean percent of patients receiving inappropriate HCQ doses among pilot facilities and matched control facilities. The x-axis shows 2-week time segments during the study period from August 11, 2020, to December 6, 2021; the y-axis shows the percent of patients with higher than recommended HCQ doses. The vertical dotted lines denote the dates when the pilot facilities were granted access to the dashboard and when the “High Dose HCQ” definition changed from ≥5.0 mg/kg/day to ≥5.2 mg/kg/day. The orange and blue dotted lines show the average and upper/lower control limits for the 6 pilot facilities compared to 8 matched control facilities, respectively. The dots on the solid lines represent performance at each time point. CL: central line; LCL: lower control limit; HCQ: hydroxychloroquine; UCL: upper control limit.

Multimedia Appendix 7
Modified Xbar control chart showing mean percent of patients receiving inappropriate HCQ doses among pilot facilities and all other facilities nationally. The x-axis shows 2-week time segments during the study period from August 11, 2020, to December 6, 2021; the y-axis shows the percent of patients with higher than recommended HCQ doses. The vertical dotted lines denote the dates when the pilot facilities were granted access to the dashboard and when the “High Dose HCQ” definition changed from ≥5.0 mg/kg/day to ≥5.2 mg/kg/day. The orange and blue dotted lines show the average and upper/lower control limits for the 6 pilot facilities compared to all other 122 facilities, respectively. The dots on the solid lines represent performance at each time point. CL: central line; LCL: lower control limit; HCQ: hydroxychloroquine; UCL: upper control limit.

References


17. Site facility name and complexity. United States Department of Veterans Affairs. URL: https://tinyurl.com/4b4s68zy [accessed 2022-01-20]


Abbreviations

CDW: corporate data warehouse
EHR: electronic health record
HCQ: hydroxychloroquine
ITS: interrupted time series
OCT: optical coherence tomography
OLS: ordinary least square
SLE: systemic lupus erythematosus
VHA: Veterans Health Administration

©Anna Montgomery, Gary Tarasovsky, Zara Izadi, Stephen Shiboski, Mary A Whooley, Jo Dana, Iziegbhe Ehiorobo, Jennifer Barton, Lori Bennett, Lorinda Chung, Kimberly Reiter, Elizabeth Wahl, Meera Subash, Gabriela Schmajuk. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 12.05.2023. 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.
Integrated Personal Health Record in Indonesia: Design Science Research Study

Nabila Clydea Harahap¹, Dr; Putu Wuri Handayani¹, Dr; Achmad Nizar Hidayanto¹, Prof Dr
Faculty of Computer Science, University of Indonesia, Depok, Indonesia

Corresponding Author:
Nabila Clydea Harahap, Dr
Faculty of Computer Science
University of Indonesia
Kampus UI Depok, Pondok Cina, Beji
Depok, 16424
Indonesia
Phone: 62 8571652699
Email: nabila.clydea@ui.ac.id

Abstract

Background: Personal health records (PHRs) are consumer-centric tools designed to facilitate the tracking, management, and sharing of personal health information. PHR research has mainly been conducted in high-income countries rather than in low- and middle-income countries. Moreover, previous studies that proposed PHR design in low- and middle-income countries did not describe integration with other systems, or there was no stakeholder involvement in exploring PHR requirements.

Objective: This study developed an integrated PHR architecture and prototype in Indonesia using design science research. We conducted the research in Indonesia, a low- to middle-income country with the largest population in Southeast Asia and a tiered health system.

Methods: This study followed the design science research guidelines. The requirements were identified through interviews with 37 respondents from health organizations and a questionnaire with 1012 patients. Afterward, the proposed architecture and prototype were evaluated via interviews with 6 IT or eHealth experts.

Results: The architecture design refers to The Open Group Architecture Framework version 9.2 and comprises 5 components: architecture vision, business architecture, application architecture, data architecture, and technology architecture. We developed a high-fidelity prototype for patients and physicians. In the evaluation, improvements were made to add the stakeholders and the required functionality to the PHR and add the necessary information to the functions that were developed in the prototype.

Conclusions: We used design science to illustrate PHR integration in Indonesia, which involves related stakeholders in requirement gathering and evaluation. We developed architecture and application prototypes based on health systems in Indonesia, which comprise routine health services, including disease treatment and health examinations, as well as promotive and preventive health efforts.

(JMIR Med Inform 2023;11:e44784) doi:10.2196/44784

KEYWORDS
personal health record; integrated; Indonesia; design science; mobile phone

Introduction

Background

Current trends in health informatics encourage the transition from institution-centric to patient-centric health care [1]. The use of IT is not only for patients in health care settings but also for all individuals who want to maintain health and are involved in disease prevention and health promotion [1]. Personal health records (PHRs) are consumer-centric tools designed to facilitate the tracking, management, and sharing of personal health information [1]. PHRs contain medical data and information about a patient that are managed by the patients themselves [2]. PHRs form a trend from information controlled by the health system to information controlled by individuals [3].

In its simplest form, a PHR is a stand-alone application (stand-alone PHR) and is not connected to other systems [4]. In a more complex form, the health information provided by the PHR is linked to the electronic health record or electronic health record.
medical record (tethered PHR) [4]. Furthermore, PHRs can be connected to various health data sources to obtain and transmit data (integrated PHR) [4]. An integrated PHR is the most ideal form of PHR as implementing PHRs in this way has the potential to improve the quality, accessibility, and delivery of health services [3].

A previous review on the implementation of PHRs shows that PHR research has mainly been conducted in high-income countries rather than in low- and middle-income countries [5]. Few studies that have been conducted in low- and middle-income countries aim to propose PHR applications for certain purposes, such as pediatric vaccination [6], or specific diseases, such as metabolic syndrome management [7], chronic heart failure [8], and kidney transplant [9]. These studies focused on the usability of PHRs and did not describe integration with other systems or applications. A study by Abdulnabi et al [10] described PHR interoperability by designing a distributed PHR model. However, there was no stakeholder involvement in exploring PHR requirements.

Using design science, we complement gaps from previous studies by developing a PHR model that is integrated with various systems, and we involve relevant stakeholders to explore the requirements and evaluate the proposed PHR model. We conducted the research in Indonesia, a low- to middle-income country with the largest population in Southeast Asia [11,12]. In Indonesia, health services are delivered by the public and private sectors. In the public sector, health facilities comprise hospitals (general and specialty) and pusat kesehatan masyarakat or primary health centers (Puskesmas). In the private sector, health facilities comprise hospitals and primary care clinics. In addition, there is the Social Security Agency for Health or Badan Pelaksana Jaminan Sosial Kesehatan (BPJS Kesehatan) that administers the national health insurance program (Jaminan Kesehatan Nasional [JKN] or national health insurance). Patients with JKN must follow a tiered referral flow starting from primary care facilities as gatekeepers for JKN patients before being referred to hospitals. Without a referral letter, JKN patients are not allowed to go directly to a hospital or specialist clinic except in an emergency [13].

In addition to health efforts that focus on treating and curing diseases, there are also Healthy Family (Keluarga Sehat) and Community Healthy Life Movement (Germas) programs that are managed by the Puskesmas and focus on promotive and preventive health efforts [14,15]. Currently, health development policies in Indonesia are directed at improving access to and quality of health services, with an emphasis on increasing promotive and preventive health efforts supported by innovation and the use of technology [16]. Integrated PHRs can be an opportunity to improve access to and quality of health services in Indonesia by using IT [17].

**Objectives**

As a technological solution for integrated PHRs in Indonesia, this study developed an integrated PHR architecture and prototype in Indonesia using the design science research (DSR) approach by Hevner et al [18]. The DSR approach was chosen in this study as the goal of DSR is to focus on designing systems that not only are practical but can also contribute to knowledge. The question that will be answered in this research is as follows: How are the architectural designs and prototypes of integrated PHR applications in Indonesia? The design of the PHR model, which was developed using a DSR approach, can provide an overview for developing PHRs using scientific theory and methods. The results of this study are expected to be a guide for health facilities or health policy makers in integrating PHRs and health applications in Indonesia.

**Methods**

**Design Science**

This study was conducted using a DSR approach [18]. In DSR, the details or stages of design and development may vary even though the focus of the research is on artifact design [19]. Peffers et al [20] describe 6 activity-based methodologies for DSR, whereas Hevner et al [18] define 7 guidelines for DSR. This study follows the DSR guidelines defined by Hevner et al [18], which are grouped into 3 cycles or phases comprising the relevance cycle (identifying problems and the artifact type), design cycle (developing and evaluating the artifact), and rigor cycle (research contribution and communication; Figure 1 [5]). These guidelines have been followed by DSR studies in low- and middle-income countries to develop health information systems such as mobile health (mHealth) [19,21-27]. The guidelines by Hevner et al [18], as a necessary element in DSR, are also consistent with the DSR methodology by Peffers et al [20].

The artifacts developed in this study are architecture and application prototypes. Integrated PHR requirements were carried out based on a systematic literature review regarding functionalities and issues in the implementation of PHRs [5], literature studies on health regulations in Indonesia, interviews with health organizations, and questionnaire distribution to patients. The requirements mentioned by health organizations in the interviews were categorized and grouped into themes that were defined by Harahap et al [5] and combined with requirements from patients. The architectural design then became a reference for designing the application prototype.

This study used a purposive sampling method to select participants who had the required knowledge regarding PHR implementation. The health organizations from which participants were interviewed were health facilities (private and government hospitals, Puskesmas, and clinics), health regulators (Ministry of Health and BPJS Kesehatan), and health application vendors. Data were collected through semistructured interviews between August 19, 2020, and January 15, 2021, and between October 1, 2021, and October 11, 2021. Interviews were conducted on the web using Zoom Cloud Meetings (Zoom Video Communications) and audio recorded with the participants’ consent. Each interview lasted between 30 and 60 minutes. The interview questions are attached in Multimedia Appendix 1.
The questionnaire was distributed to patients who met the criteria of respondents being Indonesian citizens aged ≥17 years. Before the questionnaire was distributed, we conducted a readability test to ensure that all the items on the questionnaire could be understood by the respondents in terms of writing and sentence meaning. The readability test was conducted on 6 respondents from September 25, 2021, to September 28, 2021. After the readability test, we made a revision based on the input given by the respondents. Questionnaires were distributed on the web through messaging applications such as WhatsApp (Meta Platforms), Telegram (Telegram FZ LLC), and Line (Line Corporation), as well as social media such as Facebook and Twitter, from October 28, 2021, to November 20, 2021.

Questionnaire data were analyzed using descriptive statistics in Microsoft Excel (Microsoft Corp) and SPSS Statistics (version 28; IBM Corp).

To evaluate the artifacts, this study used the evaluation guidelines defined by Venable et al [28] and the evaluation criteria defined by Hevner et al [18]. The goal of the evaluation was to determine the suitability of the design for the needs of health services in Indonesia. The architectural design and application prototype were evaluated qualitatively through interviews with IT or eHealth experts. The evaluation of the architecture design aimed to assess the completeness and conformity with the health system in Indonesia. The evaluation of the application aimed to assess functionality and usability.

This study used the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines as a comprehensive checklist that covers the necessary components of qualitative research (Multimedia Appendix 2 [29]). Interview data were analyzed using content analysis techniques in NVivo (version 12; QSR International). The content analysis steps comprised decontextualization, recontextualization, categorization, and compilation [30]. In decontextualization, the authors read the transcribed text and broke down the text into smaller meaning units. Each identified meaning unit was labeled with a code. In recontextualization, the original text was reread alongside the final list of meaning units. In the categorization process, themes and categories were identified. The categorization for the requirement analysis was based on the functionalities and issues in the implementation of PHRs defined by Harahap et al [5]. The categorization for the architectural evaluation interview was carried out based on the architectural components, whereas the categorization for the prototype evaluation interview was carried out based on the implemented functions in the PHR application. At the compilation stage, the authors wrote the results of the analysis.

**Ethics Approval**

The authors obtained a letter of ethics approval from the Faculty of Computer Science, University of Indonesia, to conduct data collection with letter S-1122A/UN2.F11.D1/PDP.01/2020. The author submitted the letter to the respondents and provided a brief explanation of the study objective. Each respondent verbally provided consent to participate during the interview.

**Results**

### Artifact Development

#### Respondent Demographics

We interviewed a total of 37 respondents. The respondents were from 10 first-level health facilities (n=6, 60% Puskesmas and n=4, 40% clinics) and 15 referral-level health facilities (n=9, 60% government hospitals and n=6, 40% private hospitals); 5%...
A total of 1343 respondents filled out the questionnaire. However, there were 24.65% (331/1343) of invalid or duplicate data, so the total valid data from filling out the questionnaire were from 75.35% (1012/1343) of respondents. A total of 37.55% (380/1012) of respondents were male, 62.45% (632/1012) were female, and most (606/1012, 59.88%) lived in Greater Jakarta. Most respondents were aged 20 to 30 years (376/1012, 37.15%; Table 1).

**Table 1.** Demographics of questionnaire respondents (n=1012).

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Respondents, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>380 (37.5)</td>
</tr>
<tr>
<td>Female</td>
<td>632 (62.5)</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
</tr>
<tr>
<td>17-20</td>
<td>207 (20.5)</td>
</tr>
<tr>
<td>20-30</td>
<td>376 (37.2)</td>
</tr>
<tr>
<td>31-40</td>
<td>148 (14.6)</td>
</tr>
<tr>
<td>41-50</td>
<td>134 (13.2)</td>
</tr>
<tr>
<td>51-60</td>
<td>125 (12.4)</td>
</tr>
<tr>
<td>&gt;60</td>
<td>22 (2.2)</td>
</tr>
<tr>
<td><strong>Domicile</strong></td>
<td></td>
</tr>
<tr>
<td>Greater Jakarta</td>
<td>606 (59.9)</td>
</tr>
<tr>
<td>Java island other than Greater Jakarta</td>
<td>279 (27.6)</td>
</tr>
<tr>
<td>Outside Java island</td>
<td>78 (7.7)</td>
</tr>
<tr>
<td><strong>Education level</strong></td>
<td></td>
</tr>
<tr>
<td>Primary school</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Junior high school</td>
<td>1 (0.1)</td>
</tr>
<tr>
<td>Senior high school</td>
<td>315 (31.1)</td>
</tr>
<tr>
<td>Diploma</td>
<td>62 (6.1)</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>415 (41)</td>
</tr>
<tr>
<td>Master’s degree</td>
<td>173 (17.1)</td>
</tr>
<tr>
<td>Doctorate</td>
<td>46 (4.5)</td>
</tr>
<tr>
<td><strong>Familiarity with the use of IT</strong></td>
<td></td>
</tr>
<tr>
<td>Excellent</td>
<td>298 (29.4)</td>
</tr>
<tr>
<td>Good</td>
<td>519 (51.3)</td>
</tr>
<tr>
<td>Okay</td>
<td>191 (18.9)</td>
</tr>
<tr>
<td>Bad</td>
<td>2 (0.2)</td>
</tr>
<tr>
<td>Very bad</td>
<td>2 (0.2)</td>
</tr>
</tbody>
</table>

**Requirements**

**Health Organization Requirements**

The functions recommended by respondents from health organizations included functions related to access to health records for patients, such as viewing diagnostic data, laboratory and examination results, and medical history:

"They can view data from medical visits, such as lab results, and x-ray results. Then, you can also see the medical history, including the diagnosis." [Head of IT, General hospital (GH) 3]

Respondents also mentioned the need for a function to view health facility profiles, such as the list of services and the
availability of beds. Other recommended functions were paying for medical expenses, billing, and claiming health insurance:

- List of services that can be provided because each hospital is different. [Physician, GH8]
- If there is a bill for the patient, it will be created to be given to the patient. [Health application vendor, vendor (VDR) 1]

Respondents also suggested the need for a function for patients to manage information related to medication consumption, such as viewing the history of medication that has been consumed, ordering medication, and scheduling medication consumption:

- Ordering medicine according to a prescription, then delivery of the medicine. [Health application vendor, VDR3]
- There is a function for medication consumption. [Head of IT, primary health care (PHC) 8]

Another recommended function was a feature that patients can use to interact or communicate on the web with medical personnel in health facilities. Communication can occur through chat, messages related to health consultations, or video calls:

- Communication to hospital and telemedicine with video. [Head of IT, private hospital (PH) 3]

In addition, respondents mentioned the need for features for patients to manage appointments with medical personnel at health facilities. With this feature, patients can register themselves for treatment at health facilities, including choosing a physician:

- There is an online registration for patients...there must be information on what time they should be treated. [Head of IT, GH6]

Respondents also recommended a function for patients to access disease-related information and health tips, such as how to maintain a healthy lifestyle. Others suggested a function for patients to manage their health data to support preventive health efforts or disease prevention. In addition, this function could assist in the recovery from certain diseases that require ongoing health management activities. For example, patients could input data on vital signs and physical activities:

- There are health articles that can be used when other features are not used. The available information depends on patients’ health conditions. This could include articles on healthy lifestyles such as safe cosmetics, or nail care. [Head of IT planning strategy, health regulator (HR) 2]
- Monitoring tracking is also a very good opportunity because health is not only about curing, we also need preventive measures so that we don’t get sick. [Head of IT, PH2]

In addition to functional requirements, respondents mentioned the need for PHR integration. PHR functionality can be integrated with health applications or other existing data sources, such as electronic medical records in health facilities, to obtain medical summaries, such as diagnoses, laboratory and examination results, and medical history. PHRs also need to be integrated with the referral information system (sistem informasi rujukan terintegrasi) to obtain patient referral history and integration with vaccine data, especially for needs during the COVID-19 pandemic:

- It needs to be integrated with electronic medical record data from health facilities. [Member of data and information center, HR1]
- We need to integrate data from SISRUTE because it records data from several health facilities. If the patient is referred, the data should be recorded. [Head of IT, GH5]

PHRs also need to be integrated with BPJS Kesehatan for JKN patients and integrated with web-based payments for patients who seek treatment at health facilities and do not use health insurance:

- Need to be integrated with BPJS health. [Head of IT, GH3]
- For payments, it is integrated with online payments. [Health application vendor, VDR3]

PHRs can be integrated with teleconsultation applications for communication between patients and medical personnel, as well as with pharmacies for ease of ordering medicines. In addition, for the convenience of monitoring personal health data such as physical activities, PHRs need to be integrated with wearable devices. Moreover, PHRs need to be integrated with the healthy family or Program Indonesia Sehat dengan Pendedekan Keluarga application to support healthy family programs at the Puskesmas and with the national health data repository owned by the Ministry of Health:

- For prescription, the application needs to be connected to the pharmacy. [Physician, PH5]
- Regarding fitness, it is difficult to implement, unless you can access data from wearable devices such as smartwatches. [Member of data and information center, HR1]
- We have a healthy family application (PIS-PK) to record family information related to individual family members, and it is not from the results of the medical examination. [Head of IT, HR1]

Respondents also mentioned security aspects that need to be applied to PHRs, such as access control, audit trails, data encryption, and data backup. Authentication and authorization are needed in the implementation of PHRs. An audit trail is required to review who is accessing and what data have been accessed in the PHR. A data backup option is required to avoid the risk of data loss. In addition, PHRs need to implement important data encryption, such as passwords:

- There must be a data backup. [Head of IT, GH2]
- There is a log in the application to see what time the user logged in and what features were accessed. [Head of IT, PHC1]

PHRs need to implement user manuals or guidance options to help users understand the information contained in them. In addition, PHRs need to have customization options based on the availability of the internet network as several regions in Indonesia have poor internet connections:
Provide user manuals, video manuals, or readable manuals. [Health application vendor, VDR1]

For areas with poor internet network, the application should still be accessible. [Physician, GH8]

Patient Requirements
A total of 70.06% (709/1012) of respondents had used health applications, whereas 29.94% (303/1012) had never used health applications. For each respondent who had used a health application, the questions asked were the health application used, the platform used to access the health application, the length of use of the health application, the frequency of use of the health application in the last 6 months, the features used in the health application, reasons for using health applications, challenges when using health applications, organizations that must be integrated or connected with health applications, and the most important components of a health application (Multimedia Appendix 4).

Respondents were also asked to rate how important the PHR functionalities were based on the previous review (Harahap et al [5]), which comprised health records, administrative records, medication management, communication, appointment management, education, self–health monitoring, and supporting function. The functionality codes for each PHR module are summarized in Multimedia Appendix 5. Respondents were asked to provide an assessment with the following scores: 1=Very Not Important, 2=Not Important, 3=Optional, 4=Important, and 5=Very Important. The mean was then calculated for each functionality. If the mean of the functionality was <4, then the functionality did not need to be implemented in the integrated PHR model. On the basis of the results of the questionnaire, 27 functionalities had a mean of >4. However, the functionality of sending messages or chatting with support groups and family members in the PHR application (Communication_03) and the functionality to connect wearable devices with the PHR application were considered important by fewer respondents (SelfHealthMonitoring_05), with a mean of <4. Figure 2 shows the respondents’ scores for each PHR functionality.
Summary of User Requirements

The requirements of health organizations and patients were grouped into PHR modules and functionalities based on the study by Harahap et al [5], which comprised health records, administrative records, medication management, communication, appointment management, education, and self-health monitoring. In addition, there were emergency modules obtained from patients’ requirements as well as security modules and supporting functions to meet the nonfunctional requirements of PHRs. In the health record module, there were functions to view the results of medical examinations, medical history, referrals, and vaccinations. In the administrative records module, there were functions for patient profiles, health facility profiles, physician profiles, health insurance, and payments and billing. In the medication management module, there were functions for medication history, medication reminders, and medication orders. In the communication module, there was a messaging function (SMS text message or video call). In the appointment management module, there were functions for registration, appointment history, reminders or notifications, and ambulance. In the education module, there was a health article function. In the self–health monitoring module, there were health data tracking functions, health dashboards, health calculators, and early warning notifications. In the emergency module, there was an emergency contact function. In the security module, there were authentication, authorization, audit log, and backup functions. In the supporting function module, there were user manual and offline functionalities. A summary of the PHR

<table>
<thead>
<tr>
<th>HealthRecord_01</th>
<th>HealthRecord_02</th>
<th>HealthRecord_03</th>
<th>HealthRecord_04</th>
<th>HealthRecord_05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very important</td>
<td>Important</td>
<td>Optional</td>
<td>Not important</td>
<td>Very not important</td>
</tr>
<tr>
<td>443</td>
<td>490</td>
<td>493</td>
<td>474</td>
<td>413</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AdministrativeRecord_01</th>
<th>AdministrativeRecord_02</th>
<th>AdministrativeRecord_03</th>
<th>AdministrativeRecord_04</th>
<th>AdministrativeRecord_05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very important</td>
<td>Important</td>
<td>Optional</td>
<td>Not important</td>
<td>Very not important</td>
</tr>
<tr>
<td>498</td>
<td>463</td>
<td>441</td>
<td>408</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MedicationsManagement_01</th>
<th>MedicationsManagement_02</th>
<th>MedicationsManagement_03</th>
<th>MedicationsManagement_04</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very important</td>
<td>Important</td>
<td>Optional</td>
<td>Not important</td>
</tr>
<tr>
<td>461</td>
<td>462</td>
<td>359</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Communication_01</th>
<th>Communication_02</th>
<th>Communication_03</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very important</td>
<td>Important</td>
<td>Optional</td>
</tr>
<tr>
<td>421</td>
<td>422</td>
<td>233</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AppointmentManagement_01</th>
<th>AppointmentManagement_02</th>
<th>AppointmentManagement_03</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very important</td>
<td>Important</td>
<td>Optional</td>
</tr>
<tr>
<td>364</td>
<td>358</td>
<td>230</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education_01</th>
<th>SelfHealthMonitoring_01</th>
<th>SelfHealthMonitoring_02</th>
<th>SelfHealthMonitoring_03</th>
<th>SelfHealthMonitoring_04</th>
<th>SelfHealthMonitoring_05</th>
<th>SupportingFunction_01</th>
<th>SupportingFunction_02</th>
<th>SupportingFunction_03</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very important</td>
<td>Important</td>
<td>Optional</td>
<td>Not important</td>
<td>Very not important</td>
<td>Very important</td>
<td>Important</td>
<td>Optional</td>
<td>Not important</td>
</tr>
</tbody>
</table>
module and functionality based on the requirements of the respondent group is provided in Multimedia Appendix 6.

**Architecture Development**

**Overview**

This study used The Open Group Architecture Framework (TOGAF) to design the architecture of an integrated PHR system in Indonesia. On the basis of previous studies, TOGAF provides a complete process and methodology to develop architecture [31]. Moreover, TOGAF is the most suitable architectural framework for application in the health sector as it provides a complete architectural development process and can be adapted to the health sector [32,33]. The TOGAF referred to in this study is the TOGAF version 9.2. The Architecture Development Method in the TOGAF can be modified to suit specific needs [34]. In this study, the scope of architectural development was to design information system architecture. Therefore, the TOGAF components needed comprised architecture vision, business architecture, application architecture, data architecture, and technology architecture as required components in designing information system architecture [35].

**Architecture Vision**

According to the TOGAF 9.2, an architectural vision is a brief description of the target architecture that describes the business value and changes that will result from successful implementation. Architectural vision serves as a vision and boundary in the development of a more detailed architecture [34]. The value that is expected to be provided by the integrated PHR is a complete medical history by allowing patients to obtain medical information from different health facilities. Integrated PHRs can also minimize unnecessary health examinations as patients can share their medical history with their physicians so that physicians have information about previous examinations that have been carried out by patients. In addition, the integrated PHR facilitates communication between patients and physicians and helps patients with administrative activities such as registration, appointments, and payment of medical expenses. Integrated PHRs can also help patients manage health outside the health care environment, such as tracking food consumption and physical activity according to the patient’s needs and health conditions.

On the basis of the user requirements and literature review, we formulated architecture principles for integrated PHRs in Indonesia following the TOGAF 9.2 (Table 2). These comprise business, data, application, and technology principles [34]. The business principles comprise information management as everybody’s business, business continuity, service orientation, compliance with the law, and patient-centeredness. Data principles comprise data being an asset, shared, and accessible; common vocabulary and data definitions; and data security. Application principles comprise technological independence and ease of use as well as functionality completeness. The technology principles comprise interoperability and ease of access.
Table 2. Architecture principles.

<table>
<thead>
<tr>
<th>Domain and principle</th>
<th>Description</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Business</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Information management is everybody’s business | To support health services, health organizations and patients need to be involved in managing information on the PHR. | • The Open Group [34]  
• Harahap et al. [36] |
| Business continuity  | The PHR has an optional function that allows users to use it with a poor internet connection. | • The Open Group [34]  
• Harahap et al. [36] |
| Service orientation  | The services provided by the PHR are integrated with health care activities in Indonesia. | • The Open Group [34]  
• Harahap et al. [36] |
| Compliance with the law | The PHR needs to comply with all applicable laws, policies, and regulations in Indonesia. | • The Open Group [34]  
• Harahap et al. [36] |
| Patient-centeredness | The PHR is designed for patients to have health information from various sources. | • The Open Group [34]  
• Harahap et al. [36] |
| **Data**             |                                                                             |                                                 |
| Data are an asset    | The data on the PHR could assist the decision-making of health care providers or patients in managing their health. | • The Open Group [34]  
• Harahap et al. [36] |
| Data are shared      | Data can be shared between patients and health care providers.               | • The Open Group [34]  
• Harahap et al. [36] |
| Data are accessible  | Access to accurate data is needed to improve quality and efficiency in the management of patient health. | • The Open Group [34]  
• Harahap et al. [36] |
| Common vocabulary and data definition | The data on the PHR must have the same definition across health organizations to allow for data sharing. | • The Open Group [34]  
• Harahap et al. [36] |
| Data security        | A security mechanism is needed to protect the data stored or exchanged on the PHR. | • Harahap et al. [5]  
• The Open Group [34]  
• Harahap et al. [36] |
| **Application**      |                                                                             |                                                 |
| Technology independ- | The PHR can be integrated with various applications on different platforms. | • The Open Group [34]  
• Harahap et al. [36] |
| ence                 |                                                                            |                                                 |
| Functional completeness | The PHR provides functionalities to support promotive, preventive, curative, and rehabilitative health services in Indonesia. | • Harahap et al. [5,36] |
| Ease of use          | The PHR should be easy to use so that users can complete their tasks.       | • Harahap et al. [5]  
• The Open Group [34]  
• Harahap et al. [36] |
| **Technology**       |                                                                             |                                                 |
| Interoperability     | The PHR needs to have interoperability standards for sharing data among stakeholders and health information systems. | • Harahap et al. [5]  
• The Open Group [34]  
• Harahap et al. [36] |
| Ease of access       | The PHR needs to be implemented using common technology used by the community to facilitate easy access to information. | • Harahap et al. [36]  
• Kharrazi et al. [37] |

aPHR: personal health record.

**Business Architecture**

Business architecture defines the business strategy, governance, organization, and key business processes [34]. On the basis of the interviews with participants from health facilities and a review of health regulations in Indonesia, the business process for PHRs can be divided into health care business processes, self–health monitoring business processes, vaccination business processes, and home care business processes. Each of these process flows can be seen in Multimedia Appendix 7. The main parties involved in the PHR are patients, physicians, health facilities, laboratories, the Palang Merah Indonesia or Indonesian Red Cross (PMI), BPJS Kesehatan or other private health insurance companies, and pharmacies.
On the basis of the flow of the process, we added a rich picture to describe the data exchange on PHRs in the health care process (Figure 3). Patients register to make visits to health facilities or laboratories. Patients who need blood donors can also submit a blood donation request to the PMI. Patients can share their health records or personal health monitoring data with their physicians. Patients who have received health services will receive their data in the PHR. Patients can connect the PHR with wearable devices to record personal health monitoring data. Patients can connect the PHR to BPJS Kesehatan or other health insurance companies to obtain their status. Patients can communicate with their physicians or receive their prescribed medicine at pharmacies through the mHealth apps or teleconsultation applications that are connected with the PHR. Moreover, patients can receive vaccination data through the PHR, which is linked to the Ministry of Health vaccine data (satu data vaksinasi). The PHR can also connect with an integrated emergency management system (sistem penanggulangan gawat darurat terpadu) to obtain the nearest emergency service contacts.

Figure 3. Rich picture of personal health record (PHR) data exchange in the health care process. BPJS: Badan Pelaksana Jaminan Sosial Kesehatan or Social Security Agency for Health; mHealth: mobile health; PMI: Palang Merah Indonesia or Indonesian Red Cross; SPGDT: sistem penanggulangan gawat darurat terpadu or integrated emergency management system.

Application Architecture

The TOGAF 9.2 defines application architecture as a blueprint for the applications to be developed, their interactions, and their relationship to the organization’s core business processes [34].

On the basis of the results of interviews and questionnaires, parties that need to be integrated into PHRs are primary health facilities (Puskesmas and clinics), referral health facilities (government and private hospitals), health laboratories,
pharmacies, satu data vaksinasi, mHealth app or teleconsultation application providers, BPJS Kesehatan, private health insurance companies, and the Ministry of Health. In addition, PHRs need to have the option to connect with wearable devices for tracking health data and family members. On the basis of the results of the business architecture design, other parties that need to be integrated with PHRs are PMI and the integrated emergency management system (sistem penanggulangan gawat darurat terpadu). Health facilities are the parties that are most integrated with the information in the PHR. The PHR is integrated with hospital information systems (sistem informasi manajemen rumah sakit) in referral health facilities and primary health care information systems (sistem informasi puskesmas) in primary health facilities to access patient referrals. Especially for Puskesmas, the PHR can be integrated with the Program Indonesia Sehat dengan Pendekatan Keluarga information system. Figure 4 summarizes the integration of the PHR with health information systems or health care providers in Indonesia.

The modules in the PHR comprise health records, administrative records, medication management, communication, appointment management, education, self–health monitoring, security, supporting functions, and emergency. In the health record module, the functionalities comprise a medical summary (results of physical examinations and medical support as well as disease and medication history), referrals, and vaccinations. This module also adds the functions of family planning, home care, and blood donors based on the identification of activities in the business architecture. In the administrative record module, the functionalities comprise patient profiles, health facility profiles, health personnel profiles, health insurance, and payments and billing. In the medication management module, the functionalities comprise medication history, medication reminders, and medication orders. In the communication module, the functionalities comprise messaging. In the appointment management module, the functionalities comprise registration, appointment history, reminders, notifications for appointments, and ambulance services. In the education module, the functionalities comprise health articles containing information on disease problems and health tips. In the self–health monitoring module, the functionalities comprise health data tracking, health dashboards, health calculators, and
alert notifications. In the emergency module, the functionalities comprise emergency contacts.

We also designed the modules and functionalities to meet the nonfunctional requirements of PHRs. The modules comprise security and supporting functions. In the security module, the functionalities comprise authentication, authorization, audit logs, and data backup. In the supporting function module, the functionalities comprise user manual and offline functionality. The functionalities that need to be prioritized for implementation in PHRs are functions related to health management (health care, health prevention, and health promotion) and functions to support information security, whereas other functionalities such as functions related to administration and supporting functions are optional. To access the PHR, the platform used is a smartphone as, based on the results of the questionnaire, it is the most widely used platform by the public to access health applications. Figure 5 summarizes the integrated PHR system model in Indonesia.

Figure 5. Modules and functionalities of the integrated personal health record (PHR) system in Indonesia. BPJS: Badan Pelaksana Jaminan Sosial Kesehatan or Social Security Agency for Health; mHealth: mobile health; PMI: Palang Merah Indonesia or Indonesian Red Cross; SPGDT: sistem penanganan gawat darurat terpadu or integrated emergency management system.

Data Architecture
The TOGAF 9.2 defines data architecture as the logical and physical structure of the organization’s data assets and data management resources [34]. We grouped data in PHRs into 3 data categories comprising master, transaction, and reference data [38,39]. Transaction data relate to data that are recorded every time a transaction occurs, such as medical records, vaccinations, and health referrals. Master data are data that do not change and do not need to be recorded in every transaction, such as patient and health facility data. Reference data are a collection of values or classifications that can be referenced by master and transaction data.

Technology Architecture
The technology architecture for PHRs is described in the form of a high-level architecture to illustrate the technology required for PHR implementation and integration with other systems (Figure 6). Patients or individuals access the PHR through a mobile app. PHR development uses the React Native cross-platform app development that can be implemented on the Android or iOS platforms. The PHR application is accessed by users via the internet, and a firewall is used for security. The PHR server comprises an application server and a database server. The application server provides access to data for the user, whereas the database server provides the data requested by the application server [40].
An application programming interface (API) is used as an intermediary for interaction between the PHR and other information systems. The type of API used is Fast Healthcare Interoperability Resources (FHIR). FHIR is an international standard recommended by the Ministry of Health to solve the problem of data exchange in health information systems in Indonesia [17]. FHIR is flexible and can be adapted to stakeholder needs, clinical specifications, and health policies. FHIR can be used to manage a single data entity (eg, heart rate), groups of data entities (eg, vital signs, medications, and allergies), or electronic recording systems such as PHRs. Therefore, FHIR is suitable for exchanging data on PHRs as PHRs aim to collect and exchange individual health data [41].

Prototype Development
The main actors involved in health care and management activities are patients and physicians. Patients are actors who receive health services and play a role in managing their health through applications. Physicians are actors who provide health services to patients. Physicians comprise general practitioners and specialists. The functionalities developed in the prototype design are priority functions defined in the application architecture, which comprise medical summaries, referrals, vaccinations, health facility profiles, physician profiles, patient profiles, messaging, medication history, medication reminders, medication orders, health data tracking, health calculators, health articles, and notifications. An explanation of the design requirements for each function and the actors involved is presented in Table 3. We developed a high-fidelity prototype for a mobile app. Some examples of patient and physician prototype designs are shown in Figures 7 and 8, respectively.
### Table 3. Design requirements for application prototype.

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
<th>Actor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical summary</td>
<td>View the history of patient visits to health facilities, including detailed examination results</td>
<td>Patient and physician</td>
</tr>
<tr>
<td>Referral</td>
<td>View patient referral history and detailed examination results from each patient referral</td>
<td>Patient and physician</td>
</tr>
<tr>
<td>Vaccination</td>
<td>View a patient’s vaccination history</td>
<td>Patient and physician</td>
</tr>
<tr>
<td>Health facility profile</td>
<td>Search for health facilities and see the nearest health facility</td>
<td>Patient</td>
</tr>
<tr>
<td>Health facility profile</td>
<td>View detailed health facility information</td>
<td>Patient and physician</td>
</tr>
<tr>
<td>Physician profile</td>
<td>View the profile of physicians who have treated patients</td>
<td>Patient</td>
</tr>
<tr>
<td>Physician profile</td>
<td>Manage physician profile</td>
<td>Physician</td>
</tr>
<tr>
<td>Patient profile</td>
<td>Manage patient profiles, including adding family members, connecting with BPJS(^a) Kesehatan or other health insurance companies, and connecting with wearable devices</td>
<td>Patient</td>
</tr>
<tr>
<td>Patient profile</td>
<td>View a list of patients who have been treated</td>
<td>Physician</td>
</tr>
<tr>
<td>Messaging</td>
<td>View lists and details of messages between patient and physician</td>
<td>Patient and physician</td>
</tr>
<tr>
<td>Medication history</td>
<td>View a list of past or current medications</td>
<td>Patient</td>
</tr>
<tr>
<td>Medication reminder</td>
<td>Manage reminders to take medication</td>
<td>Patient</td>
</tr>
<tr>
<td>Medication order</td>
<td>Obtain the prescribed medicine at the pharmacy</td>
<td>Patient</td>
</tr>
<tr>
<td>Health data tracking</td>
<td>Record health monitoring data such as physical activity, food consumption, and others</td>
<td>Patient</td>
</tr>
<tr>
<td>Health data tracking</td>
<td>View the health monitoring dashboard that has been created by the patient</td>
<td>Patient and physician</td>
</tr>
<tr>
<td>Health calculator</td>
<td>Perform patient health calculations such as BMI</td>
<td>Patient</td>
</tr>
<tr>
<td>Health article</td>
<td>Read health articles, such as information about disease problems or health tips</td>
<td>Patient</td>
</tr>
<tr>
<td>Notification</td>
<td>Receive notifications such as incoming messages, documents sent, or reminders to perform certain activities</td>
<td>Patient and physician</td>
</tr>
</tbody>
</table>

\(^a\)BPJS: Social Security Agency for Health or Badan Pelaksana Jaminan Sosial Kesehatan.

### Figure 7. Example of the home page, medical summary, and health data tracking in the patient prototype.
Artifact Evaluation

Respondent Demographics
To evaluate the architectural design and application prototype, we conducted interviews with IT or eHealth experts. Interviews were conducted with 6 respondents: 1 (17%) respondent from the Ministry of Health, 1 (17%) academician, 1 (17%) respondent from a government hospital, 1 (17%) respondent from a private hospital, and 2 (33%) health application vendors. Interviews were conducted from April 5, 2022, to April 8, 2022, with an interview duration of 40 to 60 minutes. The information of the respondents is presented in Table 4.

Table 4. Respondent demographics.

<table>
<thead>
<tr>
<th>Respondent code</th>
<th>Sex</th>
<th>Role</th>
<th>Work experience (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>Male</td>
<td>Health application vendor</td>
<td>1-5</td>
</tr>
<tr>
<td>E2</td>
<td>Female</td>
<td>Academician</td>
<td>1-5</td>
</tr>
<tr>
<td>E3</td>
<td>Male</td>
<td>IT management in the Ministry of Health</td>
<td>&gt;10</td>
</tr>
<tr>
<td>E4</td>
<td>Male</td>
<td>IT management in a government hospital</td>
<td>&gt;10</td>
</tr>
<tr>
<td>E5</td>
<td>Male</td>
<td>IT management in a private hospital</td>
<td>&gt;10</td>
</tr>
<tr>
<td>E6</td>
<td>Male</td>
<td>Health application vendor</td>
<td>&gt;10</td>
</tr>
</tbody>
</table>

Architecture Evaluation
The evaluation was carried out to assess the suitability of the integrated PHR design for the needs of health services in Indonesia (Table 5). All respondents (6/6, 100%) stated that the architectural vision and business architecture in the PHR architectural design described the needs of health services in Indonesia. Regarding the integration in the application architecture, there were several recommendations from respondents regarding parties that needed to be integrated with the PHR but were not described in the application architecture design. Other parties that need to be integrated with the PHR include billing gateways for the payment function (respondent E1) and the Directorate General of Population and Civil Registration for patient identity (respondents E2 and E3). A respondent (E4) suggested adding a health screening function to the PHR. Another respondent (E3) suggested making the messaging and medication order functions optional as they can be connected with other applications. Regarding data architecture, respondent E6 commented that the data architecture was sufficient as long as there was an explanation of the data source in the PHR. For security, respondents suggested more options for authentication methods, such as biometrics (respondent E1) and face recognition (respondent E5). Regarding technology architecture, 100% (6/6) of the respondents stated that the use of an API was a suitable solution for integration between PHRs and other applications in Indonesia. Respondent E3 commented that FHIR was the right type of API to use for PHR implementation. For security and privacy needs, respondents also agreed with the use of firewalls in the technology architecture.
On the basis of the evaluation with IT and eHealth experts, improvements were made by adding billing gateways and Directorate General of Population and Civil Registration to the application architecture (Figure 9). Improvements were also made to the modules and functionality of the integrated PHR system in Indonesia in application architecture (Figure 10). In the medication management module, the medication order function was changed from a priority function to an optional function that can be connected with mHealth apps or teleconsultation applications in Indonesia. Improvements were also made by adding a health screening function to the self–health monitoring module. In the security module, authentication methods were added, including passwords, biometrics, and face recognition. Descriptions of each module and functionality in the PHR are summarized in Multimedia Appendix 8. Improvements to the data architecture were made to add health screening data to the transaction data category. Data categories with data groups and descriptions in the PHR are described in Multimedia Appendix 9.

Table 5. Summary of personal health record (PHR) architecture evaluation results.

<table>
<thead>
<tr>
<th>Architecture component</th>
<th>Evaluation criteria</th>
<th>Evaluation results</th>
<th>Respondent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture vision</td>
<td>Suitability for the health services in Indonesia</td>
<td>It is sufficient to describe the needs of health services in Indonesia</td>
<td>E1, E2, E3, E4, E5, and E6</td>
</tr>
<tr>
<td>Business architecture</td>
<td>Conformity with the process of health services in Indonesia</td>
<td>It is sufficient to describe the needs of health services in Indonesia</td>
<td>E1, E2, E3, E4, E5, and E6</td>
</tr>
<tr>
<td>Application architecture</td>
<td>Integration with health information systems or other parties</td>
<td>Integration with billing gateway</td>
<td>E1</td>
</tr>
<tr>
<td>Application architecture</td>
<td>Integration with health information systems or other parties</td>
<td>Integration with Dukcapil</td>
<td>E2 and E3</td>
</tr>
<tr>
<td>Application architecture</td>
<td>Completeness of functionality</td>
<td>Messaging and medication orders as optional functions</td>
<td>E3</td>
</tr>
<tr>
<td>Application architecture</td>
<td>Completeness of functionality</td>
<td>Addition of health screening function</td>
<td>E4</td>
</tr>
<tr>
<td>Application architecture</td>
<td>Completeness of functionality</td>
<td>Addition of authentication method options such as biometrics and face recognition</td>
<td>E1 and E5</td>
</tr>
<tr>
<td>Data architecture</td>
<td>Data requirements and completeness</td>
<td>Data architecture is sufficient as long as there is an explanation of the data source in the PHR</td>
<td>E6</td>
</tr>
<tr>
<td>Technology architecture</td>
<td>Technology requirements for PHR implementation</td>
<td>The architecture already describes the technology requirements for PHR implementation</td>
<td>E1, E2, E3, E4, E5, and E6</td>
</tr>
</tbody>
</table>

Dukcapil: Directorate General of Population and Civil Registration.
Figure 9. Design improvements to the data exchange in the personal health record (PHR) in Indonesia. BPJS: Badan Pelaksana Jaminan Sosial Kesehatan or Social Security Agency for Health; Dukcapil: Directorate General of Population and Civil Registration; PIS-PK: Program Indonesia Sehat dengan Pendekatan Keluarga; PMI: Palang Merah Indonesia or Indonesian Red Cross; SIMPUS: sistem informasi puskesmas or primary health care information system; SIMRS: sistem informasi manajemen rumah sakit or hospital information system; SISRUTE: sistem informasi rujukan terintegrasi or referral information system; SPGDT: sistem penanggulangan gawat darurat terpadu or integrated emergency management system.
Figure 10. Design improvements to the modules and functionalities of the integrated personal health record (PHR) system in Indonesia. BPJS: Badan Pelaksana Jaminan Sosial Kesehatan or Social Security Agency for Health; Dukcapil: Directorate General of Population and Civil Registration; mHealth: mobile health; PMI: Palang Merah Indonesia or Indonesian Red Cross; SPGDT: sistem penanggulangan gawat darurat terpadu or integrated emergency management system.

Prototype Evaluation

The evaluation of the prototype design resulted in suggestions for improvements related to the main functions developed in the prototype design (Table 6). Suggestions for improvements that need to be made were the functions of medical summary, referral, vaccination, physician profile, messaging, medication history, medication reminder, medication order, health data tracking, notification, and patient profile. In the medical summary, referral, and vaccination functions, a respondent (E3) suggested adding a patient identification number. Some examples of the prototype improvements are shown in Figure 11.

In the medical summary function, suggestions for improvement were the addition of the patient’s overall medical history (respondent E3) and the addition of the patient’s medical record number (respondent E5) to the medical summary details. The addition of the patient’s overall medical history aimed to make it easier for patients to view their complete medical history without having to look at the details of each medical summary one by one. The patient’s medical record number was intended to be the patient’s identity number at the health facility.

In the referral function, suggestions for improvement were the addition of information on the actions given before the patient was referred (respondent E2) in the referral details. Other suggestions for improvement were the addition of information on the type of referral, such as back referral (respondent E2). This information is needed for the continuous treatment of patients as health workers who treat patients need to know the complete condition of the patient.

In the vaccination function, suggestions for improvement included adding other types of vaccinations apart from COVID-19 (respondents E1, E2, E3, and E4). Additional types of vaccinations are needed so that this function can be used not only during the COVID-19 pandemic but also after it ends. In addition, in the vaccination function, it was recommended that the patient be able to view the vaccination history of family members (respondent E1).

In the physician profile function, the suggestion for improvement was changing the physician’s ID to the Surat Izin Praktik number (respondents E1, E2, and E4) in the detail of the physician profile as the Surat Izin Praktik number is the standard numbering used for every health facility in Indonesia. Another suggestion for improving this function was the deletion of information to view the recommendations of the nearest physician (respondents E3 and E5). This is because, in implementation, it will be difficult to obtain updated information on physicians’ availability around the patient’s location.
In the messaging function, the suggestions for improvement were to link this function to health applications or teleconsultation applications that already exist in Indonesia. The goal is to facilitate the implementation of the PHR application as other applications can handle the messaging function. This respondent (E3) also suggested doing the same for the medication order function.

In the medication history function, a suggestion for improvement was given by a respondent (E1) to add medication categories, such as generic or patent. The respondent also gave suggestions on the medication reminder function to add information on how to take the medication. Another suggestion for improvement in medication reminders was the need to add information on whether the medication should be fully consumed (respondent E5).

Table 6. Summary of prototype evaluation results.

<table>
<thead>
<tr>
<th>Actor and function</th>
<th>Evaluation result</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient</strong></td>
<td></td>
</tr>
<tr>
<td>Medical summary</td>
<td>• Add a medical summary dashboard to view the patient’s overall medical history (respondent E3)</td>
</tr>
<tr>
<td></td>
<td>• Add the patient’s NIK number (respondent E3)</td>
</tr>
<tr>
<td></td>
<td>• Add the patient’s medical record number to the medical summary details (respondent E5)</td>
</tr>
<tr>
<td>Referral</td>
<td>• Add reason for referral (respondent E2)</td>
</tr>
<tr>
<td></td>
<td>• Add medication treatment information to referral details (respondent E2)</td>
</tr>
<tr>
<td></td>
<td>• Add referral type (respondent E2)</td>
</tr>
<tr>
<td></td>
<td>• Add the patient’s NIK number (respondent E3)</td>
</tr>
<tr>
<td>Vaccination</td>
<td>• Add vaccine category other than COVID-19 vaccine (respondents E1, E2, E3, and E4)</td>
</tr>
<tr>
<td></td>
<td>• Add the patient’s NIK number (respondent E3)</td>
</tr>
<tr>
<td>Physician profile</td>
<td>• Change physician ID to SIP number (respondents E1, E2, and E4)</td>
</tr>
<tr>
<td></td>
<td>• Delete nearest physician recommendations (respondents E3 and E5)</td>
</tr>
<tr>
<td>Messaging</td>
<td>• Link to other health applications (respondent E3)</td>
</tr>
<tr>
<td>Medication history</td>
<td>• Add patent or generic medication category (respondent E1)</td>
</tr>
<tr>
<td>Medication reminder</td>
<td>• Add the way of taking the medication (respondent E1)</td>
</tr>
<tr>
<td></td>
<td>• Add information on whether the medication must be fully consumed or not (respondent E5)</td>
</tr>
<tr>
<td>Medication order</td>
<td>• Link to other health applications (respondent E3)</td>
</tr>
<tr>
<td>Health data tracking</td>
<td>• Add weekly health data tracking dashboard (respondent E1)</td>
</tr>
<tr>
<td></td>
<td>• Add sample health data tracking input page (respondent E5)</td>
</tr>
<tr>
<td>Notification</td>
<td>• Display changes to distinguish read from unread notifications (respondent E1)</td>
</tr>
<tr>
<td>Patient profile</td>
<td>• Add other security options such as biometrics (respondent E1)</td>
</tr>
<tr>
<td></td>
<td>• Add blood type to patient profile details (respondent E2)</td>
</tr>
<tr>
<td></td>
<td>• Add marital status to patient profile details (respondent E5)</td>
</tr>
<tr>
<td>Other</td>
<td>• Add log-in or registration options with Gmail (respondent E1)</td>
</tr>
<tr>
<td></td>
<td>• Add health screening function (respondent E4)</td>
</tr>
<tr>
<td><strong>Physician</strong></td>
<td></td>
</tr>
<tr>
<td>Patient profile</td>
<td>• Add a dashboard to view patients who need to be responded to (respondent E1)</td>
</tr>
<tr>
<td></td>
<td>• Add menu of patient health screening results (respondent E4)</td>
</tr>
<tr>
<td>Notification</td>
<td>• Add sample notification for referral (respondent E2)</td>
</tr>
<tr>
<td>Physician profile</td>
<td>• Delete the function to view the health facility profile on the physician profile (respondents E4 and E5)</td>
</tr>
</tbody>
</table>

aNIK: Nomor Induk Kependudukan (patient identification number).

bSIP: Surat Izin Praktik number (standard numbering used for every health professional or health care provider [physician, nurse, and midwife] in Indonesia).
In the health data tracking function, suggestions for improvement were the need for a weekly health dashboard to observe trends in patient health tracking (respondent E1). Another suggestion was to add a sample data input page for this function (respondent E6). In the notification function, the suggestion for improvement was the need for display adjustments to indicate the difference between read and unread notifications (respondent E1).

In the patient profile function, suggestions for improvement included the need for security options other than passwords, such as biometrics (respondent E1). In the detailed patient profile, additional information was needed, such as blood type (respondent E2) and marital status (respondent E5). In addition to the functions that have been discussed, respondent E4 gave suggestions to add a health screening function to the PHR, including independent health screening or health screening in primary health facilities.

In the evaluation of the prototype design for the physician, suggestions for improvement were the addition of a dashboard to see patients who need to be responded to (respondent E1) and the addition of patient health screening results following the suggestions for the prototype design for patients (respondent E4). Another improvement suggestion was the addition of a notification example for referral data that are sent from patients to the physician (respondent E2). Furthermore, for the user profile function, suggestions for improvement were to remove unnecessary information, such as the profile function of health facilities (respondents E4 and E5).

**Discussion**

**Principal Findings**

This study designed an integrated PHR system architecture in Indonesia and an application prototype. In Indonesia, there are various mHealth apps or teleconsultation applications, such as AloDokter, Halodoc, and Mobile JKN. AloDokter and Halodoc are connected to health care providers, mostly private clinics and hospitals, whereas Mobile JKN is for JKN patients in primary health facilities such as Puskesmas and clinics. However, the exchange of health information in these applications is one-way, from the health facility or physician to the patient. Previous research on the adoption of PHRs has also found that health facilities in Indonesia generally do not provide web-based access to patients’ health records [36]. This study provides a PHR model that is connected with various health care providers and is integrated into routine health care practice in Indonesia.

A review conducted by Hoque et al [42] stated that most health applications or mHealth research in low- and middle-income countries do not follow the design science approach. The DSR approach informs how artifact validation is carried out so that it can provide evidence that the designed artifact is useful and meets the users’ requirements [21]. This study uses a DSR approach with evaluations carried out by IT or eHealth experts so that the PHR design follows the practice of health services in Indonesia.

The architecture development based on the TOGAF can describe the need for integration into the PHR by describing who are the stakeholders involved in the PHR. Architecture development using the TOGAF can help align business processes, data, applications, and IT infrastructure [43]. The architectural design developed in this study covers the provision of essential or basic
routine health services that always exist in the community, such as health examinations, disease treatment, and immunization [44]. The development of this PHR can also support national health priorities and the Germas program, which prioritizes promotive and preventive health efforts, especially for the prevention and control of noncommunicable diseases [45].

The application prototype in this study was developed as a mobile app. The increasingly widespread use of smartphone apps in the community and the ease of access to smartphones encourage the adoption of PHRs on mobile devices, or mobile PHRs [37]. In Indonesia, the number of smartphone users has reached 167 million, or 89% of the total population [46]. The results of the questionnaire in this study on the use of health applications section also showed that smartphones were the most popular devices for patients or individuals to access health applications in Indonesia. Smartphones also offer unique features such as a camera, GPS, and touch screen that can be used to extend the usefulness of mobile PHR, such as scanning and importing paper documents, recording certain symptoms, creating videos, or scanning bar codes for medical purposes [37].

Implications

As explained in the Introduction section, previous research on PHR design in low- and middle-income countries [6-9] has involved users in exploring the needs and usability of the PHR design. However, they did not explain the integration of the PHR with other health applications. Although there is a study describing PHR integration using the distributed PHR model, this study did not involve users or stakeholders in designing the model [10]. Our study complements the gaps in previous studies in low- and middle-income countries by designing an integrated PHR in Indonesia, which involves related stakeholders in the required gathering and evaluation.

The theoretical implications of this research are the contribution to the field of PHR research by presenting design science as an approach for designing an integrated PHR system for a low- or middle-income country context that takes into account the specific characteristics of the Indonesian health care system. By using DSR, it can be ensured that the PHR model developed is based on scientific theory and methods. In addition, the DSR approach helps researchers understand existing health care systems and the needs of various stakeholders, such as patients, health care providers, and health regulators, in developing PHRs. DSR then includes evaluating the proposed PHR model, which can help identify any issues and make necessary improvements so that the designed PHR can follow the health system in Indonesia.

We developed architecture and application prototypes based on health systems in Indonesia, which comprise routine health services, including disease treatment and health examinations, as well as promotive and preventive health efforts. In addition, in this study, there are health referral functions that have not been discussed in the previous review of PHR functionality [5] and other studies in low- and middle-income countries [6-10]. This referral function is needed as the health system in Indonesia has a tiered referral program that should be followed by JKN patients [13].

The practical implication of this study is that this research is expected to be a guide for health regulators, health facilities, or health application vendors in designing an integrated PHR system in Indonesia. The architectural design in this study can provide an overview to integrate the PHR into the health services process in Indonesia, information about parties that need to be integrated into the PHR, and technology that can be used for PHR implementation. The prototype design in this study provides a guideline to implement PHR functions that focus not only on health care but also on disease prevention and health promotion.

Conclusions

The architecture design of the integrated PHR system in Indonesia refers to the TOGAF version 9.2, which is divided into five components: architecture vision, business architecture, application architecture, data architecture, and technology architecture. We developed a high-fidelity prototype for patients and physicians. The functionalities that were implemented were the priority functions defined in the architecture. The architecture evaluation stated that the architecture design had already described the needs and processes of health services in Indonesia as well as the technology needed for implementation. Improvements were made to the application architecture and data architecture to add the stakeholders that need to be integrated and the required functionality to the PHR. Prototype evaluation resulted in adding the necessary information to the functions that were developed, such as linking the medication order and messaging functions to the teleconsultation application and adding a health screening function. The limitation of this research is that the evaluation only focused on assessing the suitability of the integrated PHR model for the needs of health programs in Indonesia from the perspective of IT or eHealth experts. Future studies should be conducted to evaluate the prototype PHR from the perspective of patients and physicians as the primary users of the PHR.

Acknowledgments

This work was supported by a Pendidikan Magister Menuju Doktor untuk Sarjana Unggul grant from the Ministry of Research, Technology, and Higher Education, Republic of Indonesia, with research grant contract NKB-341/UN2.RST/HKP.05.00/2021.
Authors' Contributions
NCH designed the study, performed data collection, analyzed the data, and wrote the paper. PWH designed the study, provided writing assistance, and approved the final version to be submitted. ANH provided writing assistance and approved the final version to be submitted.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File, 20 KB - medinform_v11i1e44784_app1.docx]

Multimedia Appendix 2
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[DOCX File, 26 KB - medinform_v11i1e44784_app2.docx]

Multimedia Appendix 3
Interview respondents.
[DOCX File, 23 KB - medinform_v11i1e44784_app3.docx]

Multimedia Appendix 4
Health application use.
[DOCX File, 25 KB - medinform_v11i1e44784_app4.docx]

Multimedia Appendix 5
Functionality codes.
[DOCX File, 23 KB - medinform_v11i1e44784_app5.docx]

Multimedia Appendix 6
Summary of the personal health record module and functionality based on the requirements of health organizations and patients.
[DOCX File, 24 KB - medinform_v11i1e44784_app6.docx]

Multimedia Appendix 7
Process flows in the personal health record.
[DOCX File, 231 KB - medinform_v11i1e44784_app7.docx]

Multimedia Appendix 8
Module and functionality in the personal health record.
[DOCX File, 25 KB - medinform_v11i1e44784_app8.docx]

Multimedia Appendix 9
Data categories with data groups and descriptions in the personal health record.
[DOCX File, 23 KB - medinform_v11i1e44784_app9.docx]

References


Abbreviations

API: application programming interface
BPJS Kesehatan: Badan Pelaksana Jaminan Sosial Kesehatan or Social Security Agency for Health
COREQ: Consolidated Criteria for Reporting Qualitative Research
DSR: design science research
FHIR: Fast Healthcare Interoperability Resources
GH: General hospital
HR: health regulator
JKN: Jaminan Kesehatan Nasional or national health insurance
mHealth: mobile health
PH: private hospital
PHC: primary health care
PHR: personal health record
PMI: Palang Merah Indonesia or Indonesian Red Cross
Puskesmas: pusat kesehatan masyarakat or primary health centers
Using a Clinical Data Warehouse to Calculate and Present Key Metrics for the Radiology Department: Implementation and Performance Evaluation

Leon Liman1, MSc; Bernd May2, Dr rer nat; Georg Fette3, Dip (Informatics); Jonathan Krebs1, MSc; Frank Puppe1, Prof Dr

1Chair of Computer Science VI, Würzburg University, Würzburg, Germany
2Management und Beratung in der Medizin (MBM) Medical-Unternehmensberatung GmbH, Mainz, Germany
3Service Centre Medical Informatics, University Hospital of Würzburg, Würzburg, Germany

Corresponding Author:
Leon Liman, MSc
Chair of Computer Science VI
Würzburg University
Am Hubland
Würzburg, 97074
Germany
Phone: 49 9313189250
Email: leon.liman@uni-wuerzburg.de

Abstract

Background: Due to the importance of radiologic examinations, such as X-rays or computed tomography scans, for many clinical diagnoses, the optimal use of the radiology department is 1 of the primary goals of many hospitals.

Objective: This study aims to calculate the key metrics of this use by creating a radiology data warehouse solution, where data from radiology information systems (RISs) can be imported and then queried using a query language as well as a graphical user interface (GUI).

Methods: Using a simple configuration file, the developed system allowed for the processing of radiology data exported from any kind of RIS into a Microsoft Excel, comma-separated value (CSV), or JavaScript Object Notation (JSON) file. These data were then imported into a clinical data warehouse. Additional values based on the radiology data were calculated during this import process by implementing 1 of several provided interfaces. Afterward, the query language and GUI of the data warehouse were used to configure and calculate reports on these data. For the most common types of requested reports, a web interface was created to view their numbers as graphics.

Results: The tool was successfully tested with the data of 4 different German hospitals from 2018 to 2021, with a total of 1,436,111 examinations. The user feedback was good, since all their queries could be answered if the available data were sufficient. The initial processing of the radiology data for using them with the clinical data warehouse took (depending on the amount of data provided by each hospital) between 7 minutes and 1 hour 11 minutes. Calculating 3 reports of different complexities on the data of each hospital was possible in 1-3 seconds for reports with up to 200 individual calculations and in up to 1.5 minutes for reports with up to 8200 individual calculations.

Conclusions: A system was developed with the main advantage of being generic concerning the export of different RISs as well as concerning the configuration of queries for various reports. The queries could be configured easily using the GUI of the data warehouse, and their results could be exported into the standard formats Excel and CSV for further processing.

Keywords

data warehouse; electronic health records; radiology; statistics and numerical data; hospital data; eHealth; medical records
Introduction

Background
Examinations performed by the radiology departments of hospitals, such as creating X-ray, computed tomography (CT), magnetic resonance imaging (MRI), or ultrasound images, are fundamental for many kinds of clinical diagnoses. Therefore, optimizing the use of radiology is important for any clinician working with it as well as for any patient being examined there. Such optimization has several advantages for the hospital, such as shorter times patients need to stay there as well as the ability to perform more radiologic examinations. It also has advantages for the patient, such as shorter times to wait for the radiology appointment as well as reduced radiation exposure, if unnecessary repeated examinations of the same body region are avoided.

Objectives
This optimization requires a good overview of the various key metrics of radiologic services and their changes over time. A systematic approach for computing such metrics is building and using a radiology data warehouse. The main requirements for a radiology data warehouse solution are:

- Generic data import from the underlying radiology information system (RIS), for example, via an intermediate data format
- Tools for enriching the basic data with inferred data via a preprocessing step, which allows for more simple and compact queries on the data
- An expressive query language
- A comfortable graphical user interface (GUI) for the query language, including the ability to specify the resulting reports as tables, graphs, or a standard export format for further processing
- An efficient engine for answering queries and generating reports

These requirements are further explained in the following sections.

State of the Art
The relevance of calculating the key metrics of radiology data [1], as well as the types of metrics, that are most interesting for radiology exports [2,3] has already been described. In addition, the benefits of presenting such metrics in an easily understandable dashboard [4,5] have been explained. Although such solutions have been implemented for many different uses cases [6-9], all of them use a fixed interface to 1 or multiple specific hospital information systems and provide the user with only a fixed selection of predefined calculations. In other systems, the primary goal is to show data from individual patients [10-12], which only allows for a limited amount of filtering and no user-defined queries on the data. Other approaches use a data warehouse [13,14] to unite data from several (still fixed) hospital information systems into a unified representation and therefore allow for various user-defined queries to be executed but are missing a GUI for users to specify their queries and instead have their users either use Microsoft Excel or Structured Query Language (SQL: ISO/IEC Joint Technical Committee 1/Subcommittee 32/Working Group 3) for report creation. For importing data into a clinical data warehouse, more generic solutions exist [15,16] but without an option to calculate additional features during the import. This could make certain types of reports difficult or, depending on the query language of the clinical data warehouse, even impossible to create. These solutions are further discussed in comparison to the developed solution in the final section.

For hospitals whose data have been used during the development of this system, the state of the art for calculating key metrics of their radiology data was to do so manually in Excel. Although this allows for many different reports to be created, it has several drawbacks, which are also discussed in the final section. An intermediate result of this work has already been described [17]. This is described in more detail, together with the improvements in the final result, in the following sections.

Requirements

Generic Data Import Into a Data Warehouse
A radiology data warehouse primarily needs data of the examinations (type of modality, date and time of the request, execution, and documentation of each examination), relevant basic and radiologic patient data, the medical question for the examination as well as the radiologic diagnosis, and information about the radiologic equipment used. Since hospitals use many kinds of RIS, the use of an intermediate data format facilitates the data import and makes it independent of the internal data structure of the RIS. In this project, an Excel (or comma-separated value [CSV]) table was used as an intermediate format, in which the RIS data could be exported and from which it could then be imported into the data warehouse. If a hospital could only export its RIS data into JavaScript Object Notation (JSON) format (a proprietary one or a standard one such as Health Level Seven [HL7] Fast Healthcare Interoperability Resources [FHIR]), the relevant information from this format could also be converted into a table (using an Excel or a CSV file) that uses the structure described in the next section. All the hospitals whose data were used during the development of this system were only able to provide Excel exports of their RIS data.

Semantic Preprocessing of the Basic Data
To make queries on the radiology data easier, preprocessing of basic data is useful. Therefore, additional values were inferred from the basic data during the import into the data warehouse. Two types of preprocessing were necessary for this project: The first type was calculations performed by combining information from the basic data. Examples of such precomputed values are the difference between the time when an examination was requested and the time when it was performed as well as the time when the radiologic images were interpreted. This is usually not available in the RIS directly but can be easily computed from the individual time stamps. The second type was standardizations of the basic data. For example, the medical question for the examination could be either available as unstructured text using different wordings or as a hospital-specific code, which must be associated with a readable, standardized description, for example, by using a regular...
expression during the import. As new kinds of queries are requested, additional data may be required. Because of this, another requirement is the ability to perform an incremental update of the data warehouse with just the new data instead of deleting and reimporting everything that has already been loaded into it.

**Types of Queries and Query Language**

The developed system should be able to support a wide range of different calculations. The calculations requested by the hospitals with whose data the system has been used so far could be separated into 5 different categories, which are described here:

- **Patients, appointments, and examinations per modality:** The most common metric was the number of patients, radiology appointments, and examinations in the radiology department for each modality. Additionally, these numbers were separated between inpatients and outpatients, the department of the hospital requesting an examination, the region of the body that was examined, or the shift during the day in which the examination took place. All these numbers were used to provide a general overview of the use of the radiology department.
- **Use of radiologic devices:** A radiology department usually has many different devices for different modalities as well as often multiple devices for a single modality. To better distribute examinations and clinicians on these devices, their use is 1 of the requested calculations. The metrics for this use included the number of examinations and patients per device. Furthermore, the time for each examination as well as the vacancy between examinations were evaluated.
- **Length of a patient’s stay in the hospital:** Depending on the disease, different lengths of stay in the hospital are necessary. To evaluate whether patients were staying longer in the hospital than expected, which results in a reduced capacity for new patients, the actual stay times were compared with the ones suggested by clinical guidelines.
- **Waiting times:** Short waiting times are in the interest of both the patient and the clinician requesting a radiologic examination. Therefore, for each modality, the time between the request of an examination, the actual appointment in the radiology department, and the availability of the clinical findings after the examination were calculated and compared.
- **Multiple examinations for the same question:** To find the answer to a specific medical question, in many cases, 1 kind of radiologic examination works best. If such radiologic examination is performed directly by an experienced radiologist (who also verifies whether the requested examination makes sense for the medical question), the chances are high that only 1 examination is necessary to answer the medical question. However, if for 1 medical question, multiple examinations with the same or with different modalities are performed, the patient has increased radiation exposure and fewer radiology appointments are available for other patients. To measure this, first, all the different sequences of modalities for different kinds of medical questions were calculated. Afterward, the number of patients with such sequences were counted and compared. In addition, the total time for which a patient with such repeated examinations stays in the hospital was evaluated.

The query language used by the developed system must be able to support these kinds of queries as well as additional ones requested by the hospitals. This is also important for evaluating possible ways in which any of these metrics can be improved. For example, unnecessary multiple examinations can perhaps be explained by too few available devices for the modality recommended for a question or by missing staff to operate a device. To verify whether the measures taken are successful, the query language must also be able to analyze the change in the metrics over time.

A common set of queries for data saved in the same way furthermore allows for an easy comparison of the calculated number between different hospitals. In addition, as none of the mentioned categories depends on a specific hospital, all these calculations can be performed for any hospital (even in different countries) if it is able to provide the necessary data from its RIS.

**A Comfortable GUI for the Query Language and the Result Specification**

Although the query language should be usable in textual form, a GUI is also required to create queries in a graphical way and automatically create the corresponding textual queries. As with the query language, the GUI should also allow for the layout of the requested report to be specified. The results of queries should be shown to the user as a table or as a graph. Furthermore, the results should be exportable into the standard formats Excel and CSV so that they can be further processed. Furthermore, the GUI should make the system (with a limited amount of training) usable by the clinicians themselves and therefore should not require any knowledge of computer science.

**Efficiency Requirements**

Importing data into the data warehouse as well as creating reports using the query language on these data both should happen in a reasonable amount of time. For the initial import, the tool should not need longer than a few hours, and for querying the data, most of the queries should return their results in about 1 second, while more complex queries should not run for more than a few minutes. These requirements are necessary so that a user can quickly start to use the system and, while using it, easily try different variations of a query without a long waiting time for each result.

**Methods**

**Ethical Considerations**

In this paper only retrospective, pseudonymized patient data for patients with age groups below and above 18 years with a few attributes only about their radiologic examinations were used (dates, modality, device, localization, radiologic query, boolean values for insurance [statutory or private], boolean values for the type of stay in the hospital [inpatient or outpatient]). De-pseudonymization of the data was not possible for the authors of the study. Therefore, no ethics approval was necessary.
Concept

Processing Radiology Data for Importing It into a Data Warehouse

The data from the RIS of the hospitals were provided to the tool as an Excel, CSV, or JSON file, in which each row represents a single examination. Each column in this file is 1 attribute, and the names of these attributes are written in the first row of the file.

To map these columns to attributes in the structure of the data warehouse, a configuration file (using Excel or CSV as well) was used. This file contained 1 row for each attribute and specified the name, identifier, and data type to use (eg, numbers or texts) when importing them together with the concrete values into the data warehouse. The columns containing the required metadata (eg, identifiers) must be specified in this configuration file as well.

As mentioned in the previous section, some values for the requested reports must be calculated based on the exported RIS data. To do this, several options were offered. Additional columns were added to the RIS export performing the calculations. These were then imported into the data warehouse like any other column in the RIS data by specifying them in the configuration file. The configuration file also provided an option to replace textual values with other values, which could, for example, be used to replace an abbreviation in the RIS export with a longer form. For more advanced calculations, several programming interfaces were offered and could be implemented for any value requiring such a calculation.

All the values from the RIS, together with the calculated values, were then saved to the data warehouse, and an index was created on them for increased query performance.

Creating Reports

As soon as all the needed values were saved in the data warehouse, queries on these data were run to calculate and create the requested reports. For this purpose, a query language was used to define the structure of the report. This was done by first specifying attributes to be queried as well as constraints on the values of these attributes. In the next step, these attributes were combined with the logical operators “and”, “or,” and “not.” These single attributes or groups of attributes were then used to specify the rows and columns of the requested report. For every combination of attributes in each row and each column, a query was created, resulting in the cells of the report. If additional constraints on all cells were required, other attributes were used to specify filters. Finally, the query language specified what type of count (examinations, appointments, or patients) should be returned. All this was either specified in textual form or graphically using the GUI of the data warehouse. To create a report, the query for each cell was run on the data and, depending on the configuration, the number of examinations, appointments, or patients was returned.

By using a query language like this, it is easily possible to run modifications of a query, which is further simplified by the ability of the data warehouse to save a query and load it again later.

After a query was configured and executed, the interface of the data warehouse showed the results as a table and provided the option to export this table into the standard formats Excel and CSV. The results of some predefined queries were also shown as graphics.

Implementation

The Clinical Data Warehouse PaDaWaN

PaDaWaN (short for Patient Data Warehouse Navigator) [18] was used as the clinical data warehouse. Its core is a database containing all the used medical information and a separate index to increase the speed of queries on the data. To specify the queries, PaDaWaN uses its own query language as well as its own web interface. Furthermore, it provides the ability to export and save query results. All these parts are described in more detail in the following sections.

Database Structure

PaDaWaN stores its data in a database, which could be either a Microsoft SQL [19] or a MySQL [20] database. The table structure is based on the entity attribute value model [21]. It consists of 2 main tables, as shown in Figure 1.

The first table (DWCatalog) contains a catalog of all the possible types of information in PaDaWaN. This could represent, for example, different types of diagnoses or laboratory measurements. Each entry in this table is uniquely identified by a numeric AttributeID as well as by the combination of ExternalID (the ID in the terminology defining the entry, such as, the International Classification of Diseases [ICD] code [22] I50) and Project (the name of the whole terminology, such as “ICD”). AttributeID is automatically generated by the database and is only valid for a single installation of the system. As further explained later, only the combination of ExternalID and Project is used to identify entries from this table in a query, and therefore, only this combination must be unique among different systems if the same query should be used for all of them. For easier usage in the PaDaWaN interface, every entry has a readable name (eg, “heart failure”). The ICD terminology, for example, uses a hierarchical structure. To save this or any other hierarchy among the attributes, the ParentID field is used and contains the AttributeID of the entry, that is, the parent of the current entry. The kind of data (eg, Boolean, number, or text) that could be saved for an entry is specified with the DataType field.

In the DWInfo table, all the concrete patient data are saved. Each entry in this table is uniquely identified with an automatically generated numeric InfoID and is associated with a type of information from the DWCatalog table by its AttributeID. With the other ID fields, each entry in this table is also associated with a patient, appointment, and examination. The time and date on which a value has been recorded is saved in the MeasureTime field. The actual value (eg, the content of a patient’s discharge letter) is stored as text in the Value field.
Index Structure

To increase the speed of queries on the data in PaDaWaN’s database, it was indexed with Apache Solr [23]. Solr saves data in documents, and the schema used for PaDaWaN is shown in Figure 2.

PaDaWaN offers the ability to search for data on 3 different levels: patients, appointments, and examinations. If a search is conducted on any of these levels, all patients/appointments/examinations should be found, containing all the requested combinations of attributes and values. To accomplish this in Solr, PaDaWaN uses Solr’s feature to store documents nested in other documents. As shown in Figure 2, a document is created for each patient and each appointment. Another document is created for each examination and is stored inside the patient and appointment documents. Finally, for every value, another document is created and stored inside both examination documents. Although this approach requires more disk space as each value is saved twice, this greatly increases the speed of queries being run on the patient level compared to a document structure, where the appointments are nested inside the patient documents. Each document contains all the available IDs, as described in the previous section. Additionally, all of them contain a field named ContainingFields, which stores the AttributeIDs of all values contained in the current document. This allows a query to restrict the number of top-level documents it must search on for concrete values of the attributes. These values are stored in a field in the value document, whose name is generated by combining the attribute’s type with its ID.

Figure 2. Document structure of PaDaWaN’s Solr index using nested documents for examinations and values in separate parent documents for patients and appointments. In addition to the numeric identifiers for patients, appointments, and examinations under ContainingFields, all AttributeIDs of all values inside a document and its children are saved. The values themselves are stored in dynamic fields named with a combination of their type and AttributeID. PaDaWaN: Patient Data Warehouse Navigator.
Query Language

To specify the structure of the requested tabular result, PaDaWaN uses its own query language called Medical XML Query Language (MXQL). In the following example, the result table contains 2 rows and 2 columns. The rows contain 2 different types of modalities (X-ray and CT scan), and the columns contain 2 regions of the human body (abdomen and thoracic spine). For each combination of a row and a column, the number of matching patients from hospital A is returned. In the first cell, for example, the number of patients who undergo an X-ray examination of the abdomen is counted. This query is shown in MXQL in Figure 3 and in PaDaWaN's GUI in Figure 4. The result in Excel can be seen in Figure 5. This is a simple example used to explain the query language, PaDaWaN’s GUI, and its export capabilities, and the results shown in Figure 5 may also be retrieved directly from an RIS (depending on its capabilities).

Each query in MXQL must contain at least the following 2 elements: Query and Attribute. Query is the root XML element and contains the whole rest of the query. Attribute contains information about the catalog entry whose values should be queried. To identify this catalog entry, Attribute uses the “domain” and “extID” properties, which map to the Project and ExternalID columns of the DWCatalog database table described before. The remaining elements of the query are optional and used for more complex queries. In the example query shown in Figure 3, the Attribute elements are further constrained to only match specific values for the catalog entries. This is done with the contentOperator and desiredContent properties, where desiredContent contains a value to be matched and contentOperator the way it should be matched. The IDFilter element is used to specify on which level all the attributes in the query should be combined. In the example in Figure 3, this is set to “PID,” which means that all the attributes must have the same PatientID and that the number of matching patients should be returned by such a query. The last remaining elements of the query are DistributionRow, DistributionColumn, and DistributionFilter. They are used to return counts of multiple combinations of Attributes in a single query. Each DistributionRow becomes a row in the created result, and similarly, each DistributionColumn becomes a column. The DistributionFilter can be used to apply further constraints on all the combinations of rows and columns. Finally, the displayName property of the Attribute element can be used to provide a name for the created rows and columns. Not shown in the example is the ability of MXQL to combine multiple attributes with the logical combinations “and” and “or,” which could even be nested inside another combination. MXQL also allows for the logical operator “not” to be added to any attribute.

Here, only the MXQL features used for this project are described. A complete documentation of this query language (in German) can be downloaded from PaDaWaN’s website [24].

Figure 3. Sample query in PaDaWaN’s query language MXQL. This query returns counts of patients (specified with the filterIDType “PID”) for each combination of attributes specified as DistributionRows and DistributionColumns. In this example, the first combination would be all X-ray examinations of the abdomen. DistributionFilter restricts all the combinations to patients from hospital A. MXQL: Medical XML Query Language; PaDaWaN: Patient Data Warehouse Navigator.

```
<Query>
  <IDFilter filterTypeID="PID">
    <DistributionRow>
      <Attribute domain="generated_attributes" extID="modality_group"
               displayName="X-Ray" contentOperator="EQUALS" desiredContent="X-Ray"/>
      <Attribute domain="generated_attributes" extID="modality_group" displayName="CT"
               contentOperator="EQUALS" desiredContent="CT"/>
    </DistributionRow>
    <DistributionColumn>
      <Attribute domain="generated_attributes" extID="body_region_group"
               displayName="Abdomen" contentOperator="EQUALS" desiredContent="abdomen"/>
      <Attribute domain="generated_attributes" extID="body_region_group"
               displayName="Thoracic Spine" contentOperator="EQUALS" desiredContent="TS"/>
    </DistributionColumn>
  </IDFilter>
</Query>
```
Figure 4. Web interface of PaDaWaN with the query from Figure 3. On the left side, the catalog of available attributes is shown and can be hierarchically expanded as well as searched. On the top of the right side, the query itself can be configured by dragging items from the catalog to create rows (Zeilen in German), columns (Spalten in German), and filters. With the 3 radio buttons in the top middle, the kind of IDs to be counted can be specified. Although meaning something else in German in this project, the buttons from left to right are used for patients, appointments, and examinations. The row of buttons in the middle are used to execute a query (Suchen in German) as well as to save and load queries (Speichern and Laden in German, respectively) and to export their results. After executing a query, the bottom right of the GUI shows its result (Ergebnis in German). The remaining buttons were not used for this project. The query shown here creates rows for X-ray (Rö: short form in German) and CT scan (CT: short form in German) examinations and columns for examinations of the abdomen and the thoracic spine (BWS: short form in German). The filter then restricts everything to just examinations from the hospital (Klinik in German) A. The remaining elements of the GUI were not used for this project. CT: computed tomography; GUI: graphical user interface; PaDaWaN: Patient Data Warehouse Navigator.

Figure 5. PaDaWaN Excel export for the query from Figures 3 and 4, where the number of patients with different kinds of radiology examinations (as rows) is counted for multiple regions of the human body (as columns). CT: computed tomography; PaDaWaN: Patient Data Warehouse Navigator.

Web Interface
PaDaWaN has its own graphical web interface allowing a user to search the available attributes, graphically configure a MXQL query, and preview the result table. The interface with the MXQL query from Figure 3 looks like Figure 4.

On the left side of the interface, the content of PaDaWaN’s DWCatalog table (explained before) is shown and can be hierarchically expanded and searched. With the 3 radio buttons in the top middle of the GUI, the level on which all the attributes in the query should be found (for this project, patients, appointments, or examinations) can be configured. Via drag and drop, any attribute from the catalog can be placed in any section of the query to configure either rows (Zeilen in German), columns (Spalten in German), or filters (like that explained in the previous section). Finally, a configured query can be run by...
pressing the Search button (Suchen in German). With the Save and Load buttons (Speichern and Laden in German, respectively), a configured query can be saved and any saved query can be loaded. The next 2 buttons provide the option to export a queried result in either Excel or CSV format. The bottom of the right side of the GUI shows the tabular result (Ergebnis in German) created after running a query. By clicking any of the attributes in the query on the right side of the GUI, a dialog box appears, where, for example, the value of an attribute that should be counted can be configured. In this example these values are Rö (short in German for X-ray), CT, Abdomen, BWS (short in German for thoracic spine), and A (for the name of the hospital; Klinik in German). All the remaining buttons were not used for the queries in this project.

Export of Query Results
PaDaWaN also offers an option to export query results in either Excel or CSV format using the Excel or CSV button, respectively, in Figure 4. When the query has finished, an Excel or a CSV file is created and offered as a download. When running the query shown in MXQL in Figure 3 and in the GUI in Figure 4, the Excel export looks like that in Figure 5.

As configured in the query, each row is a different kind of radiologic examination, and each column contains a different region of the human body. As the query was configured to return the number of matching patients, the first number means that in this (small and artificially generated) data set, 19 patients underwent an X-ray examination of the abdomen.

Export of the Radiology Data
For the developed system, data from a hospital’s radiology department are needed. The system should be usable by many different hospitals with many kinds of RISs. Therefore, Excel, CSV, and JSON are used as the formats in which the RIS data can be exported and then imported from this file into PaDaWaN. As mentioned in the Introduction section, if a hospital is only able to provide its RIS data as a JSON file, this can also be transformed into a table and then saved as either an Excel or a CSV file. A part of an RIS Excel export is shown in Figure 6.

Exporting uses a simple structure, where each row represents a single examination of a patient and each column contains 1 attribute with information about the examination. The title of the attribute must be given in the first row. The only required pieces of information are the ID of the patient’s stay in the hospital, the start date and time of the examination, and the modality performed. All the remaining attributes could be different for each hospital, and the way they are imported into PaDaWaN is explained in the next section.

All the IDs that were used for this project had already been pseudonymized during the RIS export.

Import of the Radiology Data Into the Data Warehouse
The RIS export, as described in the previous section, was imported into PaDaWaN using the following steps:

- **Step 1**: A configuration file is created to specify the mapping of the RIS export columns to PaDaWaN catalog entries.
- **Step 2**: Using this configuration file, the data in the RIS export are converted to PaDaWaN database entries.
- **Step 3**: Additional precalculations are performed on the RIS data using an interface provided, and the results are saved in PaDaWaN’s database as well.
- **Step 4**: A Solr index is created on these data.

An overview of this process is shown in Figure 7. Each step is described in more detail later.
Figure 7. Overview of the process for importing radiology data into the data warehouse. First, a configuration file is created and used to import the exported RIS data into PaDaWaN’s database. On these data, additional precalculations are then performed. Finally, a Solr index is created for all the data in PaDaWaN’s database. PaDaWaN: Patient Data Warehouse Navigator; RIS: radiology information system.

**Configuration File to Map the Radiology Data to the Data Warehouse’s Structure**

An Excel (or CSV) configuration file was used to specify the mapping of the columns in the RIS export to the data structure of PaDaWaN. A configuration file for the data in Figure 6 would look like that in Figure 8.

The first row of this file must always look like that shown in Figure 8. Each of the following rows represents 1 column from the RIS export. If any of these columns should be ignored, they can be left out. The columns of the configuration file are used for specifying the PaDaWaN catalog entry that should be created (with the name from the DisplayTitle column and ExtID and DataType being directly used for database columns with the same names). The DataType “SingleChoice” is used for textual values with only a limited number of possible options (eg, modality). The ColumnName and ColumnNumber columns are used to identify a column in the RIS export. The ValueMappings column can be used to map abbreviations or codes in the RIS data to more readable names. Finally, the MetaDataType column is used to specify which columns contain which type of identifiers, the time an examination was performed, and the modality.

Figure 8. Sample of an Excel configuration file to specify the mapping between an RIS export and PaDaWaN’s data structure. The ColumnName and ColumnNumber columns must match a column in the RIS export. The DisplayTitle, ExtID, and DataType columns are mapped to the corresponding columns in PaDaWaN’s DWCatalog table. With ValueMappings, column abbreviations in the RIS export can be mapped to their longer form. The final column is used to specify which RIS column contains which type of metadata. CID: Case identifier; PaDaWaN: Patient Data Warehouse Navigator; RIS: radiology information system.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Location</td>
<td>DisplayTitle</td>
<td>ExtID</td>
<td>Data</td>
<td>Type</td>
<td>ColumnNumber</td>
</tr>
<tr>
<td>2</td>
<td>Location</td>
<td>Location</td>
<td>locatic</td>
<td>SingleChoice</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Device</td>
<td>Device</td>
<td>station</td>
<td>Text</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Start of exam</td>
<td>Start of exam</td>
<td>execSi</td>
<td>Date</td>
<td>Time</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>CID</td>
<td>Case identificcid</td>
<td>Text</td>
<td>4</td>
<td>CaseID</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Examination-é</td>
<td>Code of exar</td>
<td>taskKc</td>
<td>Text</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Examination-í</td>
<td>Name of exa</td>
<td>taskTe</td>
<td>Text</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Insurance</td>
<td>Type of insu</td>
<td>payUn</td>
<td>SingleChoice</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Type of stay</td>
<td>Type of stay</td>
<td>caseT</td>
<td>SingleChoice</td>
<td>8</td>
<td>input:inpatient;:</td>
</tr>
<tr>
<td>10</td>
<td>Modality</td>
<td>Modality</td>
<td>modal</td>
<td>SingleChoice</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

Import Process of the Radiology Data Using the Configuration File

When starting the import of the RIS export, first, the configuration file, as explained before, is read and then all the columns specified in the rows of the configuration file are imported into PaDaWaN.

For this, first, an entry in PaDaWaN’s DWCatalog table is created with the values from the configuration file. To import concrete values from any column in the RIS export into PaDaWaN’s DWInfo table, some metadata are required: PatientID, AppointmentID, and ExamID, as well as MeasureTime. These are specified with the MetaDataType column in the configuration file.

With the catalog entry and the metadata, each value in each column of the RIS export was saved into PaDaWaN’s database.

Calculating and Importing Additional Values Based on the Radiology Data

As some calculations are not possible with PaDaWaN’s query capabilities or would require complex queries, several interfaces (written in Kotlin [25]) are provided to specify additional calculations that should be executed during the RIS data import. Initially, for all these interfaces, the properties of the PaDaWaN catalog entry that should be created must be provided. Additionally, the RIS column names required for the calculation must be specified. The provided interfaces can then be used to either specify calculations that should occur for each examination (eg, calculating the shift during a day in which an examination was performed) or once for all examinations (eg,
to calculate sequences of examinations that have been performed for a single patient and for the same medical question with 1 or multiple modalities).

During the execution of all the implementations, the catalog entry specified by each implementation is created and the implemented methods to calculate the values and save them to the database are called.

Creating an Index on All Imported and Calculated Values
The last step during the import process of the RIS data is the creation of a Solr index on the data from PaDaWaN’s database. For this purpose, all the values are fetched from the database and documents in the structure described before are created. These documents are then sent to Solr, which creates its index on them.

Incremental Updates of the Data Warehouse for New or Updated Radiology Data
The process of importing all the RIS data into PaDaWaN as well as creating a Solr index on it takes some time (shown in the next section). During the work on this project, additional calculations on the RIS data, updates on existing calculations, and additional information from the RIS were needed in many cases. The whole process described in the previous section could be run again, which resulted in most values being imported or calculated again, although they did not change.

Therefore, a separate configuration file could be given to the importer, containing just the names of the attributes from the RIS export or from the implemented interfaces, that had to be processed. When using this option, just the columns and calculations of these attributes are processed and saved to the database. Afterward, Solr’s ability to perform atomic updates [26] is used. In this way, the whole documents do not have to be created and indexed again, but instead, only small parts for the updated or added attributes are deleted and then added with the new values to the existing documents.

Another possibility for new radiology data would be data from new patients. In this case, the additional data can be exported from the RIS into a separate file and then the whole import process described before can be run for just this file so that only the new data are added to the database and the index and no processing of the existing data must be done again. If a near-real-time evaluation of the data is requested by a hospital, this process can also be run immediately any time new data are added to the RIS.

Performing Calculations on the Data Using the Data Warehouse and Exporting Their Results
After the RIS data and any additional calculations on them are saved to and indexed by PaDaWaN, PaDaWaN’s web interface is used to create and run queries on these data. The usage of the interface as well as the query capabilities have already been described in the section on PaDaWaN before.

The process of creating a new query involves first specifying all the attributes for rows and columns whose combinations should be counted in the result. Optionally, additional filters can be configured for all these combinations. Next, the user chooses whether the number of matching patients, appointments, or examinations should be returned. Finally, the query is run (the matching MXQL query is automatically created by the web interface), and its result is either shown directly in the GUI or is exported to an Excel or a CSV file.

For reusing queries, PaDaWaN also provides an option to save and load queries.

PaDaWaN’s web interface uses a REST-based interface, which can also be used directly without the GUI. To do so, the query must be created as an MXQL string and can then be sent to the interface. When PaDaWaN has finished the execution of the query, the result can be received in JSON format or as an Excel or a CSV file.

Presentation of Results Calculated by the Data Warehouse
As PaDaWaN allows for exporting of results into the standard formats Excel and CSV, these results can be easily imported by many different tools to perform further calculations or to create graphics. To present the 4 most common types of calculations for the hospitals involved in this project as graphics, a simple web dashboard was created and is shown in Figure 9.

The example calculates for all inpatients the percentage of findings for 3 modalities that has been available for 1, 2, 3, or 4 days after the examination in the radiology department. These numbers are further compared between the whole radiology department and just examinations that have been requested by the neurological surgery department.

To calculate these numbers, matching PaDaWaN queries were created and saved. When opening this dashboard, the queries were loaded and executed, and the numbers were extracted from PaDaWaN’s result table.
Results

Technical Evaluation

In the following sections, details about the used data themselves as well as about the import and report creation process are presented.

Table 1. Information about the used radiology exports of 4 different hospitals from Germany.

<table>
<thead>
<tr>
<th>Details</th>
<th>Hospital A</th>
<th>Hospital B</th>
<th>Hospital C</th>
<th>Hospital D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital sites, n</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Time of data</td>
<td>2018</td>
<td>2019 to September 2020</td>
<td>2018 to June 2021</td>
<td>2019 to 2021</td>
</tr>
<tr>
<td>Patients, n</td>
<td>13,603</td>
<td>125,732</td>
<td>N/A(^a^)</td>
<td>N/A(^a^)</td>
</tr>
<tr>
<td>Appointments, n</td>
<td>28,886</td>
<td>384,186</td>
<td>307,174</td>
<td>241,148</td>
</tr>
<tr>
<td>Examinations, n</td>
<td>52,542</td>
<td>487,474</td>
<td>599,481</td>
<td>296,614</td>
</tr>
<tr>
<td>Values imported, n</td>
<td>2,50,001</td>
<td>11,650,688</td>
<td>15,014,221</td>
<td>5,339,024</td>
</tr>
<tr>
<td>Values generated, n</td>
<td>555,859</td>
<td>18,848,459</td>
<td>8,151,85</td>
<td>2,974,740</td>
</tr>
<tr>
<td>RIS(^b^) export size (MB)</td>
<td>14.9</td>
<td>75.7</td>
<td>56.1</td>
<td>21.7</td>
</tr>
</tbody>
</table>

\(^a^\)N/A: not applicable. The data from hospitals C and D contained no patient identifier, so the number of patients could not be specified.

\(^b^\)RIS: radiology information system.

Figure 9. Simple web dashboard with graphics for the 4 most common types of calculations on the radiology data of the hospitals involved in this project. The graphics in this figure show what percentage of findings for a radiological examination is available in up to 1, 2, 3, or 4 days. This is given for the 3 most common modalities and is compared between examinations in the whole radiology department and only the ones requested by neurological surgery. CT: computed tomography; MRI: magnetic resonance imaging.
The data were provided as an Excel export from the RISs of the hospitals. The last 3 hospitals had multiple radiologic sites in different cities. In addition, the time for which the data were exported was different, ranging from 1 year for the first hospital to 3.5 years for the third one. Only in the exports of the first 2 hospitals was a (pseudonymized) patient identifier included, so the number of patients could only be calculated for these 2 hospitals. Each value in a single cell of the RIS exports was imported into the data warehouse, and their number is specified in Table 1. For comparison, the number of values that were generated during the import is also specified. Numbers related to the import process are presented in the next section. Finally, in the last row, the size of the Excel files exported from the RIS is shown.

**Process of Importing the Radiology Data Into the Data Warehouse**

For the data of all 4 hospitals, a separate virtual machine (running in the internal network of the university of Würzburg) was created, and each of them was configured with 4 CPUs and 32 GB of RAM and stored on a solid-state drive (SSD). Inside of these machines was installed Ubuntu 20.04.4, together with MySQL 8.0.28, Java 11.0.14, and Solr 8.11.1. PaDaWaN’s web interface was run on an Apache Tomcat [27] 10.0.18 server. On these virtual machines, the RIS exports were imported into PaDaWaN, resulting in the numbers shown in Table 2, which are discussed in the next section.

In the second section of Table 2, numbers related to the additionally performed calculations are shown. These are the number of calculated attributes, the average time needed to calculate and save them to the database, and the total time for calculating and saving all these values.

The second-to-last section of Table 2 starts with the time needed to create a Solr index of all the imported values as well as the time needed for the whole import process of each hospital. In the last 2 rows, the size of the created database and Solr index is specified.

**Creating Reports on the Radiology Data With the Data Warehouse**

Many reports were created using the data of all 4 hospitals, depending on the requirements of each hospital. Three reports...
of different complexities, which were requested by most of the hospitals and were possible with the data provided by all of them, were created to show the time PaDaWaN needed to calculate those reports and export them as an Excel file. For each report, the number of matching examinations was restricted with MXQL to only include the data of 1 year. In all 3 reports, the 4 most common types of modalities for the hospitals (X-ray, CT, MRI, and ultrasound) were used. The following reports were created:

- **Report 1**: Number of examinations performed for the 4 modalities (as rows of the query) and for the types of stay in the hospital (inpatient or outpatient, as columns of the query)
- **Report 2**: Number of examinations performed for the 4 modalities (as rows of the query) and for the different hours of the day (from 8:00 a.m. to 7:00 p.m., as columns of the query)
- **Report 3**: Number of examinations requested by all the different organizational units of each hospital (as rows of the query) for the 4 modalities (as columns of the query)

The numbers related to the creation of these reports are shown in Table 3 and are discussed in the next chapter.

One Solr query was created for each possible row-column combination, which is why the number of executed Solr queries for each report equaled the number of rows multiplied by the number of columns. These numbers were identical for all 4 hospitals in the first 2 reports, as they used the same rows and columns. In the last report, 1 row was created for each organizational unit of the hospital, resulting in different numbers of rows for each hospital.

Table 3 also shows the average time in milliseconds Solr needed for each single query, as well as the total time to run all the Solr queries and export their results as an Excel file.

### Table 3. Numbers related to the process of creating 3 reports of different complexities with the data warehouse.

<table>
<thead>
<tr>
<th>Report and details</th>
<th>Hospital A</th>
<th>Hospital B</th>
<th>Hospital C</th>
<th>Hospital D</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Report 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rows, n</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Columns, n</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Solr queries, n</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Time** (ms), mean (SD)</td>
<td>79.5 (111.6)</td>
<td>185.3 (264.0)</td>
<td>90.3 (117.3)</td>
<td>95.3 (180.7)</td>
</tr>
<tr>
<td>Total time</td>
<td>645 ms</td>
<td>1 s 489 ms</td>
<td>728 ms</td>
<td>767 ms</td>
</tr>
<tr>
<td><strong>Report 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rows, n</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Columns, n</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Solr queries, n</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>Time** (ms), mean (SD)</td>
<td>18.3 (7.7)</td>
<td>67.8 (23.3)</td>
<td>22.6 (12.4)</td>
<td>21.8 (13.7)</td>
</tr>
<tr>
<td>Total time</td>
<td>904 ms</td>
<td>3 s 275 ms</td>
<td>1 s 110 ms</td>
<td>1 s 69 ms</td>
</tr>
<tr>
<td><strong>Report 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rows, n</td>
<td>48</td>
<td>804</td>
<td>396</td>
<td>2054</td>
</tr>
<tr>
<td>Columns, n</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Solr queries, n</td>
<td>192</td>
<td>3216</td>
<td>1584</td>
<td>8216</td>
</tr>
<tr>
<td>Time** (ms), mean (SD)</td>
<td>11.6 (4.2)</td>
<td>25.1 (3.2)</td>
<td>9.1 (3.4)</td>
<td>9.8 (3.3)</td>
</tr>
<tr>
<td>Total time</td>
<td>2 s 315 ms</td>
<td>82 s 233 ms</td>
<td>15 s 201 ms</td>
<td>84 s 758 ms</td>
</tr>
</tbody>
</table>

*Average time for the execution of each Solr query.*

**Comparison With the Creation of Reports Directly in Excel**

Before using the developed tool, all 4 hospitals created such reports directly in Excel. To evaluate possible improvements compared to the report creation in Excel, this manual process was performed for new reports of different complexities with the largest data set (of hospital B) with the data of 1 year.

When the reports are created directly in Excel, nothing needs to be imported. Nevertheless, to simplify the calculations on the data, all the RIS exports of the considered year were combined into a single Excel file. The calculations otherwise executed during the import process were performed directly in Excel by using Excel formulas in new columns. As all these calculations were executed on each opening of the Excel file, all the RIS data, together with the calculated values, were then copied to another Excel file so that working with the data was faster.

The reports themselves could be created directly in Excel in many ways. If just single numbers are required, Excel’s built-in filter capabilities can be used. To create the reports for this evaluation, Excel formulas were used to define the value of
each cell. These formulas were then copied to all the other cells, and their restrictions were adapted according to each row and column of the report that had to be created.

The results of this comparison are discussed in the next section.

**User Feedback**

Because the usage of the developed system consisted of various reports requested by the participating hospitals, their feedback was evaluated by describing the requests that could and that could not be created on the data provided by them.

In general, the requests could be divided into those of interest to all hospitals and special requests by an individual hospital. Of general interest was, for example, the number of patients, appointments, and examinations; the use of devices; repeated examinations of the same body region; and the waiting time for an examination. Specialized reports were mainly created for hospital B, which has the largest radiology department among the participating hospitals. The concrete reports that were created for each hospital are listed next.

**Reports for Hospital A**

- Specifically for patients with multiple myeloma or a hepatocellular carcinoma the number of patients for each modality and quarter of the year as well as for each type of stay in the hospital and each clinical department requesting a radiologic examination for such patients has been counted.
- For the same two types of diseases the number of patients with repeated examinations using the same or different modalities was counted.
- For two clinical departments requesting radiologic examinations the time between the request and the availability of the radiologic report has been evaluated.

**Reports for Hospital B**

Each of the following reports was requested for each site of the hospital as well as for regular radiology and neuroradiology:

- The number of patients, appointments, and examinations for each modality was counted. In separate reports, these numbers were further split by each shift and hour during the day or the body regions listed for the next report.
- For repeated examinations using the same or different modalities for the same body region, the number of patients was counted. The body regions of interest for this hospital were the abdomen, the cervical spine, the thoracic spine, the lumbar spine, the ankle joint, the knee joint, the hip joint, the shoulder joint, and the liver. In another report, the total time for these sequences of modalities was evaluated.
- For each modality, the time between the request for an examination and the actual start of it, as well as the availability of the radiologic report, was evaluated. In a similar way, the duration for just the examination itself was evaluated.
- The use of each radiologic device was evaluated by the duration for just the examination itself as well as the duration from the start of an examination until the start of the next examination using the same device.
- For another report, the number of patients at the radiology department for the first time or using each modality for the first time during a year was counted.
- The number of examinations was also counted for patients not older than 18 years and for the following special treatments: osteodensitometry, teleradiology, mammography, and nuclear medicine.
- Specifically for radiologic examinations of the spine, the difference between the actual time a patient stayed in the hospital in comparison to the time recommended by a clinical guideline was evaluated.

**Reports for Hospital C**

- The number of appointments and examinations for each modality was counted for each site of the hospital and each year. These numbers were calculated separately for each hospital department requesting a radiologic examination, for each shift during the day, for each type of stay in the hospital, and for each type of insurance a patient has.
- Only the number of MRI examinations was calculated for each hospital site and year separated by the following body regions: spine, abdomen, upper abdomen, pelvis, small intestine, joints, soft tissues of the neck, hand, foot, chest, skull, shoulder, and heart.

**Reports for Hospital D**

- For each site of the hospital, each year, type of stay in the hospital, and shift, the number of appointments and examinations was counted for each modality.
- Because for this hospital, the names of external private medical practices requesting radiologic examinations were also provided, the number of appointments and examinations requested by each of them was also counted for each hospital site, year, and modality.

With these reports, all the requests of hospitals A and B and some of the requests of hospitals C and D could be fulfilled. As mentioned before, hospitals C and D were not able to provide patient identifiers along with the rest of their exported data, and therefore, no number of patients and no sequences of multiple examinations (as they usually do not occur during the same appointment) could be calculated. The data provided by hospitals C and D also contained no time stamps except the time an examination started, and therefore, no time differences between, for example, the request for an examination and the actual appointment or the availability of the radiologic report, could be evaluated.

However, as long as the hospitals were able to provide the necessary data, all their requests could be fulfilled, and therefore, all of them were satisfied with the developed system.

**Discussion**

**Principal Findings: Used Data**

As shown by the numbers in the previous section, the imported data was diverse, with different numbers of years and attributes. The number of generated values was different as well (depending on the requested reports). However, some hospitals were not able to provide all the data for their requested reports,
such as hospitals titled C and D, which could not (or only with a lot of effort) provide a patient identifier, which resulted in the inability to create any reports with counts of patients. Nevertheless, due to its configurable and modular approach, the developed system can be used for these RIS exports as well, only requiring the creation of a new configuration file as well as some implementations of the interfaces for additional calculations. For hospital B, by far, the maximum number of reports was created, which resulted in the number of generated values exceeding the number of imported ones. One of these calculations, for example, was to count how many patients encountered multiple examinations for the same medical question with the same or different modalities. This directly pointed out multiple cases in which, for example, X-ray examinations had been conducted, followed by a CT or MRI examination, where only a CT or MRI examination would have been necessary, resulting in unnecessary radiation exposure for the patients as well as unnecessary radiology appointments.

Process of Importing the Radiology Data Into the Data Warehouse

When comparing the different numbers related to the import process, we found a correlation between the number of generated and calculated values and the time the developed tool needed to process them. However, even the RIS export with the maximum imported and calculated values (of hospital B) needed only about 1 hour 10 minutes for the whole process, making it fast to use even when installed in a new environment. For most reports, this time is only needed once, and multiple different reports can be created with the system afterward. If adaptations are needed (like for additional calculations during the import process), the mechanism for incremental updates can be used so that the time until the adaptations can be used for reports is even shorter. The storage required for the database and the Solr index together (31 GB for the largest data set of hospital B) can be easily found on many existing systems, and therefore, in most cases, no additional drives need be bought when using this tool. As shown in Table 2, the sizes of the database and Solr index were nearly identical for 3 of the hospitals. The difference between these sizes for hospital C was the reason that the data provided by it as well as the calculated values were mostly Boolean values. Although they are saved in similar form as other types of values in the database, the Solr index does not need to save and process any concrete textual or numeric value for them, resulting in the Solr index being a lot smaller than the database.

Creating Reports on the Radiology Data With the Data Warehouse

For many kinds of reports, the developed system can calculate and export them in a few seconds, as shown in Table 3. This allows a user to quickly iterate and try multiple configurations of a query. By using the preview option of PaDaWaN’s interface, intermediate results do not always have to be exported to Excel or CSV. Even the third report in Table 3 could be created in less than 1.5 minutes for all 4 hospitals, although many single Solr queries were necessary for them. In all these reports, the queries were similar, resulting in the average time for each query becoming shorter with the total number of queries. Another observation from the created reports is that with a larger Solr index (hospital B has the largest one), the average time for each Solr query more than doubles compared to the reports created for the other hospitals but still goes down to 25 ms during the creation of the third (and largest) report.

Comparison With the Creation of Reports Directly in Excel

When comparing the developed tool with report creation directly in Excel, except for the combination of data from multiple Excel files into 1 (which is not necessary for the developed tool, because the RIS data are combined into 1 database and Solr index), no import of data are required, making this step faster and easier in Excel. However, for the calculation of additional values depending on the type of calculation, the required Excel formulas can get quite complex and are therefore more difficult to develop and maintain compared to a calculation written using the Kotlin method. To circumvent this disadvantage, Excel’s ability to add scripts [28] can be used. During the creation of reports, the main disadvantage of using only Excel is the requirement for complex formulas, making the whole report more difficult to create and maintain compared to configuring a query in PaDaWaN’s web interface. Especially the addition/deletion of an attribute to/from any row, column, or filter is easy in PaDaWaN’s GUI, while this requires a user to adapt the Excel formulas in every cell. Therefore, PaDaWaN allows for easy ad hoc adaptation of reports even while discussing them with clinicians. Such ad hoc adaptation also benefits from the fast execution times of most PaDaWaN queries, as explained before. The drawbacks resulting from the use of Excel formulas for creating reports can be partially overcome by using Excel’s feature to create pivot tables. This allows a user, in a similar way to PaDaWaN, to configure rows, columns, and filters of the requested table as well as the kind of numbers (eg, counts of patients or examinations) that should be returned. The disadvantages of this feature are that no logical combinations of attributes for a single column, row, or filter can be specified and would require additional precalculated columns. It also lacks PaDaWaN’s ability for advanced searches on textual data directly as part of the query [18]. Altogether this evaluation showed that although most kinds of reports can be somehow created with Excel, especially more complex queries are difficult to configure and maintain there, while this can be done easily in PaDaWaN’s GUI. A limitation of the developed tool is the requirement for an initial setup and for additional training of the clinicians on how to use it, while Excel is a tool that is already installed in many hospitals and many clinicians are already familiar with its usage.

Comparison With Alternative Solutions

In addition to the creation of reports directly in Excel, other solutions already exist that provide a user with key metrics on medical data. A comparison of the solutions introduced in the State of the Art section is presented in Table 4.
The first 4 solutions provide dashboards for the following uses cases: general imaging use [6], ordered and performed imaging studies for the emergency department [7], scheduled and in-progress examinations in pediatric radiology [8], and various metrics on orders, acquisition, interpretation, and reporting of radiologic images [9]. Although all of them could present their results as graphics, no additional queries (in addition to the ones predefined for the graphics) could be performed. Furthermore, the solutions only work with 1 or multiple specific RISs.

The next 3 solutions work in a different way, as their primary purpose is to display information about individual patients who are currently treated (or about to be treated) in the radiology department [10-12]. The values shown by them can be filtered, for example, by a specific type of examination, but no real queries on these data can be defined by the user. The graphics provided by Henkel et al [11] are limited to single patients and show, for example, the history of 1 of their laboratory values. Munbodh et al [12] also provide predefined graphics for the total number of examinations of different kinds during the past month. Although they all use intermediate storage for all the patients’ data from the hospitals and RISs, no additional calculations on these data can be performed during the import, and of the 3, only the solution provided by Henkel et al [11] is not tied to a specific hospital or RIS.

The next 2 solutions use a data warehouse as a business intelligence tool for the radiology department [13] and combine radiology with pathology data [14]. Although both solutions allow for user-defined queries to be executed, these queries must be specified using Excel or SQL and not via a GUI. They also cannot create graphics from the query results and are tied to a specific RIS. Additionally, no precalculations on the RIS data can be performed and saved in the data warehouse.

The last 2 solutions are importers for i2b2 [15] and for i2b2 as well as PaDaWaN [16] and are not dependent on a specific RIS. They can use all the capabilities of these data warehouse solutions, including the ability for user-defined queries via a GUI and to show some predefined graphics based on these queries. However, because no additional calculations can be performed and saved to the data warehouse during the import, some kinds of reports are difficult or even impossible to create (eg, the evaluation of repeated examinations of the same body region). The developed system has the capability of additional calculations and therefore supports the most diverse kinds of reports.

Verification of the Calculated Metrics

As mentioned in the Introduction section, there are 2 main purposes of calculating all these metrics: From a patient’s perspective, the waiting time for an appointment in the radiology department should be as short as possible and the exposure to radiation should be as low as possible. The hospital, however, wants to maximize its profits. By reducing the time until a diagnosis has been made, the patients can stay in the hospital for fewer days, which therefore allows the hospital to treat more patients. In addition, by eliminating or at least reducing unnecessary repeated examinations of the same region of the body, the radiology department also has the capacity to treat more patients. Treating more patients results in more profit for the hospital and shorter waiting times for patients. The reduction in unnecessary examinations also lowers the patients’ exposure to radiation. Thus, both main purposes of the developed system can be achieved if metrics concerning the waiting times for appointments and diagnosis as well as repeated examinations can be calculated. The way this can be done has already been described before. For a hospital to be able to improve these metrics, it is also important that potential reasons for longer waiting times and repeated examinations be evaluated. Just from the RIS data themselves, this is, for example, possible by comparing the use of different radiologic devices to check whether the purchase of additional devices is necessary. Another possibility is the comparison of different hospital departments that request radiologic examinations. If a metric is significantly worse for one department compared to others, maybe the communication between this department and radiology needs to be improved. However, the developed system is not limited to RIS data alone. By combining these data with other data from the hospital information system in the data warehouse, additional possibilities for improvements can be found. For example, by comparing the diagnosis made by the radiology department with the final clinical diagnosis, the quality of the radiologic diagnosis can be evaluated. In addition, because the data warehouse can store data from multiple years, along with their time stamps, the developed system also supports the verification if any measure taken results in the desired improvement of specific metrics.

Table 4. Comparison of different existing solutions to calculate key metrics of radiology data (or of medical data in general in the last column).

<table>
<thead>
<tr>
<th>Solution</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[6-9]</td>
</tr>
<tr>
<td>Graphical results</td>
<td>Yes</td>
</tr>
<tr>
<td>Graphical query definition</td>
<td>No</td>
</tr>
<tr>
<td>User-defined queries</td>
<td>No</td>
</tr>
<tr>
<td>Additional calculations during import</td>
<td>N/Aa</td>
</tr>
<tr>
<td>Import independent of a specific RISb</td>
<td>No</td>
</tr>
</tbody>
</table>

aN/A: not applicable. Because these solutions operate directly on a radiology information system (RIS), no intermediate storage is used and therefore no additional attributes can be saved to it.

bRIS: radiology information system.
Limitations of and Bias in the Calculated Metrics

As all the metrics calculated by the developed system are based on data from the RIS and the hospital information system, their quality directly depends on the quality of the hospital’s documentation. As this is not verifiable by the developed tool, it could only be assumed that this is done correctly. As mentioned in the previous paragraph, it is also important for the developed system that potential reasons for, say, longer waiting times for an appointment be evaluated. If, for example, a difference between the radiologic diagnosis and the final clinical diagnosis is not documented, it cannot be checked as a potential reason for longer waiting times or unnecessary examinations. In general, inferring conclusions from the calculated metrics can be difficult if potential causes are not documented. For example, if for a medical question, the ideal radiologic examination would be MRI, it may not be conducted, because it is too expensive or enough devices are not available. This could, for example, be further validated by comparing the performed examination with the one recommended by clinical guidelines (like the ones provided by the German Radiation Protection Commission [29]). Another potential bias in the calculated metrics may result from radiologic examinations that are performed externally (eg, at a private medical practice). If not properly documented, this is a missed indicator for the need for additional devices or employees. By combining RIS data with additional data from the hospital in the data warehouse, the text search capabilities of the data warehouse can potentially be used to find information about such external examinations in the patient’s discharge letter. Additionally, the interpretation of the metrics could be different among different hospitals, resulting in limitations of the comparability of the calculated metrics between them. Although, for example, a waiting time of a few days for an appointment in the radiology department could be acceptable for one hospital, it could be unacceptable for another one.

Conclusion

To summarize, the developed system has its main advantages in being generic concerning the export of different RISs as well as concerning the configuration of queries for various reports. To use it, the only requirement is the ability of an RIS to create an Excel, CSV, or JSON export. This can then be imported by creating a simple configuration file in Excel or as a CSV file as well. During the import process, additional values can be calculated by implementing several provided interfaces. If further values should be added later, this is easily possible with the ability to use incremental updates. Various reports with any combination of and restriction on the imported attributes can then be graphically configured using PaDaWaN’s web interface. Finally, the results of these reports can be exported into the standard formats Excel and CSV so that they can be easily processed with many different tools.

The whole tool as a Docker [30] image, a sample RIS export, and a configuration file are publicly available on PaDaWaN’s website [31].

The developed tool can in the future be further enhanced by, for example, adding the ability to calculate other numbers than the count of patients, appointments, or examinations, such as the average of numeric values found by a query. To improve the presentation of the results, the current ability to create graphics for some predefined reports can also be extended to be configurable by a user and therefore allow for the creation of many kinds of graphical reports.

Conflicts of Interest

None declared.

References


23. Learn more about Solr. Apache Software Foundation. URL: https://solr.apache.org [accessed 2022-08-09]


31. Download – Chair of Computer Science VI – artificial intelligence and applied computer science. Institut für Informatik - Universität Würzburg. URL: https://www.informatik.uni-wuerzburg.de/en/is/research/padan-data-query-tool/download/ [accessed 2022-08-09]

**Abbreviations**

**CT:** computed tomography  
**JSON:** JavaScript Object Notation  
**GUI:** graphical user interface  
**MRI:** magnetic resonance imaging  
**MXQL:** Medical XML Query Language  
**PaDaWaN:** Patient Data Warehouse Navigator  
**RIS:** radiology information system
Assessment and Improvement of Drug Data Structuredness From Electronic Health Records: Algorithm Development and Validation

Ines Reinecke1, BA, MA; Joscha Siebel1; Saskia Fuhrmann2,3, Dr Rer Nat; Andreas Fischer3, MA; Martin Sedlmayr2, Prof Dr, Dr Rer Nat; Jens Weidner1, MA; Franziska Bathelt1, Dr Rer Nat

1Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany
2Center for Evidence-Based Healthcare, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany
3Hospital Pharmacy, University Hospital Carl Gustav Carus, Dresden, Germany

Corresponding Author:
Ines Reinecke, BA, MA
Institute for Medical Informatics and Biometry
Carl Gustav Carus Faculty of Medicine
Technische Universität Dresden
Fetscherstraße 74
Dresden, 01307
Germany
Phone: 49 35145887 ext 12975
Email: ines.reinecke@uniklinikum-dresden.de

Abstract

Background: Digitization offers a multitude of opportunities to gain insights into current diagnostics and therapies from retrospective data. In this context, real-world data and their accessibility are of increasing importance to support unbiased and reliable research on big data. However, routinely collected data are not readily usable for research owing to the unstructured nature of health care systems and a lack of interoperability between these systems. This challenge is evident in drug data.

Objective: This study aimed to present an approach that identifies and increases the structuredness of drug data while ensuring standardization according to Anatomical Therapeutic Chemical (ATC) classification.

Methods: Our approach was based on available drug prescriptions and a drug catalog and consisted of 4 steps. First, we performed an initial analysis of the structuredness of local drug data to define a point of comparison for the effectiveness of the overall approach. Second, we applied 3 algorithms to unstructured data that translated text into ATC codes based on string comparisons in terms of ingredients and product names and performed similarity comparisons based on Levenshtein distance. Third, we validated the results of the 3 algorithms with expert knowledge based on the 1000 most frequently used prescription texts. Fourth, we performed a final validation to determine the increased degree of structuredness.

Results: Initially, 47.73% (n=843,980) of 1,768,153 drug prescriptions were classified as structured. With the application of the 3 algorithms, we were able to increase the degree of structuredness to 85.18% (n=1,506,059) based on the 1000 most frequent medication prescriptions. In this regard, the combination of algorithms 1, 2, and 3 resulted in a correctness level of 100% (with 57,264 ATC codes identified), algorithms 1 and 3 resulted in 99.6% (with 152,404 codes identified), and algorithms 1 and 2 resulted in 95.9% (with 39,472 codes identified).

Conclusions: As shown in the first analysis steps of our approach, the availability of a product catalog to select during the documentation process is not sufficient to generate structured data. Our 4-step approach reduces the problems and reliably increases the structuredness automatically. Similarity matching shows promising results, particularly for entries with no connection to a product catalog. However, further enhancement of the correctness of such a similarity matching algorithm needs to be investigated in future work.

(JMIR Med Inform 2023;11:e40312) doi:10.2196/40312

KEYWORDS

secondary usage; Observational Medical Outcomes Partnership; OMOP; drug data; data quality; Anatomical Therapeutic Chemical; ATC; RxNorm; interoperability
Introduction

Background
Over the last decade, the amount of electronically available data in the health care domain has increased enormously worldwide. Much of the data is generated during the processing of administrative claims, through documentation processes in electronic health records (EHRs) performed during patient treatments, or via data feeds from mobile devices providing patient-reported outcomes. Therefore, it is not surprising that real-world data (RWD) are becoming more important for health care research. RWD studies can be considered complementary to randomized controlled trials (RCTs), as they allow the results of RCTs to be confirmed in much larger cohorts and over a longer period. Compared with RCTs, RWD studies allow for better external validity and better generalizability, and they not only offer opportunities for long-term surveillance of drug products but also are cost-effective and time-saving [1].

Drug surveillance systems, such as the US Food and Drug Administration’s Sentinel initiative, are critical for promoting postmarket drug safety [2-8]. The European Medicines Agency has also started to establish research infrastructure based on RWD to support pharmacovigilance [9]. In addition, the European Health Data and Evidence Network [10] emerged to establish transnational research networks based on a common data model that enables standardized RWD and methods for observational studies to generate real-world evidence. Recently, the European Health Data and Evidence Network has begun to collaborate with the European Medicines Agency to address COVID-19 [11].

However, the original purpose of RWD generation during patient treatment is not primarily aimed at its use in research. Therefore, notable problems have been identified regarding the replication and validity of observational research results based on RWD. To ensure the reliability and robustness of the results from RWD research, these issues have to be addressed, as they become even more important when observational studies are conducted across countries at large scale.

Data harmonization, the use of international standards and terminologies, a common data model, methods, and tools for data analyses that increase the reproducibility of results are needed [12]. These gaps are being addressed by the International Observational Health Data Sciences and Informatics community, which provides the common data model called the Observational Medical Outcomes Partnership (OMOP) and standardized analysis tools based on OMOP. It also includes standardized vocabularies that contain translations between national terminologies and internationally acknowledged terminologies, for example, Systematized Nomenclature of Medicine-Clinical Terms, Logical Observation Identifiers Names and Codes, Anatomical Therapeutic Chemical (ATC) classification, and RxNorm [13]. OMOP allows RWD to be stored in the same way, regardless of data origin, thus ensuring the use of RWD in international, large-scale observational studies. Compared with similar projects such as Informatics for Integration Biology and the Bedside or the National Patient-Centered Clinical Research Network, Observational Health Data Sciences and Informatics-OMOP meets the needs of observational RWD studies well [14]. Many RWD studies on OMOP have shown that drug data at the ingredient level are sufficient to answer their research questions [15]. Although drug data with details on dosage and units for drug exposure can be important for observational research on drug effectiveness and drug safety with the same drug at different doses, the availability of the drug ingredient is the least common denominator and the basic requirement for drug-related RWD studies on OMOP. Therefore, drug prescription data must be available in a structured format that does not necessarily include the name of the drug product but at least the ingredient information. For drug utilization research, the World Health Organization recommends the use of ATC classification, which divides drugs into different groups based on the organ or system on which they act [16]. ATC classification includes a hierarchy based on 5 different levels, with ATC level 5 being the chemical substance that represents the active ingredient of a drug product [17]. Each approved drug product on the market is assigned a specific ATC level 5 code. The National Institutes of Health Collaboratory recommends assessing and reporting the quality of EHR data for clinical reuse in terms of data completeness, accuracy, and consistency [18]. Weiskopf et al [19] also determined that the completeness and correctness of data are of special importance for data quality improvement.

Objective
To the best of our knowledge, there is no existing approach to systematically analyze and improve the structuredness of drug prescription data for observational research. Thus, in this study, we systematically analyzed the structuredness of EHR drug prescriptions to determine the ratio between structured drug prescription data containing ATC code level 5 and free-text drug prescriptions without an available standard concept based on the 14 ATC groups of level 1. In addition, we presented an approach to improve the structuredness of drug prescription data by introducing an automatic detection method for ATC code determination. To ensure the robustness and accuracy of the results of automatic detection, we introduced a validation step based on existing text-mining algorithms.

Methods
Study Details
This retrospective, noninterventional study systematically reviewed drug prescriptions based on real-world observational data at the University Hospital Carl Gustav Carus Dresden (UKD), Germany. This study was based on fully anonymized data and did not include any correlations with individual patients. All inpatient drug prescriptions, including acute medications, from 2016 to 2020 were included in the study, without restriction to specific conditions or treatments. The original data were recorded in the ORBIS hospital information system from Dedalus, using the ORBIS module, “KURV,” that represents the patient curve including medication data. A total of 1,768,153 drug prescriptions were reviewed from the hospital information system records. Drug prescription data from other systems (eg, intensive care units and chemotherapy) were excluded because data in those systems were completely
structured and stored in separate backend systems. The data used for this study were provided by the Data Integration Centre at the UKD, which was established with funding from the German Federal Ministry of Education and Research as part of the Medical Informatics Initiative in Germany.

Ethics Approval

The study was approved by the Ethics Committee of the Technical University of Dresden as a retrospective, observational, noninterventional, nonhuman subject study (SR-EK-521112021).

Data Set Details

The following 2 data sets were used: drug prescription data (data set 1) and drug product catalog data (data set 2). The drug product catalog was exported from the UKD Enterprise-Resource-Planning system on November 16, 2021, and contained the drug product name, drug ingredient name, ATC level 5 code, drug dose and unit information, and legacy products. In addition, 2 other data sets were derived from data set 1. First, an aggregated data set (data set 3) was generated based on the grouped data set 1 for all the unstructured drug prescription entries. The grouping activity to create data set 3 was performed on the MEDICATION column of data set 1 by grouping all entries in the data element MEDICATION using the Python library Pandas and its groupby function. Frequency information for each unique MEDICATION record was added to data set 3. The data set (data set 4) contained a subset of the first 1000 most frequent entries from data set 3 and additional results from the manual evaluation step.

All the metadata elements of the data sets presented above that are relevant to this study are illustrated and described in detail in Table 1. Drug prescriptions selected from the drug catalog data are labeled as structured data (eg, “IBUPROFEN STADA 600 mg Zäpfchen | [Ibuprofen natrium, Ibuprofen]”) based on the contents of the STRUCTURE column of data set 1. Drug prescriptions that were not selected from the drug product catalog are designated as unstructured data (eg, “Ibuprofen 600” and “Ibuprofen”).

https://medinform.jmir.org/2023/1/e40312

JMIR Med Inform 2023 | vol. 11 | e40312 | p.528

(page number not for citation purposes)
Table 1. Description of relevant data set with its metadata elements.

<table>
<thead>
<tr>
<th>Data set and data element</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DS</strong>1 initial data set with all drug prescriptions<strong>a</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDICATION</td>
<td>String</td>
<td>Free text or predefined value, chosen from an available fixed drop-down menu that contains product names, derived from the drug product catalog, when creating a new drug prescription</td>
</tr>
<tr>
<td>YEAR</td>
<td>Number</td>
<td>Extracted from the prescription start date information for further statistical analyses</td>
</tr>
<tr>
<td>STRUCTURE</td>
<td>Boolean</td>
<td>TRUE if MEDICATION was chosen from the drug catalog or FALSE if the free text was entered</td>
</tr>
<tr>
<td>ATC&lt;sub&gt;b&lt;/sub&gt; L5</td>
<td>String</td>
<td>ATC code level 5 available in case STRUCTURE is true otherwise empty</td>
</tr>
<tr>
<td><strong>DS2 drug catalog data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Product_name</td>
<td>String</td>
<td>Product name as listed in the ERP&lt;sup&gt;c&lt;/sup&gt; system</td>
</tr>
<tr>
<td>Ingredient_name</td>
<td>String</td>
<td>Ingredient name as listed for the product</td>
</tr>
<tr>
<td>Atc_code</td>
<td>String</td>
<td>ATC code level 5</td>
</tr>
<tr>
<td><strong>DS3 grouped DS1 by MEDICATION data element</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDICATION</td>
<td>String</td>
<td>Grouped unstructured free-text entries</td>
</tr>
<tr>
<td>FREQUENCY</td>
<td>Number</td>
<td>Summed up the occurrence of the MEDICATION text field to determine the most relevant free-text drug prescriptions</td>
</tr>
<tr>
<td>Step1</td>
<td>String</td>
<td>Algorithm 1 result as an ATC code or empty if no match</td>
</tr>
<tr>
<td>Step2</td>
<td>String</td>
<td>Algorithm 2 result as an ATC code or empty if no match</td>
</tr>
<tr>
<td>Step3</td>
<td>String</td>
<td>Algorithm 3 result as an ATC code or empty if no match</td>
</tr>
<tr>
<td><strong>DS4 most frequent 1000 entries of DS3 (sorted by frequency)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDICATION</td>
<td>String</td>
<td>Grouped unstructured free-text entries</td>
</tr>
<tr>
<td>FREQUENCY</td>
<td>Number</td>
<td>Summed up the occurrence of the medication text field</td>
</tr>
<tr>
<td>Step1</td>
<td>String</td>
<td>Algorithm 1 result as an ATC code or empty if no match</td>
</tr>
<tr>
<td>Step2</td>
<td>String</td>
<td>Algorithm 2 result as an ATC code or empty if no match</td>
</tr>
<tr>
<td>Step3</td>
<td>String</td>
<td>Algorithm 3 result as an ATC code</td>
</tr>
<tr>
<td>Eval1</td>
<td>Boolean</td>
<td>Algorithm 1 evaluation result</td>
</tr>
<tr>
<td>Eval2</td>
<td>Boolean</td>
<td>Algorithm 2 evaluation result</td>
</tr>
<tr>
<td>Eval3</td>
<td>Boolean</td>
<td>Algorithm 3 evaluation result</td>
</tr>
<tr>
<td>True12</td>
<td>Boolean</td>
<td>TRUE if the same result for algorithm 1+2</td>
</tr>
<tr>
<td>True13</td>
<td>Boolean</td>
<td>TRUE if the same result for algorithm 1+3</td>
</tr>
<tr>
<td>True23</td>
<td>Boolean</td>
<td>TRUE if the same result for algorithm 2+3</td>
</tr>
<tr>
<td>True123</td>
<td>Boolean</td>
<td>TRUE if the same result for algorithm 1+2+3</td>
</tr>
<tr>
<td>CORRECT</td>
<td>String</td>
<td>Corrected ATC code. In case no algorithm determined the correct result, entered manually in the evaluation step</td>
</tr>
<tr>
<td>COMMENTS</td>
<td>String</td>
<td>Any comments or additional information if needed</td>
</tr>
<tr>
<td>FINAL</td>
<td>String</td>
<td>Finally determined ATC code for all entries or labels in case no ATC code could be determined (labels are introduced in the methods validation section in detail)</td>
</tr>
</tbody>
</table>

<sup>a</sup>DS: data set.  
<sup>b</sup>ATC: Anatomical Therapeutic Chemical.  
<sup>c</sup>ERP: Enterprise-Resource-Planning.

**Data Analysis**

**Overview**

Data analysis consisted of a 4-step process, as shown in Figure 1. The first step of the process was an initial data quality analysis to determine the overall ratio of structured to unstructured drug prescriptions.

To improve the structure of the drug prescriptions, 3 existing algorithms were applied to automatically identify correct ATC codes for the unstructured drug prescriptions. The identified ATC codes were then manually reviewed by experts.
(pharmacists and medical information scientists) and checked for correctness. This step also included the identification of existing patterns that can help conclude the reliability of the automatically identified ATC codes for unstructured data. Finally, the results of the previous 3 steps were consolidated to assess the degree of improvement achieved in unstructured drug prescriptions. To ensure expert coverage of the entire process, an interdisciplinary team of pharmacists, computer scientists, and medical informatics researchers was formed.

**Figure 1.** Data analysis 4-step approach. ATC: Anatomical Therapeutic Chemical; NLP: natural language processing.

**Initial Data Assessment**
Initially, the ratio of structured to unstructured drug prescriptions was determined for data set 1. For this purpose, the STRUCTURE data element of data set 1 was used to subdivide the data into 2 groups. If the value of STRUCTURE was TRUE, the record was considered to be structured; otherwise, it was unstructured. Subsequently, the unstructured subset of drug prescriptions was grouped by the data element MEDICATION as data set 3, and the frequency was calculated and added as the data element FREQUENCY.

The first manual review of the grouped drug prescriptions (data set 3) was done by the interdisciplinary team of experts to identify records that are not drug prescriptions but other instructions, such as orders for blood counts or other laboratory and measurement orders (eg, “BGA”—laboratory request for blood gas analysis, “BE”—request to nurses for taking a blood sample, and “BB”—laboratory request for blood count). This task resulted in a set of rules (Multimedia Appendix 1) to allow the automated search and identification of medication entries that needed to be excluded for further steps.

**Improvement**
Unstructured drug prescriptions (usually provided as free text) were used as inputs for the improvement step. Preprocessing of the drug prescription was not performed previously. In this step, 3 different algorithms were implemented to automatically identify ATC level 5 codes based on the MEDICATION text.

The algorithms were based on a different mechanism for matching the MEDICATION text of data set 1 with the product catalog data elements INGREDIENT_NAME and PRODUCT_NAME of data set 2, as described in detail in Table 2.

**Table 2.** An overview of algorithms for Anatomical Therapeutic Chemical (ATC) code identification for unstructured drug prescriptions.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Mechanism</th>
<th>Data input for comparison</th>
<th>Result data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>String comparison</td>
<td>MEDICATION, Ingredient_name</td>
<td>ATC code</td>
</tr>
<tr>
<td>2</td>
<td>String comparison</td>
<td>MEDICATION, Product_name</td>
<td>ATC code</td>
</tr>
<tr>
<td>3</td>
<td>Similarity matching</td>
<td>MEDICATION, INGREDIENT_NAME, and PRODUCT_NAME</td>
<td>ATC code + similarity score</td>
</tr>
</tbody>
</table>
Algorithms 1 and 2 rely on simple string comparisons to recognize either the ingredient name or product name within the drug prescription. Algorithm 3 performs natural language processing (NLP) based on similarity matching between the data element MEDICATION in data set 1 and the 2 data elements PRODUCT_NAME and INGREDIENT in data set 2 with the Python library FuzzyWuzzy [20] using Levenshtein distance because it has shown promising results in other health care research areas [21,22]. The best similarity score result was 100, which meant that the components of the string MEDICATION were entirely contained in INGREDIENT_NAME or PRODUCT_NAME. The lower the similarity score, the less similar the MEDICATION string is compared with the drug catalog entries. This algorithm provided up to 3 possible ATC codes, sorted in descending order based on their similarity scores. To determine the most promising method of the FuzzyWuzzy library for our implementation, we defined that the word order in the data element MEDICATION is irrelevant and can be different from the compared strings in INGREDIENT_NAME and PRODUCT_NAME. All words from the entry of the data element MEDICATION must be included in the entry of INGREDIENT_NAME or PRODUCT_NAME, but not vice versa. This led to the implementation of the method token_set_ratio. This method tokenizes both strings to be compared, changes the upper case to the lower case, and removes punctuation. It then sorts the tokens alphabetically and split them into 2 groups: the intersection group (tokens that are the same in both strings) and the remainder group (tokens that differ in compared strings). The token_set_ratio method compares the intersection group with the intersection and remainder of the first string and then the same with the remainder of the other string and finally takes the highest results of this comparison as the final result. As shown in the following example (Textbox 1), the token_set_ratio method provides the best results concerning the given requirements.

**Textbox 1.** An example of the token_set_ratio method.

```python
import fuzzy

d1 = "Stada paracetamol"
d2 = "paracetamol Stada 400 mg"

Print("Token Set Ratio: ", fuzz.token_set_ratio(d1.lower(),d2.lower()))
```

The algorithms were applied to data sets 1 and 3. The results of the algorithms in data set 3 were also used in data set 4. The concordance between the results for each permutation (algorithms 1+2, 1+3, 2+3, and 1+2+3) was also calculated. The complete source can be accessed on Zenodo [23].

**Validation**

The validation step consisted of manual checks of the automatically generated ATC codes by the same interdisciplinary team as in the previous steps. It was performed on a subset of the most common free-text prescriptions. To maintain the validation effort proportionate to the benefit, a minimum target was defined for the manual validation process of unstructured drug prescriptions to cover at least 80% of structured and manually validated unstructured entries combined. During the validation step, information was added to each algorithm to determine whether the correct ATC code, wrong ATC code, or no ATC code was identified. If no algorithm identified the correct ATC code, it was determined by manual validation when possible. If an entry was found to generally have no drug prescription, it was marked as an additional entry without drug prescription with the keyword “nomed.” For drug prescriptions that require further specification to determine the exact ATC level 5 code, the manual review checks whether the ATC level 4 or 3 code can be determined based on the free text of the drug prescription, otherwise the entry was flagged as unspecified with the keyword “unspec.” All unstructured drug prescription entries for which no validation of the automatically generated ATC codes was performed were marked with the keyword “no_eval.”

The results of the manual validation were summarized to identify any patterns that can help improve the robustness of the results of automatically detected ATC codes based on the total findings and correctness of each algorithm, the concordance level between the results of algorithms 1, 2, and 3, and the Levenshtein similarity score for algorithm 3. For algorithm 3, we used a 2-tailed t test implemented in Python to determine whether there was a significant difference between the means of the Levenshtein similarity score for the correct and incorrect results.

In addition, the incorrect results were examined in more detail by the interdisciplinary team to identify patterns that would reveal important reasons and similarities related to the ingredients of concern (ATC) to the greatest extent possible.

**Final Data Assessment**

For the final data assessment, the results of the step improvement and validation documented in data set 4, including correctly identified ATC level 5 codes or “nomed,” “unspec,” or
“no_eval” labels, were merged with the original drug prescription data from data set 1. Thus, the final data assessment was executed based on the algorithm results and manual validation. The total number of drug prescription records was determined for each of the 14 ATC groups, including the proportion of structured versus unstructured data per ATC group. In addition, the total number of unique ATC level 5 codes, including their structuredness, as well as the most frequent ATC level 5 codes used in drug prescription of the data set 1 are presented. This allows a ranking of the structuredness based on the ATC groups and ATC codes.

Results

Initial Data Assessment
The initial assessment revealed 843,980 (n=1,768,153, 47.73%) drug prescriptions in the data set 1) structured drug prescriptions. The proportion of unstructured drug prescriptions that required further investigation was 52.27% (924,173 drug prescription entries). A small set of rules, for example, all drug prescription entries starting with laboratory or measurement orders (Multimedia Appendix 1), identified a total of 160,896 (9.1% of all drug prescriptions) entries as no drug prescription data and reduced the unstructured drug prescriptions requiring review for the next steps to 763,277 (43.17% of all drug prescriptions).

Improvement
The quantitative performance of the algorithms was very different, as each algorithm returned a different number of results. Algorithm 3 provides an ATC code for all unstructured drug prescriptions, owing to its implementation and nature. Algorithm 1 (based on ingredient matching) identified ATC codes for 8048 unique free texts. Multiplied by the frequency of each text entry, this yielded a total of 244,718 (32.06%) drug prescriptions of the total 763,277 unstructured drug prescriptions. The quantitative outcome performance of algorithm 2 (based on the drug product) is lower than that of algorithm 1, as it identified ATC codes for 6744 unique free texts. This represents a total of 126,100 (16.52%) drug prescriptions of the total 763,277 unstructured drug prescriptions. At this point, no statement can be made about the correctness of the algorithm results, but the analysis of the match rate between all algorithms shows matching rates for the total number of unstructured drug prescriptions and the most frequent 1000 free-text entries as illustrated in Figure 2.

Validation
The manual validation step was performed on the most frequent 1000 free-text entries, which already covered 66.56% (615,129/924,173) of all unstructured drug prescriptions, as shown in Figure 3. Together with the proportion of structured drug prescriptions (843,980/1,768,153, 47.73%) and entries without medication (166,307/1,768,153, 9.4%) that were

Figure 2. Match rates of algorithm results calculated for all data sets under inspection.

![Figure 2](https://medinform.jmir.org/2023/1/e40312)

Figure 3. Manual validation on the most frequent 1000 free-text entries.
identified during the initial data analysis, the structuredness could potentially be increased to 85.18% (1,506,059/1,768,153) of all medication prescriptions.

For the most frequent 1000 free-text entries (data set 4), algorithm 1 returned 286 (28.6%) correct results, 1 (0.1%) incorrect result, and no results for 713 (71.3%) entries. Algorithm 2 returned 142 (14.2%) correct results, 6 (0.6%) incorrect results, and no results for 852 (85.2%) unique entries. Algorithm 3 returned 765 (76.5%) correct results and 235 (23.5%) incorrect results. We also determined the correctness in terms of the result match rates between the algorithms, as shown in Figure 2, for data set 4. After the manual validation of data set 4, the returned ATC codes were always correct if all algorithms or algorithms 1 and 2 returned the same results.

Figure 3. Percentage of the most frequent 1000 of all unstructured drug prescriptions.

For the matching results of algorithms 1 and 3, we noted a minor discrepancy and identified 5 incorrect results out of 286, which were related to sodium chloride drug prescriptions in 4 out of 5 cases, and another incorrect result was returned for the ingredient aciclovir. The manual review revealed that an ATC code could not be provided because of missing details that was due to the ATC code at ATC level 1 varying by route of administration (eg, oral, parenteral, and conjunctival). For the matching results of algorithms 2 and 3, we identified only 1 incorrect result related to the ingredient telmisartan because the drug prescription was for a combination drug (telmisartan and diuretics), whereas the algorithms identified only the ATC code of the single ingredient telmisartan.

For the data set 4, a significant difference in the means of the Levenshtein similarity score was found between the correct and incorrect results (see Table 3 for descriptive statistics) with a $P$ value of $2.4 \times 10^{-47}$, which is well below the significance level $\alpha (.05)$. This means that the higher the Levenshtein similarity score, the higher the probability of result correctness. Speaking in terms of absolute numbers, for entries with a Levenshtein similarity score >84.28, the results can be assumed to be correct with a low error rate.

| Table 3. Descriptive statistics for Levenshtein similarity score for correct and wrong results. |
|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| **Descriptive statistics** | **Algorithm 3** | **Wrong** |
| **Count** | 766 | 234 |
| **Frequency, n (%)** | 416,585 | 67.18 (14.86) |
| **Values, mean (SD)** | 84.28 (15.52) | 84.598 |
| **Range (%)** | 21-100 | 29-100 |
| **Percentile** | | |
| 25th percentile | 76 | 55 |
| 50th percentile | 87 | 63 |
| 75th percentile | 96 | 75 |
The 234 incorrect results returned only by algorithm 3, with 4.78% (84,598/1,768,153) of drug prescriptions, can be categorized into four groups: (1) manually identified additional entries without drug prescriptions for which the applied rules did not work; (2) specification generally not possible owing to missing information; (3) restriction to ATC level 3 or 4 because of nonspecific drug prescription information; and (4) other reasons. We found 16 entries (multiplied by frequency=5411) with no additional drug prescription entries. For an additional 11 (multiplied by frequency=2187) entries, no ATC code could be provided because the dosage form or dose was missing. For 2 drug prescriptions (multiplied by frequency=2610) of insulin therapies, there was a restriction to ATC level 3. Another 2 (multiplied by frequency=887) entries were restricted to ATC level 4 sodium chloride prescriptions.

For the other 203 misidentified entries, we examined the subset of 26 entries where algorithm 3 returned results with a Levenshtein similarity value of ≥80 because it is an indication of correctness but unfortunately did not apply to all results. A small group of 26 results with a Levenshtein similarity value of ≥80 was still incorrect. The main reason for the errors in these results was that the ATC codes for the ingredients differed by dosage form and when combined in a drug product, as shown in Table 4. Most incorrect results (15 out of 26) were caused by the absence of the ingredient dosage form in the free text, especially for sodium chloride, prednisolone, dimetindene, aciclovir, and hydrocortisone. The full data set 4 with all the algorithm outcome quality data elements listed in Table 1 is available in Multimedia Appendix 2.

### Table 4. Wrong results of algorithm 3 with Levenshtein similarity score ≥80.

<table>
<thead>
<tr>
<th>Medication free text</th>
<th>Wrong result</th>
<th>Levenshtein similarity score</th>
<th>Correct result</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASS RATIOPHARM 100 mg TAH Tabletten[(Acetylsalicylsäure)]</td>
<td>N02BA01</td>
<td>89</td>
<td>B01AC06</td>
<td>Similarity of words</td>
</tr>
<tr>
<td>Prednisolon</td>
<td>S01CA53</td>
<td>100</td>
<td>H02AB06</td>
<td>Dosage form</td>
</tr>
<tr>
<td>MAGNESIUM VERLA 300 Orange Granulat[(Magnesium-Ion)]</td>
<td>A12CC05</td>
<td>100</td>
<td>V06XX02</td>
<td>Similarity of words</td>
</tr>
<tr>
<td>ARILIN 500 Filmtablatten</td>
<td>(Metronidazol)</td>
<td>G01AF01 100</td>
<td>P01AB01</td>
<td>Similarity of words</td>
</tr>
<tr>
<td>Candesartan HEXAL, comp 16 mg/12.5 mg Tabletten</td>
<td><em>(Candesartan)</em></td>
<td>C09CA06 89</td>
<td>C09DA26</td>
<td>Combination product</td>
</tr>
<tr>
<td>Heparin</td>
<td>C05BA03</td>
<td>100</td>
<td>B01AB01</td>
<td>Dosage form</td>
</tr>
<tr>
<td>PREDNISOLON</td>
<td>S01CA53</td>
<td>100</td>
<td>H02AB06</td>
<td>Dosage form</td>
</tr>
<tr>
<td>FENISTIL Injektionslösung</td>
<td>(Dimetinden)</td>
<td>D04AA13 100</td>
<td>R06AB03</td>
<td>Dosage form</td>
</tr>
<tr>
<td>ACIC 250 PI Via Pulver z.Herst.e.Infusionsl.</td>
<td><em>(Aliclo vir)</em></td>
<td>D06BB03 100</td>
<td>J05AB01</td>
<td>Dosage form</td>
</tr>
<tr>
<td>NaCl 0.9%</td>
<td>B05CB01</td>
<td>100</td>
<td>B05BB11</td>
<td>Dosage form</td>
</tr>
<tr>
<td>VALSARTAN HEXAL, comp.160mg/12.5mg Filmtablatten</td>
<td><em>(valsartan)</em></td>
<td>C09CA03 100</td>
<td>C09DA23</td>
<td>Combination product</td>
</tr>
<tr>
<td>Prednisolon mg</td>
<td>S01CA53</td>
<td>88</td>
<td>H02AB06</td>
<td>Dosage form</td>
</tr>
<tr>
<td>NaCl 0.9%</td>
<td>B05CB01</td>
<td>100</td>
<td>B05BB11</td>
<td>Dosage form</td>
</tr>
<tr>
<td>ACIC 200 Tablettren</td>
<td><em>(Aliclo vir)</em></td>
<td>D06BB03 100</td>
<td>J05AB01</td>
<td>Dosage form</td>
</tr>
<tr>
<td>ACIC 500 PI Via Pulver z.Herst.e.Infusionsl.</td>
<td><em>(Aliclo vir)</em></td>
<td>D06BB03 100</td>
<td>J05AB01</td>
<td>Dosage form</td>
</tr>
<tr>
<td>Simvastatin</td>
<td>C10BA02</td>
<td>100</td>
<td>C10AA01</td>
<td>No combination product</td>
</tr>
<tr>
<td>Candesartan HEXAL, comp 8 mg/12.5 mg Tabletten</td>
<td><em>(Candesartan)</em></td>
<td>C09CA06 89</td>
<td>C09DA26</td>
<td>Combination product</td>
</tr>
<tr>
<td>NaCl 0.9%</td>
<td><em>(Natrium-Ion, Chlorid)</em></td>
<td>B05CB01 100</td>
<td>B05BB11</td>
<td>Dosage form</td>
</tr>
<tr>
<td>C) FENISTIL 1 Ampulle als Bolus</td>
<td><em>(Dimetinden)</em></td>
<td>D04AA13 100</td>
<td>R06AB03</td>
<td>Dosage form</td>
</tr>
<tr>
<td>HCT</td>
<td>C09DX01</td>
<td>100</td>
<td>C03AA03</td>
<td>Shortness of text</td>
</tr>
<tr>
<td>Allopurinol</td>
<td>M04AA51</td>
<td>100</td>
<td>M04AA01</td>
<td>Combination product</td>
</tr>
<tr>
<td>Prednisolon 5 mg</td>
<td>S01CA53</td>
<td>81</td>
<td>H02AB06</td>
<td>Dosage form</td>
</tr>
<tr>
<td>HYDROCORTISON 10 mg Jenapharm Tabletten</td>
<td><em>(Hydrocortison)</em></td>
<td>S01BA02 81</td>
<td>H02AB09</td>
<td>Dosage form</td>
</tr>
<tr>
<td>Simvastatin 20 mg</td>
<td>C10BA02</td>
<td>100</td>
<td>C10AA01</td>
<td>No combination product</td>
</tr>
<tr>
<td>NaCl 0.9%</td>
<td>B05CB01</td>
<td>100</td>
<td>B05BB11</td>
<td>Dosage form</td>
</tr>
</tbody>
</table>
Final Data Assessment

Compared with the initial data assessment in which we could only distinguish between structured and unstructured drug prescriptions, we were able to perform a percentage distribution between structured and unstructured prescriptions for each of the 14 ATC level 1 groups after applying the algorithm. The final results are presented in Table 5, which shows the number of structured drug prescriptions versus the number of unstructured drug prescriptions per ATC level 1 group. The total number of drug prescriptions per ATC level 1 group, including percentages, provides an overview of the most and least frequently prescribed drugs sorted by the 14 ATC level 1 groups. For completeness, we added 3 additional rows to Table 5 containing the number of unstructured entries identified as other orders (no_med), the number of unstructured entries identified as unspecified entries (unspec), and the remaining unstructured data for which no validation was performed; thus, no statement on the correct ATC code was possible. ATC level 1 group “N – Nervous system” was the most common group with 24.1% (322,286/1,337,565) of the initial data set 1, followed by “B – Blood and blood forming organs,” “A – Alimentary tract and metabolism,” and “C – Cardiovascular system” with approximately 19% each.

Figure 4 illustrates the structuredness of the data for each of the 14 ATC level 1 groups. The ATC level 1 group with the most structured data was “S – Sensory organs” with 98.03% (5077/5179) structured data, followed by the group “H – Systemic hormonal drugs, excluding sex hormones and insulins” with 79.9% (51,296/64,199) structured data. ATC level 1 groups “R – Respiratory system,” “C – Cardiovascular system,” “J – Anti-infective for systemic use,” “V – Various,” “B – Blood and blood forming organs,” and “N – Nervous system,” ranged from 61% to 70% structured data. ATC group “P – Antiparasitic products, insecticides, and repellents” had the lowest percentage of structured drug prescriptions at only 23.4% (342/1461). In total, 742 ATC level 5 codes (ingredients) were identified in the drug prescription data. The structuredness of the ingredients varied widely, showing a wide range of structuredness among ingredients, as shown in Figure 5, where each of the 742 ATC level 5 codes (ingredients) is represented by a single dot. The y-axis represents the degree of the structuredness between 0% and 100%. The x-axis represents the frequency of each ATC level 5 code in data set 1, with a limitation of 85.18% (1,506,059/1,768,153) structured and evaluated unstructured data. There were only 4 ATC level 5 codes that were each used more than 45,000 times in drug prescriptions, namely N02BB02 (metamizole), B05BB01 (sodium chloride), A02BC02 (pantoprazole), and N02AA05 (oxycodone).

Table 5. Number of Anatomical Therapeutic Chemical (ATC) codes by ATC level 1 for structured, unstructured, and combined data.

<table>
<thead>
<tr>
<th>ATC 1st level</th>
<th>Structured drug prescriptions (n=843,980), n/N (%)</th>
<th>Unstructured drug prescriptions for evaluated subset (n=924,173), n/N (%)</th>
<th>Total number (n=1,768,153), n/N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N – Nervous system</td>
<td>197,831/322,286 (61.38)</td>
<td>124,455/322,286 (38.62)</td>
<td>322,286/1,337,565 (24.1)</td>
</tr>
<tr>
<td>B – Blood and blood forming organs</td>
<td>164,032/251,120 (65.32)</td>
<td>87,088/251,120 (34.68)</td>
<td>251,120/1,337,565 (18.77)</td>
</tr>
<tr>
<td>A – Alimentary tract and metabolism</td>
<td>137,988/250,543 (55.08)</td>
<td>112,555/250,543 (44.92)</td>
<td>250,543/1,337,565 (18.73)</td>
</tr>
<tr>
<td>C – Cardiovascular system</td>
<td>170,703/247,629 (68.93)</td>
<td>76,926/247,629 (31.07)</td>
<td>247,629/1,337,565 (18.51)</td>
</tr>
<tr>
<td>J – Anti-infective for systemic use</td>
<td>60,844/88,659 (68.63)</td>
<td>27,815/88,659 (31.37)</td>
<td>88,659/1,337,565 (6.63)</td>
</tr>
<tr>
<td>H – Systemic hormonal drugs, excluding sex hormones and insulins</td>
<td>51,296/64,199 (79.9)</td>
<td>12,903/64,199 (20.1)</td>
<td>64,199/1,337,565 (4.8)</td>
</tr>
<tr>
<td>M – Musculo-skeletal system</td>
<td>12,083/36,819 (32.82)</td>
<td>24,736/36,819 (67.18)</td>
<td>36,819/1,337,565 (2.75)</td>
</tr>
<tr>
<td>R – Respiratory system</td>
<td>19,686/28,148 (69.94)</td>
<td>8462/28,148 (30.06)</td>
<td>28,148/1,337,565 (2.1)</td>
</tr>
<tr>
<td>V – Various</td>
<td>9639/14,672 (65.7)</td>
<td>5033/14,672 (34.3)</td>
<td>14,672/1,337,565 (1.1)</td>
</tr>
<tr>
<td>L – Antineoplastic and immunomodulating agents</td>
<td>8670/14,538 (59.64)</td>
<td>5868/14,538 (40.36)</td>
<td>14,538/1,337,565 (1.09)</td>
</tr>
<tr>
<td>G – Genito urinary system and sex hormones</td>
<td>3662/8778 (41.71)</td>
<td>5116/8778 (58.28)</td>
<td>8778/1,337,565 (0.66)</td>
</tr>
<tr>
<td>S – Sensory organs</td>
<td>5077/5179 (98.03)</td>
<td>102/5179 (1.97)</td>
<td>5179/1,337,565 (0.39)</td>
</tr>
<tr>
<td>D – Dermatologicals</td>
<td>2127/3534 (60.19)</td>
<td>1407/3534 (39.81)</td>
<td>3534/1,337,565 (0.26)</td>
</tr>
<tr>
<td>P – Antiparasitic products, insecticides, and repellents</td>
<td>342/1461 (23.41)</td>
<td>1119/1461 (76.59)</td>
<td>1461/1,337,565 (0.11)</td>
</tr>
<tr>
<td>no med</td>
<td>0/1,768,153 (0)</td>
<td>166,307/1,768,153 (9.41)</td>
<td>166,307/1,768,153 (9.41)</td>
</tr>
<tr>
<td>unspec</td>
<td>0/1,768,153 (0)</td>
<td>2187/1,768,153 (0.12)</td>
<td>2187/1,768,153 (0.12)</td>
</tr>
<tr>
<td>Total validated</td>
<td>843,980/1,768,153 (47.73)</td>
<td>662,079/1,768,153 (37.44)</td>
<td>1,506,059/1,768,153 (85.18)</td>
</tr>
<tr>
<td>Not validated</td>
<td>0/1,768,153 (0)</td>
<td>262,094/1,768,153 (14.82)</td>
<td>262,094/1,768,153 (14.82)</td>
</tr>
</tbody>
</table>
Figure 4. Structuredness of drug prescriptions by ATC groups for 85.18% of initial data set DS1. ATC: Anatomical Therapeutic Chemical.
Figure 5. Structuredness of drug prescription by ATC L5 (a) total unstructured data and (b) by each ATC L1 group. A: Alimentary tract and metabolism; ATC: Anatomical Therapeutic Chemical; B: Blood and blood forming organs; C: Cardiovascular system; D: Dermatological; G: Genito urinary system and sex hormones; H: Systemic hormonal drugs, excluding sex hormones and insulins; J: Anti-infective for systemic use; L: Antineoplastic and immunomodulating agents; L1: level 1; L5: level 5; M: Musculo-skeletal system; N: Nervous system; P: Antiparasitic products, insecticides, and repellents; R: Respiratory system; S: Sensory organs; V: Various.

Together, these 4 ATC level 5 codes accounted for 14.79% (261,460/1,768,153) of the total data in the data set 1. Pantoprazole had the lowest level of structured data for these 4 ingredients (10,490/65,861, 15.93%), whereas oxycodone had the highest level of structured data (44,952/46,434, 96.81%). Multimedia Appendix 3 contains the complete list of all ingredients with their ATC level 5 codes, number of structured, unstructured, and total drug prescriptions. In addition, it includes the percentage strength of each ingredient, ordered by the total number of drug prescriptions, starting with the largest number.

Discussion

Principal Findings

Our 4-step approach ensures data quality assessment as recommended by Zozus et al [18]. We provide transparency by reporting the structuredness of drug prescriptions. In addition, our approach improved the structuredness and thus the completeness of drug prescription data. This leads to better usability for secondary use on research infrastructure, such as the OMOP based ATC codes accompanied by manual review.

The initial analysis of the drug data showed a ratio of 52.3% unstructured to 47.7% structured drug prescriptions. With the algorithms presented, the structuredness could be increased to 85.1% with 1 level of evaluation. For the evaluation of the initial data set 1, manual examination of the most frequent 1000 free-text entries was sufficient, and we were able to achieve the targeted minimum coverage of 80%. Algorithm 3, which was based on similarity matching, was found to quantitatively outperform the other 2 algorithms, providing results for all unstructured drug prescriptions. In terms of the reliability of the results, algorithm 3 had a correctness rate of only 76.5%.
Therefore, the evaluation phase was critical to manually correct all incorrectly derived ATC codes. In addition, the manual evaluation process was critical in identifying patterns that can be used to determine the reliability of the algorithms based on additional factors, such as algorithm-to-algorithm matches and the Levenshtein similarity score of algorithm 3. When all 3 methods or algorithms 1 and 2 yielded the same ATC code, the algorithm results were considered correct in each case. The scoring process yielded a percentage of incorrect results (approximately 1.5%) when algorithms 2 and 3 or 1 and 3 yielded the same results.

The patterns identified are a good indication that can help increase the reliability of the results without further manual evaluation, compared with the overall correctness of 76.5% (765/1000) found for algorithm 3 results. The Levenshtein similarity score revealed another trend for algorithm 3: incorrect results had a significantly lower mean similarity score than correct results. The exceptions have been isolated to a small number of ingredients, reasons such as missing dosage form and dosage information, and cases where the same ingredients were used for single and combination drug products, resulting in separate ATC codes. The quality of the RWD has a major impact on the results of observational studies that rely on it. It is important to ensure that RWD data are suitable for use in observational studies [24] and that any limitations or quality concerns are explicitly stated [25].

Limitations

Currently, the analyzed data set is limited to inpatient drug prescriptions from the University Hospital Carl Gustav Carus Dresden. No drug prescription data from intensive care medicine were included, and no other institution has used our technique yet. Outliers or rare patterns may have gone undetected because the study was limited to the first 1000 free-text prescriptions. Although this covers most of the data, the results of the algorithm for the remaining free-text entries are yet to be evaluated. This study does not include an outcome evaluation of additional drug prescription entries based on identified patterns. Currently, the method is limited to determining ATC codes for unstructured drug prescriptions and does not consider other terminologies such as RxNorm.

Comparison With Prior Work

Most studies evaluating the quality of RWD data refer to the dimensions of completeness and accuracy compared with predefined gold standard data that vary by publication, as identified by Weiskopf et al [19]. We did not define our gold standard based on RWD sources but used the internationally recognized and widely used terminology ATC as standardized terminology and provided a method to automatically determine the appropriate ATC for drug prescription data that are unstructured and available only as free text. Wang et al [26] developed a rule-based data quality system with >6000 criteria for plausibility testing (eg, pregnancy is not plausible in male patients) but did not address data harmonization by mapping unstructured free-text data to defined terminologies for research. Unlike Schmidt et al [27] and Kahn et al [25], our study not only focus on data quality assessment but also defines the absence of structure in the data as free text without a corresponding ATC code and builds on previous research by proposing a method to improve unstructured data by automatically annotating the appropriate ATC code.

The high proportion of free-text or unstructured drug prescriptions was due to the hospital’s prescription system and local conditions. According to previous studies on the data structure in RWD [28], this is a widespread challenge in Germany. However, the issues of dealing with unstructured data in EHR records that prevent interoperability are widespread as stated by Kruse et al [29,30] in their systematic reviews of existing literature on the use of EHR data that need to be addressed to ensure “fitness for use” in general. Compared with the well-established Unified Medical Language System MetaMap [31,32], which has been used by industry and academia for many years, our NLP approach focuses on a lightweight implementation. On the one hand, this limits the configuration possibilities, but on the other hand, it reduces the computational efforts and promotes the performance of the ATC code recognition. Because MetaMap focuses only on English and does not support German drug catalogs, our approach closes this gap and can be adapted to other languages as well.

Future Work

The presented 4-step approach can be applied to any RWD with unstructured data such as conditions, procedures, or test results. This approach will be tested in the future on other sites that provide drug data and product lists with ATC codes. In the next phases, further research will be conducted on pattern recognition to enable reliable prediction of the accuracy of results for specific ATC codes rather than manually checking them. In addition, new NLP-based algorithms will be implemented to improve the overall reliability of the results. Furthermore, our approach can be applied to other hospital sites that participate in the German Medical Informatics Initiative [33,34] in the following steps. Our approach is not limited to the German language. Because the only requirement is to provide a common list of ingredients or drug products for comparison with unstructured free text, this can work for any other language if compared texts are available in the same language.

Conclusions

RWD observational research requires a high level of data structuredness. Even more critical is the awareness of limitations as well as transparency of the level of structure of the data on which the research is based. Using drug prescriptions as a first use case, we were able to investigate and improve the structure of RWD, which can be applied to other RWDs in the future. Although the presented methods require manual verification to ensure that the results are correct, the methodology is promising and can be used to improve structuredness of data.
Acknowledgments
This study is part of the project MIRACUM, funded by the German Ministry of Education and Research (FKZ 01ZZ1801L).

Authors’ Contributions
IR and FB worked on the conceptualization and methodology. JS and IR worked on the software. IR, SF, AF, and JS worked on the evaluation. IR, FB, JW, and JS analyzed the data. IR, FB, JS, and MW curated the data. FB and IR wrote the original draft. JW, SF, AF, and MS reviewed and edited the draft. IR and FB worked on the visualization. FB and MS were responsible for supervision. IR was responsible for project administration. All authors read and agreed to the published version of the manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Small set of rules.
[XLSX File (Microsoft Excel File), 9 KB - medinform_v11i1e40312_app1.xlsx]

Multimedia Appendix 2
Data set 4.
[XLSX File (Microsoft Excel File), 131 KB - medinform_v11i1e40312_app2.xlsx]

Multimedia Appendix 3
Complete list of ATC L 5 with frequency and proportion of structured versus unstructured entries.
[XLSX File (Microsoft Excel File), 32 KB - medinform_v11i1e40312_app3.xlsx]

References


Abbreviations

ATC: Anatomical Therapeutic Chemical
**EHR**: electronical health record  
**NLP**: natural language processing  
**OMOP**: Observational Medical Outcomes Partnership  
**RCT**: randomized controlled trial  
**RWD**: real-world data  
**UKD**: University Hospital Carl Gustav Carus Dresden

Edited by C Lovis; submitted 15.06.22; peer-reviewed by A Lamer, M Pedrera Jiménez, B Ru; comments to author 07.09.22; revised version received 27.09.22; accepted 18.11.22; published 25.01.23.

Please cite as:  
Reinecke I, Siebel J, Fuhrmann S, Fischer A, Sedlmayr M, Weidner J, Bathelt F  
Assessment and Improvement of Drug Data Structuredness From Electronic Health Records: Algorithm Development and Validation  
JMIR Med Inform 2023;11:e40312  
URL: https://medinform.jmir.org/2023/1/e40312  
doi: 10.2196/40312  
PMID: 36696159

©Ines Reinecke, Joscha Siebel, Saskia Fuhrmann, Andreas Fischer, Martin Sedlmayr, Jens Weidner, Franziska Bathelt. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 25.01.2023. 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.
An Ontology-Based Approach for Consolidating Patient Data Standardized With European Norm/International Organization for Standardization 13606 (EN/ISO 13606) Into Joint Observational Medical Outcomes Partnership (OMOP) Repositories: Description of a Methodology

Santiago Frid1,2, MSc, MD; Xavier Pastor Duran1,2, MD, PhD; Guillem Bracons Cucó3, MSc; Miguel Pedrera-Jiménez4, MSc, PhD; Pablo Serrano-Balazote5, MD; Adolfo Muñoz Carrero6, PhD; Raimundo Lozano-Rubi1,2, MD, PhD

1Medical Informatics Unit, Hospital Clínic de Barcelona, Barcelona, Spain
2Clinical Foundations Department, Universitat de Barcelona, Barcelona, Spain
3Fundació Clínic per a la Recerca Biomèdica, Barcelona, Spain
4Data Science Unit, Hospital 12 de Octubre, Madrid, Spain
5Direction of Planification, Hospital 12 de Octubre, Madrid, Spain
6Unit of Investigation in Telemedicine and Digital Health, Instituto de Salud Carlos III, Madrid, Spain

Corresponding Author:
Santiago Frid, MSc, MD
Clinical Foundations Department
Universitat de Barcelona
Casanova 143
Barcelona, 08036
Spain
Phone: 34 934035258
Email: santifrid@gmail.com

Abstract

Background: To discover new knowledge from data, they must be correct and in a consistent format. OntoCR, a clinical repository developed at Hospital Clínic de Barcelona, uses ontologies to represent clinical knowledge and map locally defined variables to health information standards and common data models.

Objective: The aim of the study is to design and implement a scalable methodology based on the dual-model paradigm and the use of ontologies to consolidate clinical data from different organizations in a standardized repository for research purposes without loss of meaning.

Methods: First, the relevant clinical variables are defined, and the corresponding European Norm/International Organization for Standardization (EN/ISO) 13606 archetypes are created. Data sources are then identified, and an extract, transform, and load process is carried out. Once the final data set is obtained, the data are transformed to create EN/ISO 13606–normalized electronic health record (EHR) extracts. Afterward, ontologies that represent archetyped concepts and map them to EN/ISO 13606 and Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) standards are created and uploaded to OntoCR. Data stored in the extracts are inserted into its corresponding place in the ontology, thus obtaining instantiated patient data in the ontology-based repository. Finally, data can be extracted via SPARQL queries as OMOP CDM–compliant tables.

Results: Using this methodology, EN/ISO 13606–standardized archetypes that allow for the reuse of clinical information were created, and the knowledge representation of our clinical repository by modeling and mapping ontologies was extended. Furthermore, EN/ISO 13606–compliant EHR extracts of patients (6803), episodes (13,938), diagnosis (190,878), administered medication (222,225), cumulative drug dose (222,225), prescribed medication (351,247), movements between units (47,817), clinical observations (6,736,745), laboratory observations (3,392,873), limitation of life-sustaining treatment (1,298), and procedures (19,861) were created. Since the creation of the application that inserts data from extracts into the ontologies is not yet finished, the queries were tested and the methodology was validated by importing data from a random subset of patients into the ontologies using a locally developed Protégé plugin (“OntoLoad”). In total, 10 OMOP CDM–compliant tables (“Condition_occurrence,”...
Conclusions: This study proposes a methodology for standardizing clinical data, thus allowing its reuse without any changes in the meaning of the modeled concepts. Although this paper focuses on health research, our methodology suggests that the data be initially standardized per EN/ISO 13606 to obtain EHR extracts with a high level of granularity that can be used for any purpose. Ontologies constitute a valuable approach for knowledge representation and standardization of health information in a standard-agnostic manner. With the proposed methodology, institutions can go from local raw data to standardized, semantically interoperable EN/ISO 13606 and OMOP repositories.

KEYWORDS
health information interoperability; health research; health information standards; dual model; secondary use of health data; Observational Medical Outcomes Partnership Common Data Model; European Norm/International Organization for Standardization 13606; health records; ontologies; clinical data

Introduction
The term primary use of health data encompasses the generation and use of data within the context of individual health care in hospitals and physicians’ offices to serve direct care needs [1]. The term secondary use of health data is defined by the American Medical Informatics Association as “non-direct care use of PHI [personal health information] including but not limited to analysis, research, quality/safety measurement, public health, payment, provider certification or accreditation, and marketing and other business including strictly commercial activities” [2]. Although they can be further categorized [3], one of the main types of secondary uses is research.

Clinical data sharing for research is highly relevant from a scientific, economic, and ethical perspective [4]. The overwhelming increment in the volume of available data is directly related with the emergence of a new paradigm of scientific methodology in which massive amounts of data are processed and analyzed for obtaining knowledge through machine learning and data mining algorithms [5].

Despite the growth of big data technologies and the use of artificial intelligence, in order to discover new knowledge from data, they must be correct and in a consistent format, which requires a great amount of resources for cleaning, binding, and organizing them. The semantics of data is a key component regarding the aforementioned challenges. To use the electronic health record (EHR) data for different projects, it must maintain its semantics and context, independently of any particular use case. This is especially important in research, where EHR reuse processes are often based on black boxes on which the final data customer is unaware of how the data uploaded to their research database were recorded, extracted, and transformed [6].

A common health information standard should be used in both primary and secondary use to share clinical information in a way that it can be unequivocally interpreted, both syntactically and semantically, by 2 or more systems. European Norm/International Organization for Standardization (EN/ISO) 13606 is a health information standard that seeks to define a rigorous and stable architecture for communicating the health records of a single patient, preserving the original clinical meaning. It is based on a dual model that includes a reference model (RM; with the necessary components and their constraints to represent EHR extracts) and an archetype model (AM; for the formalization of clinical-domain concepts according to the RM) [7,8]. Archetypes allow the formal representation of the structure of clinical information and its meaning (through terminology binding) so that it is automatically processable by information systems.

Furthermore, the EN/ISO 13940 norm [9] provides a conceptual framework centered in the clinical process. This norm, based on a clinical perspective, defines the system of concepts that are necessary for achieving continuity in the caregiving process, including both the content and the context of the health activities. This ample norm defines the concepts relative to health care actors, health problems, sanitary activities, health care processes, sanitary planification, time-related concepts, responsibilities, and information management.

Moreover, the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) defines a common format (data model), as well as a common representation (terminologies, vocabularies, and coding schemes), to allow systematic analyses of disparate observational databases using a library of standard analytic routines that have been written based on the common format [10]. The OMOP CDM is considered by several authors as the most adequate data model for sharing data in EHR-based longitudinal studies [11-13].

This paper describes the work carried out between Hospital Clínic de Barcelona (HCB), Hospital 12 de Octubre (H12O), and Instituto de Salud Carlos III (ISCIII), which seeks to consolidate clinical data of hospitalized patients with COVID-19 from different hospitals in joint repositories, structured with EN/ISO 13606 and then normalized according to the OMOP CDM.

The aim of this study is to design and implement a scalable methodology based on the dual-model paradigm and the use of ontologies to consolidate clinical data from different organizations in a standardized repository for research purposes without loss of meaning. This implies a series of particular objectives such as (1) to define a set of relevant clinical and
biochemical variables of patients hospitalized with COVID-19, (2) to model a set of standardized archetypes based on EN/ISO 13606 to communicate such information, (3) to conceptually represent those clinical variables by means of ontologies, (4) to generate EN/ISO 13606–standardized EHR extracts of COVID-19 patients, and (5) to map and transform the source data to create OMOP CDM–compliant tables.

Methods

Ethical Considerations
This study was approved by the Hospital Clínica de Barcelona Ethics Committee for Investigation with Drugs (HCB/2018/0573).

Cohort Inclusion Criteria
We included in this project patients with COVID-19 admitted to the emergency room (ER) or hospitalized between February 17, 2020 (beginning of the first wave in Spain), and February 15, 2022 (end of the sixth wave in Spain).

Methodology
The following methodology comprises a series of steps in order to achieve the study’s objectives.

Step 1: Definition of Clinical Variables, Data Structures, and EN/ISO 13606 Archetypes
The first step consists in deciding the clinically relevant variables that should be included. Afterward, the data structures must be defined, including the fields, their descriptions, and the standardized terminologies and classifications to be used.

Since OMOP CDM is intended for secondary use of data (specifically, for biomedical research), its granularity is somewhat reduced when compared to raw data captured in hospitals. For this reason, the Medical Informatics Unit (MIU) at HCB decided to first standardize the data according to EN/ISO 13606, in order to have semantically interoperable EHR extracts with the maximum level of detail.

Therefore, the MIU at HCB and the Data Science Unit at H12O defined the EN/ISO 13606 archetypes to be used, modeled with the software LinkEHR [14] created by VeraTech for Health. The data types used are those established by the standard’s RM.

This RM has multiple components, including the entry (a result of 1 clinical action, 1 observation, 1 clinical interpretation, or 1 intention) and its elements (the leaf node of the EHR hierarchy, containing a single data value). In our project, the archetypes modeled at the entry level of the RM were the following: diagnosis, episodes, limitation of life-sustaining treatment, administered medication, cumulative drug dose, prescribed medication, movements between units, clinical observations, laboratory observations, patients, health problems, and procedures. These archetypes were registered under a Creative Commons license (ID 2204210968527), so that any user who follows the license terms can share and adapt them [15]. Figure 1 shows a mind map of the diagnosis entry archetype as an example.

Figure 1. Mind map of the EN/ISO 13606 “diagnosis” archetype in Spanish, modeled with LinkEHR. The “diagnosis” entry has 6 elements: episode_id, diagnosis, diagnosis_datetime, patient_id, diagnosis_id, and source. Each of them has its corresponding data type. EN/ISO: European Norm/International Organization for Standardization.

Step 2: Identification of Data Sources and Extract, Transform, and Load
Afterward, the corresponding data sources must be identified, in order to carry out the extract, transform, and load (ETL) process. In our case, these sources were (1) structured data from HCB’s health information system (HIS), SAP; (2) unstructured data from HCB’s HIS. A collaborative work with Barcelona Supercomputing Center (BSC) allows for the recognition of clinical entities through natural language processing and its extraction as normalized structured data; (3) outpatient setting structured data from Agència de Qualitat i Avaluació Sanitàries de Catalunya.

Since the last 2 sources come from separate projects whose description is besides the objective of this paper, we will focus on the first one. Archetypes created in the previous step were used as templates for identifying data in the aforementioned sources. Periodic meetings were held with the Information Technology Department at HCB to identify the location of the data and the transformations needed to obtain the structured data defined in the previous step. Once this was achieved, the tables were loaded into a MySQL database hosted on a dedicated server of the MIU.
**Step 3: Creation of EN/ISO 13606 EHR Extracts From Source Data**

Once the final data set is obtained, data must be transformed to create EHR extracts normalized according to EN/ISO 13606. This transformation includes mapping of local variables to standardized nomenclatures and classifications (Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT), International Classification of Diseases 10—Clinical Modification (ICD-10-CM), Logical Observation Identifiers Names and Codes (LOINC)), assigning readable descriptions to local codes, and categorizing certain concepts (e.g., grouping hospital units according to the level of care).

This process is performed by mapping archetypes to the implicated information systems, without the need to modify them. This approach allows the automation of data extraction and the reuse of this methodology for other use cases with very little effort, which constitutes one of the great advantages of dual-model strategies.

In our case, we carried out this process with the help of VeraTech for Health, our technical partners, using LinkEHR, thus creating extracts on our dedicated server and constituting an EN/ISO 13606 standardized clinical repository. **Figure 2** shows a test example of an EN/ISO 13606 EHR extract (without real-patient data). In this extract, the ICD-10-CM code H40.9 (unspecified glaucoma) is being communicated, alongside its date and time of record and the ID of the clinical episode it pertains to.

**Figure 2.** Anonymized, normalized EN/ISO 13606 EHR extract of diagnosis in Spanish. EHR: electronic health record; EN/ISO: European Norm/International Organization for Standardization.

**Step 4: Creation of Ontologies**

Traditionally, clinical concepts and the relationships between them have been poorly developed in HISs. The MIU at HCB developed OntoCR, an ontology-based clinical repository, conforming to EN/ISO 13606 standard [16,17]. The use of ontologies allows for the definition of a conceptual architecture centered on the representation of the clinical process, while the use of EN/ISO 13606 allows syntactic and semantic interoperability between systems. EN/ISO 13940 was also used to define the generic concepts needed to achieve continuity of care, representing both the content and the context of the health care services.

One of the main advantages of ontologies is their flexibility to perform changes with minimum use of resources, adapting to an ever-changing environment. Likewise, ontologies allow the addition of conceptual layers, thus mapping locally defined concepts to health information standards, facilitating the communication of information without loss of meaning.

A relational database (OWL-DB) is used for storing ontologies and instantiated data, designed according to the Web Ontology Language (OWL) specification [18]. The ontologies in this project were created using Protégé, a free, open-source ontology editor created by Stanford University that fully supports OWL and Resource Description Framework (RDF) specifications from the World Wide Web Consortium [19]. A plug-in developed by our team, the OWL-DB plugin, connects Protégé with the OWL-DB module at the storage level.

These ontologies were conceptualized in 3 different layers. The first one describes the concepts modeled in the archetypes, with the classes and properties that describe the data structure defined in the first phase. Data types according to the EN/ISO 13606 RM were used.
In the next layer, we used a locally created ontology that reproduces the EN/ISO 13606 RM and AM. By creating an additional ontology that maps the archetypal concepts to the EN/ISO 13606 model, we structured our data according to the standard. In this layer, each entry-level archetype is represented in a separate ontology.

As with EN/ISO 13606, we created an ontology that models the OMOP CDM and afterward mapped archetypal concepts to the corresponding meta-class of the standard. So, the third layer consists of ontologies for each archetype that reproduce concepts according to the OMOP CDM structure. Figure 3 shows these 3 ontologies. The left image (ontology of the AM of diagnosis) depicts the class “Diagnosis” with its properties diagnosis, diagnosis_id, episode, diagnosis_datetime, source, and patient_id. In the upper-right image, a new ontology was created where the class “Diagnosis” was modeled as a subclass of “iso13606: Entry,” thus inheriting its properties defined in the RM. Finally, in the lower right image, a third ontology maps the property diagnosis with OMOP CDM’s meta-class “condition_source_value.”

Figure 3. Ontologies of the archetype model of diagnosis (left) and its mapping to the EN/ISO 13606 structure (upper right) and the OMOP CDM (lower right) in Spanish, edited with Protégé software. EN/ISO: European Norm/International Organization for Standardization; OMOP CDM: Observational Medical Outcomes Partnership Common Data Model.

Afterward, these ontologies must be loaded into a production environment of OntoCR so as to generate the structure that can receive instantiated data of patients and store it.

**Step 5: Integration of EN/ISO 13606 Extracts Into the Ontology-Based Clinical Repository and Extraction of Data as OMOP CDM–Compliant Tables**

Once the ontological structure is ready to receive the data, the EHR extracts must be inserted into the repository, thus incorporating the normalized, instantiated data. We initially explored the possibility of adapting a preexisting application programming interface (API) that was used for the same purpose in a previous project. However, the resources needed for its adaptation were significantly elevated, and its scalability reduced. Therefore, we decided to work on an application that identifies each archetype node within the extract and inserts it into its counterpart in the OWL file. This is facilitated by the representation in the ontologies of each archetype, their nodes, and the data types used (compliant with EN/ISO 13606).

Finally, data stored in the ontology-based clinical repository needs to be extracted through SPARQL queries, a language used for graph databases. Since archetypal concepts have been previously mapped to the OMOP CDM, by performing these queries, the extraction process is simplified. If there are cases in which data needs to be transformed to fit the CDM, such transformations can be included in the queries or carried out via SQL queries once relational tables are obtained.

In Figure 4, a SPARQL query for extracting data for the OMOP CDM PERSON table is shown. Attributes that are not present in the ontological repository must still be included in the SELECT clause so as to create the corresponding table column without any instantiated data. Since EN/ISO 13606 data types were used in the extracts and modeled in the ontologies, they were also represented in the queries (see the lower lines of SPARQL code).
Data anonymization is performed by the IT department at this level using an institutional software solution. This way, EN/ISO 13606 extracts contain identified data that can be used for primary uses, while OMOP CDM tables are anonymized for secondary uses.

Once obtained, the anonymized data can be consolidated in a joint OMOP CDM repository with other institutions that use the same standard (in our case, H12O). OMOP CDM has a large number of tables, divided into 6 groups: standardized clinical data, standardized health system data, standardized derived elements, standardized health economics, standardized metadata, and standardized vocabularies. Our OMOP CDM repository contains the following tables, which are part of the standardized clinical data: “Condition_occurrence,” “Death,” “Device_exposure,” “Drug_exposure,” “Measurement,” “Observation,” “Observation_period,” “Person,” “Visit_detail,” and “Visit_occurrence.”

Figure 5 shows an overview of the whole process. The knowledge modeling starts with the creation of EN/ISO 13606 archetypes based on clinical concepts, which are then represented in ontologies that map them to EN/ISO 13606 RM and OMOP CDM. These ontologies are uploaded to OntoCR without instantiated patient data yet.
The data-related processes begin with the archetype-based extraction of raw data from our local system into a MySQL database and its transformation to create EN/ISO EHR extracts, which are then inserted into OntoCR via an application specifically developed for this project. SPARQL queries are performed against this ontological repository to obtain an OMOP CDM repository that is consolidated with H12O in a joint one.

**Ontologies**

Another interesting result of this study is the development of the ontologies that represent OMOP CDM, as well as their mappings to EN/ISO 13606 AM and RM. This process was carried out by members of the MIU at HCB after carefully reading the pertinent documentation of these standards and designing the optimal way of using them to represent clinical concepts.

Furthermore, representing clinical variables by means of ontologies is another way of reusing clinical information. With the creation of new ontologies for each project at HCB, where we have developed the ontology-based clinical repository OntoCR, we continue to extend our clinical knowledge representation.

**EN/ISO 13606 Extracts**

Table 1 shows the correspondence between EHR archetypes, OMOP CDM tables, number of extracts created throughout the study, and the number of COVID-19 patients they pertain to. We have included the diagnoses recorded in the episodes of the study period as well as the historical ones. Health problem...
entries are part of the aforementioned project with BSC to extract clinical entities from unstructured texts through natural language processing, so they will not be included in this table.

**Table 1.** Correspondence between EHR\(^a\) archetypes, OMOP CDM\(^b\) tables, number of extracts created throughout the study, and the number of patients they pertain to.

<table>
<thead>
<tr>
<th>EHR archetype</th>
<th>OMOP CDM table</th>
<th>EN/ISO(^c) 13606 extracts, n</th>
<th>Patients, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>“Person”</td>
<td>6803</td>
<td>6803</td>
</tr>
<tr>
<td>Episode</td>
<td>“Visit_occurrence”</td>
<td>13,938</td>
<td>6791</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>“Condition_occurrence”</td>
<td>190,878</td>
<td>6799</td>
</tr>
<tr>
<td>Cumulative drug dose</td>
<td>“Drug_exposure”</td>
<td>262,770</td>
<td>5630</td>
</tr>
<tr>
<td>Administered medication</td>
<td>“Drug_exposure”</td>
<td>262,770</td>
<td>5630</td>
</tr>
<tr>
<td>Prescribed medication</td>
<td>“Drug_exposure”</td>
<td>341,986</td>
<td>5639</td>
</tr>
<tr>
<td>Movements between units</td>
<td>“Visit_detail”</td>
<td>47,817</td>
<td>6791</td>
</tr>
<tr>
<td>Clinical observation</td>
<td>“Measurement”</td>
<td>6,736,745</td>
<td>5973</td>
</tr>
<tr>
<td>Laboratory observation</td>
<td>“Measurement”</td>
<td>3,392,873</td>
<td>6001</td>
</tr>
<tr>
<td>Limitation of life-sustaining treatment</td>
<td>“Observation”</td>
<td>1298</td>
<td>1142</td>
</tr>
<tr>
<td>Procedure</td>
<td>“Procedure_occurrence”</td>
<td>19,861</td>
<td>4994</td>
</tr>
</tbody>
</table>

\(^a\)EHR: electronic health record.
\(^b\)OMOP CDM: Observational Medical Outcomes Partnership Common Data Model.
\(^c\)EN/ISO: European Norm/International Organization for Standardization.

**OMOP CDM–Compliant Clinical Tables**

We still do not have the final number of records in our OMOP tables, since the initial approach of adapting the preexisting API had to be replaced by the creation of the application that inserts data from the extracts into the ontologies. However, an OMOP database for a random small subset of patients was successfully created to test the queries and validate the methodology. This was performed using a locally developed Protégé plugin (“OntoLoad”) that imports a set of data from a relational database into the ontologies [17]. **Table 2** describes the OMOP tables that were created.

**Table 2.** OMOP CDM\(^a\)-compliant clinical tables created for a random small subset of patients.

<table>
<thead>
<tr>
<th>OMOP CDM table</th>
<th>Patients, n</th>
<th>Records, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Condition occurrence”</td>
<td>121</td>
<td>864</td>
</tr>
<tr>
<td>“Death”</td>
<td>110</td>
<td>110</td>
</tr>
<tr>
<td>“Device_exposure”</td>
<td>3</td>
<td>56</td>
</tr>
<tr>
<td>“Drug_exposure”</td>
<td>106</td>
<td>5609</td>
</tr>
<tr>
<td>“Measurement”</td>
<td>3</td>
<td>2091</td>
</tr>
<tr>
<td>“Observation”</td>
<td>3</td>
<td>195</td>
</tr>
<tr>
<td>“Observation_period”</td>
<td>897</td>
<td>897</td>
</tr>
<tr>
<td>“Person”</td>
<td>922</td>
<td>922</td>
</tr>
<tr>
<td>“Visit_detail”</td>
<td>250</td>
<td>772</td>
</tr>
<tr>
<td>“Visit_occurrence”</td>
<td>897</td>
<td>971</td>
</tr>
</tbody>
</table>

\(^a\)OMOP CDM: Observational Medical Outcomes Partnership Common Data Model.

**Discussion**

**Principal Results**

This study proposes a methodology for standardizing clinical data, thus allowing its reuse without any change in the meaning of the modeled concepts. Although the focus of this paper is health research, our methodology suggests that the data be initially standardized according to EN/ISO 13606 to obtain EHR extracts with a high level of granularity that can be used for any purpose, as previous studies have suggested [20]. Afterward, its transformation to OMOP CDM–compliant tables allows its consolidation in joint repositories for research purposes. Although EN/ISO 13606 was chosen because of the operational mechanisms it offers for data exchange, due to the flexibility and standard-agnostic nature of our methodology, there is complete independence regarding any specific standard. Thus,
by modeling ontologies of other standards and mapping them to local variables, we may, for example, carry out transformations between EN/ISO 13606, OpenEHR [21], Fast Healthcare Interoperability Resources (FHIR) [22], OMOP CDM, and Informatics for Integrating Biology and the Bedside (i2b2) [23] with the minimum use of resources and without the need for changes in the database structure. Health information standards such as EN/ISO 13606 and OpenEHR allow the modeling and formalization of clinical knowledge through their RMs and archetypes [24], and ontologies are precisely a tool for carrying out such tasks. This is what makes them ideal in the context of an implementation of a dual-model strategy, allowing the representation of concepts in the health domain, its standardization, and the storage of instantiated patient data.

Furthermore, ontologies provide several advantages for the conceptualization of entities in a domain. It explicitly represents domain knowledge, allows the application of inference processes, enables the reuse of domain knowledge, allows data aggregation, and detects new associations between concepts [17].

It is clear for us that loading normalized data onto clinical repositories (instead of ad hoc data loading) provides many benefits. It is possible to reuse the same interoperability standards used in health care, adapting them to this new paradigm [25]. This approach allows the availability of clinical data for further single- or multicenter research.

We would like to highlight the vital importance of continuous collaborative research. This study is framed within a continued line of research since 2009 between HCB, ISCIII, and H12O. In this line of collaborative research, a standardized and transparent process has been designed and implemented for obtaining standardized data models for research from EHR raw data. Hence, in the first stage, the basis for a semantically interoperable clinical information management system based on EN/ISO 13606 was defined, proving that clinical information residing in heterogeneous systems could be normalized, combined, and communicated without loss of meaning. In the second stage, a common information model that reflects the clinical process and the relationships between the clinical records components was developed. In the third stage, a normalized information model based on EN/ISO 13606 archetypes was implemented and applied to local information systems for specific clinical use cases. With this model, it is possible to construct and order information recovered from these complex systems for the exchange of integral health and social information of patients and to use it for secondary purposes.

Comparison With Prior Work

Many of the requisites of clinical data repositories for primary use are common to those for secondary use, such as normalized clinical information models, controlled terminologies, identification of actors, and contextual information. Developments carried out for primary use repositories are also profitable for secondary uses, and the progresses derived from secondary uses accelerate the advances in shared clinical records. A lot of work has been reported in this field throughout the globe in the last years, which has led to developing policies, repository models and its application in the form of competitive projects [2,26,27].

It is very usual for researchers to resort to the generation of their own data for research and its manual introduction into data management systems. It is also quite common for them to use general purpose tools, particularly spreadsheets, as data management systems [28], while there is perception of a high need of additional support for the analysis of high volumes of data. This represents a significant problem, since these applications cannot guarantee the consistency of data, and they present difficulties for sharing and consolidating data and a limited capability of data exploitation.

Different methodologies have been proposed to create OMOP repositories from raw data. Some approaches are based on a simple mapping of local variables to their OMOP CDM counterparts, an alignment of vocabularies using the Athena tool provided by OHDSI and an ETL process through SQL scripts [29]. Other authors have proposed transforming source data to RDF, carrying out a semantic mapping (in some cases, using an ontological representation of OMOP CDM), and loading it to a data store [30,31].

Likewise, other standard-agnostic approaches have been reported in the literature. The ongoing INFOBANCO project of the Madrid Region [32] seeks to create a platform for the management, persistence, exchange, and reuse of health data focused on applying each health information standard for the purpose it was intended to, offering multiple interoperability and exploitation services suited for specific use cases [24]. Furthermore, the 3-pillar strategy of the Swiss Personalized Health Network [33] pursues a semantically interoperable clinical data landscape based on a multidimensional encoding of concepts, an RDF-based storage and transport of the instances of these concepts and a conversion of RDF to any target data model.

Strengths and Limitations

This study has many strengths that are worth mentioning. On the one hand, it describes a real-world collaborative effort between 3 health care institutions in Spain to model, share, and consolidate standardized patient data. Furthermore, the standard-agnostic nature of the proposed methodology leads to a significant scalability, allowing transformation between different health information standards and common data models. The software used in our methodology (LinkEHR, Protégé, and Liferay) either have a free version or are open source, which make them accessible to low-income areas and institutions with limited funding for interoperability projects.

We must also mention the limitations of this study. First of all, the ontology-based clinical repository used in our institution was developed throughout many years, and it might not be a suitable approach for institutions that seek a rapid implementation of a methodology. This can limit the external validity of the study. Moreover, since the tool to insert data from standardized extracts into the ontologies is not ready yet, we still have not completed the creation of OMOP CDM tables. However, an OMOP CDM database for a small subset of
patients was successfully created to test the queries and validate the methodology.

Next Steps
The MIU team at HCB is working on creating the ontological representation of different health information standards (FHIR and OpenEHR) and CDMs (i2b2, International Cancer Genome Consortium Accelerating Research in Genomic Oncology (ICGC Argo) [34], and Clinical Data Interchange Standards Consortium (CDISC) [35]). This will extend the current metamodel and allow us to carry out multistandard transformations, which will also help us compare the performance of such standards for different scenarios.

Conclusions
Semantic interoperability plays a very important role within HISs, providing meaning and clinical context to the clinical information and allowing for better clinical decision-making and research. This study has demonstrated that ontologies constitute a valuable approach for knowledge representation and standardization of health information in a standard-agnostic manner. With the proposed methodology, institutions can go from local raw data to standardized, semantically interoperable EN/ISO 13606 and OMOP repositories.

Acknowledgments
This study is also framed within the Spanish Secretary of State for Telecommunications and Digital Infrastructure’s “Plan de Impulso de las Tecnotencias del Lenguaje” (Plan TL). We would like to thank the Instituto de Salud Carlos III (ISCIII), VeraTech For Health, and Barcelona Supercomputing Center for their collaboration on this project. This work was supported by the ISCIII and cofunded by the European Union (grant PI18/00890, PI18/00981, and PI18/00019).

Conflicts of Interest
None declared.

References


32. infobank. Comunidad de Madrid. URL: https://cpsanidadcm.org/infobanco/ [accessed 2023-02-08]


34. Accelerating research in genomic oncology. ICGC ARGO. URL: https://www.icgc-argo.org/ [accessed 2023-02-08]

35. Standards. CDISC. URL: https://www.cdisc.org/standards [accessed 2023-02-08]

Abbreviations

AM: archetype model
API: application programming interface
BSC: Barcelona Supercomputing Center
Edited by A Benis; submitted 23.11.22; peer-reviewed by R Sánchez de Madariaga, P Azevedo Marques; comments to author 20.12.22; revised version received 28.12.22; accepted 05.01.23; published 08.03.23.

Please cite as:
An Ontology-Based Approach for Consolidating Patient Data Standardized With European Norm/International Organization for Standardization 13606 (EN/ISO 13606) Into Joint Observational Medical Outcomes Partnership (OMOP) Repositories: Description of a Methodology
JMIR Med Inform 2023;11:e44547
URL: https://medinform.jmir.org/2023/1/e44547
doi:10.2196/44547
PMID:36884279

©Santiago Frid, Xavier Pastor Duran, Guillem Bracons Cucó, Miguel Pedrera-Jiménez, Pablo Serrano-Balazote, Adolfo Muñoz Carrero, Raimundo Lozano-Rubí. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 08.03.2023. 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.
Original Paper

Data-Driven Identification of Unusual Prescribing Behavior: Analysis and Use of an Interactive Data Tool Using 6 Months of Primary Care Data From 6500 Practices in England

Lisa EM Hopcroft¹, PhD; Jon Massey¹, PhD; Helen J Curtis¹, DPhil; Brian Mackenna¹, MPharm; Richard Croker¹, MSc; Andrew D Brown¹, MPharm, MPych; Thomas O’Dwyer¹, BSc; Orla Macdonald², MPharm; David Evans¹, MPhil; Peter Inglesby¹, MPhil; Sebastian CJ Bacon¹, BA; Ben Goldacre¹, MRCPsych; Alex J Walker¹, PhD

¹Bennett Institute for Applied Data Science, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom
²Oxford Health NHS Foundation Trust, Oxford, United Kingdom

Corresponding Author:
Alex J Walker, PhD
Bennett Institute for Applied Data Science
Nuffield Department of Primary Care Health Sciences
University of Oxford
Radcliffe Primary Care Building
32 Woodstock Road
Oxford, OX2 6GG
United Kingdom
Phone: 44 01865617855
Fax: 44 01865617855
Email: alex.walker@phc.ox.ac.uk

Abstract

Background: Approaches to addressing unwarranted variation in health care service delivery have traditionally relied on the prospective identification of activities and outcomes, based on a hypothesis, with subsequent reporting against defined measures. Practice-level prescribing data in England are made publicly available by the National Health Service (NHS) Business Services Authority for all general practices. There is an opportunity to adopt a more data-driven approach to capture variability and identify outliers by applying hypothesis-free, data-driven algorithms to national data sets.

Objective: This study aimed to develop and apply a hypothesis-free algorithm to identify unusual prescribing behavior in primary care data at multiple administrative levels in the NHS in England and to visualize these results using organization-specific interactive dashboards, thereby demonstrating proof of concept for prioritization approaches.

Methods: Here we report a new data-driven approach to quantify how “unusual” the prescribing rates of a particular chemical within an organization are as compared to peer organizations, over a period of 6 months (June-December 2021). This is followed by a ranking to identify which chemicals are the most notable outliers in each organization. These outlying chemicals are calculated for all practices, primary care networks, clinical commissioning groups, and sustainability and transformation partnerships in England. Our results are presented via organization-specific interactive dashboards, the iterative development of which has been informed by user feedback.

Results: We developed interactive dashboards for every practice (n=6476) in England, highlighting the unusual prescribing of 2369 chemicals (dashboards are also provided for 42 sustainability and transformation partnerships, 106 clinical commissioning groups, and 1257 primary care networks). User feedback and internal review of case studies demonstrate that our methodology identifies prescribing behavior that sometimes warrants further investigation or is a known issue.

Conclusions: Data-driven approaches have the potential to overcome existing biases with regard to the planning and execution of audits, interventions, and policy making within NHS organizations, potentially revealing new targets for improved health care service delivery. We present our dashboards as a proof of concept for generating candidate lists to aid expert users in their interpretation of prescribing data and prioritize further investigations and qualitative research in terms of potential targets for improved performance.
Introduction

There is recognition that evidence-based decision-making in the National Health Service (NHS) in England is critical to maintaining standards of care while reducing NHS spending [1] and the UK government has recently consulted on wide-ranging plans to “digitize, connect, and transform the health and care sector,” with a key priority being data-driven innovation. Flagship initiatives such as Getting It Right First Time [2] and RightCare [3] focus on identifying and addressing unwarranted variation in the NHS. Such initiatives can be limited in their scope in that the “data-driven” element of the work often focuses on assessing performance relative to recommendations that are defined prospectively rather than employing hypothesis-free data-driven methodologies to make objective assessments as to where opportunities for improvement might exist.

Monthly prescription data for every general practice in England has been made available to the public since 2010, via the NHS Business Services Authority [4]. This data set includes product and month of prescribing, the number of items prescribed and the total quantity, making it very amenable to detailed analysis for the purposes of original research [5-9] and systematic audits and reviews [10,11]. These data, made navigable via interactive dashboards [12,13], are commonly used by NHS staff—in particular, medicines optimization (MO) teams—to monitor performance on key prescribing indicators, compare performance to peer organizations, inform prioritization of work streams, estimate the impact and feasibility of interventions, or create customized outputs according to local priorities. Mining these data systematically for unusual prescribing behavior could help identify where service delivery improvements are possible in the absence of human bias or expectation. Such “unbiased” or “hypothesis-free” approaches might aid local decision makers when designing appropriate interventions and policies.

The value of deploying systematic analyses to large prescribing data sets has been demonstrated elsewhere. Using regional prescribing claims data from Germany, researchers were able to identify practices prescribing more “third-level” medications (ie, not first- or second-line treatments) than expected using funnel plots and mixed effects models [14]. Our own group has successfully deployed similar outlier detection methodology on a national scale to show that the prescribing of 2 antipsychotic drugs, in very limited use nationally, is concentrated in 2 small geographic regions of England [15]. More complex outlier analysis of wholesale codeine time series data has identified significant shifts in supply occurring around the time of regulatory changes (specifically, the up-scheduling of low-dose codeine products from over-the-counter to prescription-only) [16].

We run OpenPrescribing [13], a website that allows public interrogation and visualization of primary care prescription data at multiple administrative levels in the NHS in England. We have previously deployed novel methodologies to identify changes over time in any one of the 80 measures implemented in OpenPrescribing, providing monthly alerts to notify practitioners when their prescribing rates deviate from the norm and may require clinician attention [17]. These measures have been selected on the basis of clear guidance being available from health authorities and are subject to initial and continuing review by clinicians, pharmacists, and epidemiologists. OpenPrescribing has 20,000 unique users every month and thousands of subscribers to our innovative organization email alerts service [17].

We set out to develop new hypothesis-blind data science techniques to identify unusual prescribing behavior, thereby providing proof of concept for such an analysis and illustrating potential opportunities for service improvement. Using this approach, we have no hypothesis with regard to where interesting patterns might be found (ie, which clinical area or which organization), we only have an expectation of what would constitute an interesting pattern in the data. We applied this methodology to 6 months of national prescribing data to identify outliers at multiple administrative levels of the NHS in England during that time period, presenting the most extreme outliers in each organization for the consideration of expert users to prioritize for further review, qualitative research, and interpretation within the local context.

Methods

Study Design

Prescribing practice was analyzed by conducting a retrospective cohort study using prescribing data from all English NHS general practices, primary care networks (PCNs), clinical commissioning groups (CCGs), and sustainability and transformation partnership (STPs; Textbox 1).
Data Source

Data for the period June 1, 2021, to December 1, 2021, were extracted from the OpenPrescribing database; this 6-month study period was used so as to smooth out short-term fluctuations (by aggregating multiple months of data) while keeping to a relatively recent time frame (so that the data remain relevant). OpenPrescribing imports openly accessible prescribing data from the large, monthly files published by the NHS Business Services Authority, which contain data on cost and items prescribed for each month for every typical general practice and CCG in England, dating back to mid-2010 [4,22]. These data are published only at the level of organization; patient-level data are not made available. Detailed methods for the creation of OpenPrescribing, including data management, aggregation, and cleaning, are available elsewhere [23]. The monthly prescribing data sets contain 1 row for each different medication and dose in each prescribing organization in NHS primary care in England, describing the number of items (ie, prescriptions issued) and the total cost. These data are sourced from community pharmacy claims data and, therefore, contain all items that were dispensed. All available prescribing data were extracted for institutions identified as “typical” general practices; all other organizations, such as prisons or specialist community clinics, were excluded using NHS Digital organization data [24]. We limited our analysis to the 2369 chemicals from chapters 1-15 of the British National Formulary (BNF) to exclude chapters not following a chemical and subparagraph structure, those which largely cover nonmedicinal products such as dressings (see Textbox 2 for further information regarding prescribing terminology).

Textbox 2. Prescribing terminology.

- The public prescribing data made available by the National Health Service Business Services Authority uses a pseudo-British National Formulary (BNF) classification. The most granular level of data is at “presentation” level, which includes information on the prescription medicine, brand, strength, and formulation. This data can then be grouped using the pseudo-BNF hierarchy, using products, chemical substances, subparagraphs, paragraphs, sections, and chapters, with decreasing specificity. Chapters are defined according to body system, for example, gastrointestinal system, cardiovascular system, and respiratory system.
- “Chemical” in this context refers to the standard International Nonproprietary Name (INN) for the active constituent of the medicine and does not include any further specification by preparation, dose, or brand. BNF subparagraphs can be used to identify groups of chemicals belonging to the same class.
- The majority of chemicals have all available preparations included in a single chemical definition; for example, atorvastatin preparations (including liquid and tablets) are included in 2.12: Cardiovascular system—lipid-regulating drugs (chemical code: 0212000B0).
- However, there are some instances where the same chemical is used in different body systems with system-specific presentations, and therefore the same INN will appear multiple times in a chapter that most reflects its use. For example, the INN dexamethasone appears in 3 separate chapters within the pseudo-BNF hierarchy and therefore will have separate chemical groupings:

6.3: Endocrine system—corticosteroids (endocrine), which include oral and parenteral preparations (chemical code: 0603020G0)
11.4: Eye—corticosteroids and other anti-inflammatory preparations, which include ocular preparations (chemical code: 1104010I0)
12.1 Ear, Nose, and Oropharynx—drugs acting on the ear, which ear preparations (chemical code: 12101050)

Outlier Detection

We were interested in detecting outliers with regard to chemicals (see Textbox 2 for further information). We first calculate a prescription rate for each chemical in each practice; specifically, we calculate the number of prescriptions containing our chemical of interest and divide this by the number of prescriptions containing chemicals of the same BNF subparagraph, for example, all statin prescriptions as a proportion of all lipid-regulating drugs. This captures the prescribing rate for the chemical of interest as compared to all drugs in the same class in a single practice. This ratio is calculated across all practices, and the mean and SD are calculated. The ratios in each practice are then reexpressed as $z$ scores using this mean and SD. A $z$ score is the number of SDs that a given data point is away from the mean. The $z$ scores are used to rank all chemicals within a practice in terms of their outlier status (the most extreme outliers occupying the top and bottom of this ranked list).

This process is repeated at 3 higher administrative levels—STP, CCG, and PCN—to generate the equivalent ranked list of
prescribing outliers for these larger organizations. Results for all 4 administrative levels are presented, as each organization retains some decision-making power with regard to prescribing. At the practice or PCN level this will be individual or group general practitioner decisions based on their practice population, but MO teams at the STP and CCG level will also monitor prescribing behavior to inform prescribing policy (and formulary) for these wider geographic regions.

**Visualization of Organization-Level Results**

An interactive dashboard has been created at OpenPrescribing [25] for each organization, where data describing 20 of the most extreme outliers are summarized as follows: 10 where prescribing in the organization is higher than other peer organizations, and 10 where prescribing in the organization is lower than other peer organizations. Tables are provided for both sets, which summarize the following values for each chemical: *Chemical Items* and *Subparagraph Items* are the number of prescriptions for the chemical and BNF subparagraph, respectively; *Ratio* is the *Chemical Items* as a proportion of *Subparagraph Items* for the chemical in the organization of interest; *Mean* and *SD* summarize this ratio over all organizations; and *z score* is the *Ratio* reexpressed as a *z* score. This same information is described visually by a density plot (provided in the Multimedia Appendices), where the distribution of ratios across all organizations is captured by a blue line, with the ratio for the organization of interest indicated by a vertical red line. Densities are generated using the Seaborn kdeplot() function, setting the bandwidth for smoothing as suggested by Scott [26].

**User Feedback**

Links to early prototypes were shared directly with a group of interested clinicians and pharmacists by email. Any feedback gained was used to inform the iterative development of the tool and proposed visualizations of the results. Further to this, the tool was shared more widely (via Twitter), and formal feedback was collected via a Google form (Textbox 3). Additional unstructured feedback was compiled from direct emails and mentions on social media.

<table>
<thead>
<tr>
<th>Textbox 3. Outlier detection feedback form.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Respondent details:</strong></td>
</tr>
<tr>
<td>• Email. Free text</td>
</tr>
<tr>
<td>• Which organization’s report are you giving feedback on? Free text</td>
</tr>
<tr>
<td>• Please describe your relationship to the organization (eg, doctor, practice nurse, or commissioner). Free text</td>
</tr>
<tr>
<td><strong>Understandability:</strong></td>
</tr>
<tr>
<td>• Does this report make sense to you? Yes or No</td>
</tr>
<tr>
<td>• Any further comments on the understandability of the report(s). Free text</td>
</tr>
<tr>
<td><strong>Interest:</strong></td>
</tr>
<tr>
<td>• Is it interesting? Yes or No</td>
</tr>
<tr>
<td>• Any further comments on the interestingness of the report(s). Free text</td>
</tr>
<tr>
<td><strong>Utility:</strong></td>
</tr>
<tr>
<td>• Is it useful? Yes or No</td>
</tr>
<tr>
<td>• Any further comments on the usefulness of the report(s). Free text</td>
</tr>
<tr>
<td><strong>Individual items:</strong></td>
</tr>
<tr>
<td>• Thinking about where your prescribing is higher than most, please describe any observations you have on any individual items. Free text</td>
</tr>
<tr>
<td>• Thinking about where your prescribing is lower than most, please describe any observations you have on any individual items. Free text</td>
</tr>
<tr>
<td><strong>Improvements:</strong></td>
</tr>
<tr>
<td>• What, if anything, would you change about the report(s)? Free text</td>
</tr>
</tbody>
</table>

**NHS Devon CCG Case Study Details**

RC (who, in addition to his role at the Bennett Institute, is also Deputy Director for MO at NHS Devon) emailed a link to the dashboard containing sparkline graphs for NHS Devon to pharmacist colleagues in his MO team. These graphs provided new insights to the team, which would have been impractical to achieve using existing data analysis workflows (eg, custom queries in OpenPrescribing or ePACT2). The MO team met to discuss what the causes behind the deviation might be in each case. Where it could not be determined that there was a clinically justifiable reason for being an outlier, the MO team gathered further relevant prescribing data from routine sources such as OpenPrescribing, ePACT2, and PrescQIPP. This allowed deeper exploration of prescribing patterns related to the outlier chemical (eg, trends over time and the rate at which alternative medications were prescribed). The MO team continues to
investigate these data to decide whether an intervention is appropriate.

**Software and Reproducibility**

Data management was performed using Python 3.8.1 and Google BigQuery, with analysis carried out using Python. Code for data management and analysis is archived on the internet [27] and dashboards are available on the OpenPrescribing website [25].

**Patient and Public Involvement**

We publicized this tool via social media and actively sought feedback from interested health care professionals and members of the public to inform its iterative development via a survey (see User Feedback section above). We will continue to seek and consider feedback via these same channels as the tool is developed. We have developed a publicly available website [13] through which we invite any patient or member of the public to contact us regarding this study or the broader OpenPrescribing project.

---

**Table 1.** Summary statistics for z scores calculated for outlying chemicals across the 4 administrative levels. Outlying chemicals are those occurring in the top 10 (ie, “Higher than most”) or bottom 10 (ie, “Lower than most”) by z score in at least one organization at the corresponding administrative level.

<table>
<thead>
<tr>
<th>Organization type</th>
<th>Unique chemicals, n</th>
<th>Higher than most</th>
<th>Lower than most</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Maximum</td>
<td>Median</td>
<td>Minimum</td>
</tr>
<tr>
<td>STP&lt;sup&gt;a&lt;/sup&gt; (n=42)</td>
<td>680</td>
<td>6.33</td>
<td>5.42</td>
</tr>
<tr>
<td>CCG&lt;sup&gt;b&lt;/sup&gt; (n=106)</td>
<td>1138</td>
<td>10.20</td>
<td>5.79</td>
</tr>
<tr>
<td>PCN&lt;sup&gt;c&lt;/sup&gt; (n=1257)</td>
<td>1416</td>
<td>2528.09</td>
<td>5.28</td>
</tr>
<tr>
<td>Practice (n=6476)</td>
<td>1346</td>
<td>6825.50</td>
<td>5.23</td>
</tr>
</tbody>
</table>

<sup>a</sup>STP: sustainability and transformation partnership.

<sup>b</sup>CCG: clinical commissioning group.

<sup>c</sup>PCN: primary care network.

While the median values for the “higher than most” outlying chemicals are similar, the IQR (Q3-Q1) values demonstrate that variation between peer organizations decreases with the size of the organization; the least amount of variation is observed between STPs, and the most amount of variation is observed between practices. More outlying chemicals are identified in smaller organizations (PCNs and practices). With regard to outlying chemicals identified as being prescribed at lower rates compared to peer organizations, both the median and IQR of the z scores are very similar across all organization types. For both sets of outlying chemicals, the most extreme outliers occur further away from the mean as the organization size decreases; the maximum value for the “higher than most” outlying chemicals increases with the size of the organization and the minimum value for the “lower than most” outlying chemicals decreases with the size of the organization. The z scores for “higher than most” outlying chemicals are more extreme than the “lower than most” outlying chemicals in all organization types.

**Results**

**Outlier Detection**

We developed interactive dashboards for every practice in England to highlight unusual prescribing. The outlying chemicals (ie, the 10 chemicals ranked highest and 10 chemicals ranked lowest by z score) identified using our methodology are described in Table 1. Both counts of unique chemicals and summary statistics of z scores are provided at each of the 4 administrative levels. Those outlying chemicals that are “higher than most” will all have positive z and as such are summarized using the maximum, median, Q1 and Q3; similarly, outlying chemicals that are “lower than most” will have negative z scores, and are summarized using the minimum, median, Q1 and Q3. A measure of the variation in the z score amongst all organizations at the same administrative level can be obtained by calculating the Inter Quartile Range (IQR), defined as Q3-Q1.

NHS Devon CCG is the fifth largest CCG in England to highlight unusual prescribing. The outlying chemicals (ie, the 10 chemicals ranked highest and 10 chemicals ranked lowest by z score) identified using our methodology are described in Table 1. Both counts of unique chemicals and summary statistics of z scores are provided at each of the 4 administrative levels. Those outlying chemicals that are “higher than most” will all have positive z and as such are summarized using the maximum, median, Q1 and Q3; similarly, outlying chemicals that are “lower than most” will have negative z scores, and are summarized using the minimum, median, Q1 and Q3. A measure of the variation in the z score amongst all organizations at the same administrative level can be obtained by calculating the Inter Quartile Range (IQR), defined as Q3-Q1.

NHS Devon CCG is the fifth largest CCG in England, commissioning health care for 1.2 million people in the southwest of England. The top 10 chemicals that are prescribed at higher rates here compared to other CCGs are shown in the top portion of Table 2, while the top 10 chemicals that are prescribed at lower rates are shown in the bottom portion of Table 2 (a listing of the specific products and a sparkline plot, showing graphically where the ratio value for this CCG occurs in the context of the same ratio in all CCGs, are provided in Multimedia Appendix 1). These prescribing outliers for this CCG have been reviewed by the local MO team, to provide likely explanations for the outlier prescribing.
Table 2. The outlier detection dashboard for NHS\textsuperscript{a} Devon clinical commissioning group (CCG)\textsuperscript{b}.

<table>
<thead>
<tr>
<th>BNF\textsuperscript{c} chemical (number of products)</th>
<th>Chemical items, n</th>
<th>BNF subparagraph</th>
<th>Subparagraph items, n</th>
<th>Ratio</th>
<th>Mean</th>
<th>SD</th>
<th>z score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prescribing where NHS Devon CCG is higher than most</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levobupivacaine hydrochloride (1)</td>
<td>130</td>
<td>Local anesthetics</td>
<td>20,482</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>10.2</td>
</tr>
<tr>
<td>Gripe mixtures (1)</td>
<td>1</td>
<td>Sodium bicarbonate</td>
<td>56</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>8.34</td>
</tr>
<tr>
<td>Gluten free pastas (3)</td>
<td>4</td>
<td>Foods for special diets</td>
<td>9199</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7.97</td>
</tr>
<tr>
<td>Epoetin zeta (1)</td>
<td>2</td>
<td>Hypoplastic, hemolytic, and renal anemias</td>
<td>18</td>
<td>0.11</td>
<td>0</td>
<td>0.01</td>
<td>7.52</td>
</tr>
<tr>
<td>Flumetasone pivalate (1)</td>
<td>333</td>
<td>Otitis externa</td>
<td>19,724</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>6.98</td>
</tr>
<tr>
<td>Gluten free or wheat free cereals (1)</td>
<td>2</td>
<td>Foods for special diets</td>
<td>9199</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.81</td>
</tr>
<tr>
<td>Levofoxacin (2)</td>
<td>1372</td>
<td>Quinolones</td>
<td>4754</td>
<td>0.29</td>
<td>0.05</td>
<td>0.04</td>
<td>6.61</td>
</tr>
<tr>
<td>Liquefied phenol (1)</td>
<td>1</td>
<td>Phenolics</td>
<td>3</td>
<td>0.33</td>
<td>0.01</td>
<td>0.06</td>
<td>5.83</td>
</tr>
<tr>
<td>Ruxolitinib (1)</td>
<td>2</td>
<td>Other antineoplastic drugs</td>
<td>1494</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5.10</td>
</tr>
<tr>
<td>Ferrous gluconate (1)</td>
<td>10,437</td>
<td>Oral iron</td>
<td>90,095</td>
<td>0.12</td>
<td>0.03</td>
<td>0.02</td>
<td>3.66</td>
</tr>
<tr>
<td><strong>Prescribing where NHS Devon CCG is lower than most</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sodium bicarbonate (3)</td>
<td>55</td>
<td>Sodium bicarbonate</td>
<td>56</td>
<td>0.98</td>
<td>1</td>
<td>0</td>
<td>–8.34</td>
</tr>
<tr>
<td>Ciprofloxacin (6)</td>
<td>2989</td>
<td>Quinolones</td>
<td>4754</td>
<td>0.63</td>
<td>0.86</td>
<td>0.05</td>
<td>–4.22</td>
</tr>
<tr>
<td>Dexamethasone (2)</td>
<td>13,061</td>
<td>Otitis externa</td>
<td>19,724</td>
<td>0.66</td>
<td>0.79</td>
<td>0.05</td>
<td>–2.71</td>
</tr>
<tr>
<td>Fexofenadine hydrochloride (6)</td>
<td>33,711</td>
<td>Antihistamines</td>
<td>169,747</td>
<td>0.20</td>
<td>0.36</td>
<td>0.07</td>
<td>–2.33</td>
</tr>
<tr>
<td>Oral rehydration salts (8)</td>
<td>1942</td>
<td>Oral sodium and water</td>
<td>5010</td>
<td>0.39</td>
<td>0.65</td>
<td>0.11</td>
<td>–2.27</td>
</tr>
<tr>
<td>Betamethasone esters (12)</td>
<td>1672</td>
<td>Topical corticosteroids</td>
<td>141,063</td>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
<td>–2.20</td>
</tr>
<tr>
<td>Fusidic acid (1)</td>
<td>359</td>
<td>Antibacterials</td>
<td>13,283</td>
<td>0.03</td>
<td>0.08</td>
<td>0.03</td>
<td>–2.12</td>
</tr>
<tr>
<td>Senna (9)</td>
<td>39,769</td>
<td>Stimulant laxatives</td>
<td>110,838</td>
<td>0.36</td>
<td>0.55</td>
<td>0.1</td>
<td>–2.03</td>
</tr>
<tr>
<td>Ticagrelor (3)</td>
<td>2285</td>
<td>Antiplatelet drugs</td>
<td>467,104</td>
<td>0</td>
<td>0.02</td>
<td>0.01</td>
<td>–2.02</td>
</tr>
<tr>
<td>Lactulose (2)</td>
<td>19,621</td>
<td>Osmotic laxatives</td>
<td>127,773</td>
<td>0.15</td>
<td>0.28</td>
<td>0.06</td>
<td>–1.89</td>
</tr>
</tbody>
</table>

\textsuperscript{a}NHS: National Health Service.

\textsuperscript{b}The results of our outlier detection methodology are provided as interactive dashboards; here, the 10 chemicals where prescribing in National Health Service (NHS) Devon CCG is higher than most and the 10 chemicals where prescribing in NHS Devon CCG is lower than most, are presented. British National Formulary (BNF) chemical is the chemical of interest (number of products indicates how many products are represented by the BNF chemical). Chemical items provide the number of prescribing items containing this chemical. BNF subparagraph is the BNF subparagraph to which the chemical belongs, and subparagraph items is the number of prescribing items containing an item belonging to this BNF subparagraph. Ratio, Mean, SD, and z score place the chemical items count in the context of the subparagraph items count as described in the Methods section.

\textsuperscript{c}BNF: British National Formulary.

Focusing on the results for flumetasone pivalate, we can see that 1.7% (n=333) of the 19,724 “Otitis externa” items contain flumetasone pivalate and that this is 6.98 SDs above the mean for all CCGs (the sparkline plot provided in Multimedia Appendix 1 demonstrates visually where this 1.7% falls [red line] in the distribution across all CCGs [blue line]).

Several of the chemicals prescribed more often in NHS Devon CCGs than other CCGs are defined as first-line treatments in local formularies, for example, flumetasone pivalate [28] and levofoxacin [29]. Corresponding patterns of underprescribing can be seen in the “lower than most” results table for similar chemicals, specifically, ciprofloxacin (an alternative to levofoxacin) and dexamethasone (an alternative to flumetasone pivalate).

The lower prescribing rates for fusidic acid reflect a change in this CCG to prescribe this chemical by specialist recommendation only [30], due to rising costs [31] and a narrow spectrum of action. The lower rates of prescribing for senna and lactulose are also likely due to a formulary shift in this CCG toward macrogols [32]. Finally, the low prescribing rate of betamethasone esters is also expected as these chemicals are nonformulary in this CCG [33].

This dashboard also demonstrates a valid use for low-number results. Gluten-free pastas and cereals—something that we have previously identified as having high variability in prescribing...
rates [34]—were not recommended to be prescribed by the NHS in the study period (NHS England issued advice to CCGs in November 2018 with the recommendation to restrict gluten free prescribing to bread and flour mixes [35]), so should not appear at all. The identification of this low-number outlier via our methodology has prompted further work within NHS Devon CCG to clarify how this prescription was generated and processed.

User Feedback

Through the formal Google form and direct correspondence with interested parties, we received feedback for a prototype version of the dashboard from 6 individuals. An example of this prototype is shown in Multimedia Appendix 2, showing 5 top and bottom outlying chemicals. Several respondents indicated that the results were expected (ie, results echoed internal reporting or were aligned with local prescribing policies); while this indicates that our tool is working, 1 user did question what the added value was above existing reporting. Other users stated that the tool had revealed unexpected results worthy of follow-up.

There were multiple requests to present more than the top and bottom 5 results (eg, the top and bottom 10 or 20 results) to explore the data in more detail. Users recognized that extreme outliers could be derived from very small numbers of patients or items; some requested that results with small counts be removed, though others recognized that these may be important, particularly in practices or PCNs. There was a suggestion that users could choose to have low numbers suppressed or displayed, depending on whether their focus was systemic anomalies or rogue prescriptions. There were also requests to include other data in the results, including cost and highlighting drugs on the “Not suitable to prescribe” list.

There were other requests that were more relevant to the design of the tool than the analysis itself. The feedback demonstrated that users required more information to interpret and understand the data (ie, z scores, ratios, means, and SDs) and that with this additional explanation, more could be made of the graphical summary. There was also a request for an improved user experience regarding navigating to practices via the drop-down sections (which could be implemented as an organizational search).

We used the most common feedback to inform further development, and the released version of the dashboards now includes the top and bottom 10 outlying chemicals and optional filtering of low numbers. To provide a clear illustration of how the dashboards changed in response to user feedback, the corresponding update for Multimedia Appendix 2 is shown in Figures S1 and S2 in Multimedia Appendix 3.

Discussion

Summary

We have developed and implemented a new hypothesis-free methodology to detect unusual or “outlier” prescribing rates of chemicals in a single organization in relation to all “peer” organizations. We have applied this methodology to 6 months of national prescribing data to quantify how typical the prescribing is for individual chemicals at multiple administrative levels (practice, PCN, CCG, and STP) over the time period. We have displayed these results via interactive dashboards. We have sought and will continue to seek user feedback to inform development and incrementally improve usability and functionality.

Summary statistics demonstrate that the number of outlying chemicals increases as the size of the organization decreases and that more extreme outliers are identified among smaller organizations, demonstrating that there is more variability in prescribing behavior among practices than there is among larger administrative organizations. The data also demonstrate, however, that outliers do occur when comparing larger organizations to each other. While there is less variation between STPs, the median z score for “higher than most” and “lower than most” outliers among STPs is 5.42 and 2.35, respectively; these z scores are both more than 2 SDs from the mean. The ranking of these quantifications allows us to identify the most extreme outliers in terms of prescribing behavior at each organizational level. A case study of an individual CCG (NHS Devon) demonstrated that our methodology identified prescribing patterns that aligned with local prescribing guidance, but also detected patterns that warranted further investigation. It is not appropriate to formally assess the utility of our methodology as there are many legitimate reasons that a chemical may be an outlier in a particular organization. Some of the reasons are as follows: prescribing guidance as defined by local formulary may differ from elsewhere; local prescribing policy may place responsibility for prescribing particular drugs in secondary care rather than primary care; clinicians may be reluctant to change medication for patients who are stable on a long-established medication regime (in particular the elderly or vulnerable); or there is a justified preference for other drugs in the same class. Given the complexities of interpreting these data, we present this tool as a proof of concept and starting point for NHS organizations to perform and plan internal audits rather than a definitive reporting tool.

Strengths and Weaknesses

Our approach combines a comprehensive national prescribing data set with a well-understood system for drug classification, thereby capturing the national context at high resolution and allowing the interpretation of prescribing behavior for all chemicals at multiple administrative levels of the NHS in England, all of which retain some decision-making power with regard to prescribing. The methods used are well established and easy to understand, readily amenable to visual presentation as graphs, and allow prioritization of results by ranking. Our approach has utility in other contexts, and repurposing it to gain a greater understanding of other NHS data (eg, hospital prescriptions) would be straightforward.

We also note some limitations. First, the calculation of z scores using mean and SD assumes a normal distribution. This is more likely to be the case where numbers of items prescribed are high (aggregated to STP or CCG), but may not be the case where number of items are low (aggregated to PCN or practices, or where the items are more rarely prescribed). Second, this approach can generate very large z scores where SDs are tight.
or item numbers generally are very low. An example of this can be seen in Figure S1 in Multimedia Appendix 3; while the ratio generated by the number of prescribed items containing Sodium aurothiomalate is very low (1/114,367=8.74×10^{-6}), the tight SDs observed across the whole population of STPs translate this small value into a large z score. Expert users may be seeking out such results to identify very rare prescribing items (low number results did prove important in the NHS Devon case study), but they may also wish to suppress such results to focus on more commonly prescribed chemicals. To accommodate this and in line with our user feedback, we have implemented the option to show or hide counts of 5 or less. We also recognize that the process by which we have sought user feedback thus far could be prone to bias, in that specific users were targeted due to their expertise and familiarity with such tools so as to enable rapid development.

Findings in Context

This is one of a suite of tools that we are seeking to develop at OpenPrescribing, each of which captures variability with a view to leveraging further insight from the data sets to which we have access. We make extensive use of decile plots to place individual organizations into a wider context [5,6,36] and have applied algorithms to identify when those individual organizations start to deviate from the rest of their peers [17]. We have also used deciles to summarize financial data and estimate potential savings if “price-per-unit” costs were aligned with the lowest decile [7]. These methodologies all have the potential to support NHS organizations in England to guide audits, prioritize and shape new policies, and crucially assess the impact of those interventions with regard to patient care and cost savings.

Policy Implications and Interpretation

The Department of Health and Social Care consultation explicitly recognizes the value of near real-time data release and the potential of data-driven insights to guide targeted policy making [37]. The methodology described here contributes toward that key priority by exposing specific patterns in data that warrant attention that may have otherwise been obscured.

We do not advocate that our approach be used in isolation, but rather as a starting point for expert users to interpret within the local context and make evidence-based decisions about priorities and planning. By updating these dashboards on a regular basis, we hope to provide decision makers with near real-time feedback so as to monitor performance and respond quickly when necessary. Comprehensive coverage of the opportunities and challenges that exist in encouraging widespread adoption of these approaches across the NHS in England is provided in the Goldacre Review [38].

Future Research

Areas for further research include the implementation of a systematic and unbiased approach to collecting and inviting user feedback, enhancing results output as determined by ongoing user feedback (eg, new functionality, information, or visualizations), updating the dashboards in line with recent structural changes to the NHS in England (specifically, Integrated Care Boards replacing STPs), and consulting with patient and public involvement and engagement groups to maximize value for the patient community. The long-term aim is to incorporate regular updates as part of an organization’s page on the OpenPrescribing website; the frequency of these updates (annual vs monthly) and the extent to which historical dashboards would be available for each organization have yet to be determined but would be a focus of the enhanced user consultation described above. Ultimately, our aim would be to provide organization specific alerts to notify staff where prescribing behavior appears to be different to their peers.

Conclusions

Capturing the variability in prescribing rates among peer organizations permits the hypothesis-free identification of prescribing outliers. We have applied such an analysis to 6 months of national prescribing data and made the most extreme prescribing outliers in each organization publicly available as interactive dashboards. We intend that these dashboards prompt further qualitative analysis within the individual organizations to identify where service delivery improvements could be made.

Acknowledgments

We are grateful to wider NHS colleagues for discussions that have informed our work on this topic. Conceptualization was done by HC, BM, BG, and AJW. JM, DE, PI, and SB were involved in data curation. LEMH, JM, BM, and AJW performed formal analysis. HC, BG, and AJW were responsible for funding acquisition. LEMH, JM, HC, BM, RC, OM, and AJW conducted investigation. LEMH, JM, HC, BM, RC, and AJW carried methodology. DE, PI, SB, and TOD were involved with the resources. JM, DE, PI, SB, and TOD were responsible for software. Supervision was done by BG. LEMH, JM, HC, BM, and AJW were responsible for visualization. LEMH, JM, HC, BM, and AJW were responsible for writing original draft. LEMH, JM, HC, BM, RC, OM, BG, and AJW were responsible for writing review and editing. BG is the guarantor. This project is funded by the National Institute for Health Research (NIHR) under its Research for Patient Benefit (RfPB) Programme (grant PB-PG-0418-20036). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. Funders had no role in the study design, collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication.

Conflicts of Interest

All authors have completed the International Committee of Medical Journal Editors’ uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare the following: BG has received research funding from the Laura and John Arnold Foundation, the NHS National Institute for Health Research (NIHR), the NIHR School of Primary Care Research, the NIHR...
Oxford Biomedical Research Centre, the Mohn-Westlake Foundation, NIHR Applied Research Collaboration Oxford and Thames Valley, the Wellcome Trust, the Good Thinking Foundation, Health Data Research UK, the Health Foundation, the World Health Organization, UKRI, Asthma UK, the British Lung Foundation, and the Longitudinal Health and Wellbeing strand of the National Core Studies program; he also receives personal income from speaking and writing for lay audiences on the misuse of science. NJD and JM have been employed on grants received by BG. JM is the recipient of a doctoral studentship from the Wellcome Trust.

Multimedia Appendix 1
The outlier detection dashboard for Devon CCG (including product listings and sparkline plots). The results of our outlier detection methodology are provided as interactive dashboards; here the ten chemicals where prescribing in NHS Devon CCG is higher than most and the ten chemicals where prescribing in NHS Devon CCG is lower than most are presented. Data for each result is highlighted in grey with additional information provided below with no highlighting. BNF Chemical is the chemical of interest (all products represented by this BNF chemical are provided as additional information). Chemical Items provides the number of prescribing items containing this chemical. BNF Subparagraph is the BNF Subparagraph to which the Chemical belongs and Subparagraph Items is the number of prescribing items containing an item belonging to this BNF Subparagraph. Ratio, Mean, std and Z-score place the chemical items count in the context of the subparagraph items count as described in the methods. The sparkline plot provided as additional information for each result shows where the Ratio value for this CCG occurs (vertical red line) in the context of the same Ratio in all CCGs (summarised by the blue line). The y axis is density (see Methods).

[DOCX File, 119 KB - medinform_v11i1e44237_app1.docx ]

Multimedia Appendix 2
Prototype dashboard showing the top and bottom five outlying chemicals for Cumbria and northeast STP. BNF Chemical is the chemical of interest, Chemical Items provides the number of prescribing items containing this chemical. BNF Subparagraph is the BNF Subparagraph to which the Chemical belongs and Subparagraph Items is the number of prescribing items containing an item belonging to this BNF Subparagraph. Ratio, Mean, std, Z_Score place the chemical items count in the context of the subparagraph items count as described in the methods. The sparkline plot shows where the ratio value for this STP occurs (vertical red line) in the context of the same ratio in all STPs (summarised by the blue line).

[DOCX File, 419 KB - medinform_v11i1e44237_app2.docx ]

Multimedia Appendix 3
Example dashboard showing the top ten outlying chemicals for Cumbria and northeast STP. BNF Chemical is the chemical of interest, Chemical Items provides the number of prescribing items containing this chemical. BNF Subparagraph to which the chemical belongs and Subparagraph Items is the number of prescribing items containing an item belonging to this BNF Subparagraph. Ratio, Mean, std and Z score place the chemical items count in the context of the subparagraph items count as described in the methods. The sparkline plot shows where the ratio value for this STP occurs (vertical red line) in the context of the same Ratio in all STPs (summarised by the blue line). Example dashboard showing the bottom ten outlying chemicals for Cumbria and northeast STP. See Figure S1 for definitions of each column.

[DOCX File, 1142 KB - medinform_v11i1e44237_app3.docx ]

References
2. Getting It Right First Time (GIRFT). URL: https://www.gettingitrighthetime.co.uk/ [accessed 2022-01-27]

https://medinform.jmir.org/2023/1/e44237 JMIR Med Inform 2023 | vol. 11 | e44237 | p.562 (page number not for citation purposes)


12. ePACT2. NHS Business Services Authority. URL: https://www.nhsbsa.nhs.uk/access-our-data-products/epact2 [accessed 2023-02-08]


28. 11.3.1 Antibacterials. North & East Devon Formulary and Referral. URL: https://northeast.de

29. 5.1.12 Quinolones. North & East Devon Formulary and Referral. URL: https://southwest.de

30. 12.1.1 Otitis externa. North & East Devon Formulary and Referral. URL: https://northeast.de

31. Drug tariff for Tariff prices for Fusidic acid 1% modified-release eye drops. URL: https://openprescribing.net/tariff

32. Management of constipation in adults. North & East Devon Formulary and Referral. URL: https://northeast.de

33. 12.1.1 Otitis externa. North & East Devon Formulary and Referral. URL: https://northeast.de


Abbreviations

BNF: British National Formulary
CCG: clinical commissioning group
MO: medicines optimization
NHS: National Health Service
PCN: primary care network
STP: sustainability and transformation partnership

©Lisa EM Hopcroft, Jon Massey, Helen J Curtis, Brian Mackenna, Richard Croker, Andrew D Brown, Thomas O'Dwyer, Orla Macdonald, David Evans, Peter Inglesby, Sebastian CJ Bacon, Ben Goldacre, Alex J Walker. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.04.2023. 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.
Original Paper

Chinese Clinical Named Entity Recognition From Electronic Medical Records Based on Multisemantic Features by Using Robustly Optimized Bidirectional Encoder Representation From Transformers Pretraining Approach Whole Word Masking and Convolutional Neural Networks: Model Development and Validation

Weijie Wang¹, MS; Xiaoying Li¹, PhD; Huiling Ren¹, MS; Dongping Gao¹, PhD; An Fang¹, MS
Institute of Medical Information and Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

Corresponding Author:
Huiling Ren, MS
Institute of Medical Information and Library
Chinese Academy of Medical Sciences & Peking Union Medical College
69 Dongdan N St
Beijing, 100005
China
Phone: 86 010 52328911
Email: ren.huiling@imicams.ac.cn

Abstract

Background: Clinical electronic medical records (EMRs) contain important information on patients’ anatomy, symptoms, examinations, diagnoses, and medications. Large-scale mining of rich medical information from EMRs will provide notable reference value for medical research. With the complexity of Chinese grammar and blurred boundaries of Chinese words, Chinese clinical named entity recognition (CNER) remains a notable challenge. Follow-up tasks such as medical entity structuring, medical entity standardization, medical entity relationship extraction, and medical knowledge graph construction largely depend on medical named entity recognition effects. A promising CNER result would provide reliable support for building domain knowledge graphs, knowledge bases, and knowledge retrieval systems. Furthermore, it would provide research ideas for scientists and medical decision-making references for doctors and even guide patients on disease and health management. Therefore, obtaining excellent CNER results is essential.

Objective: We aimed to propose a Chinese CNER method to learn semantics-enriched representations for comprehensively enhancing machines to understand deep semantic information of EMRs by using multisemantic features, which makes medical information more readable and understandable.

Methods: First, we used Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking (RoBERTa-wwm) with dynamic fusion and Chinese character features, including 5-stroke code, Zheng code, phonological code, and stroke code, extracted by 1-dimensional convolutional neural networks (CNNs) to obtain fine-grained semantic features of Chinese characters. Subsequently, we converted Chinese characters into square images to obtain Chinese character image features from another modality by using a 2-dimensional CNN. Finally, we input multisemantic features into Bidirectional Long Short-Term Memory with Conditional Random Fields to achieve Chinese CNER. The effectiveness of our model was compared with that of the baseline and existing research models, and the features involved in the model were ablated and analyzed to verify the model’s effectiveness.

Results: We collected 1379 Yidu-S4K EMRs containing 23,655 entities in 6 categories and 2007 self-annotated EMRs containing 118,643 entities in 7 categories. The experiments showed that our model outperformed the comparison experiments, with $F_1$-scores of 89.28% and 84.61% on the Yidu-S4K and self-annotated data sets, respectively. The results of the ablation analysis demonstrated that each feature and method we used could improve the entity recognition ability.

Conclusions: Our proposed CNER method would mine the richer deep semantic information in EMRs by multisemantic embedding using RoBERTa-wwm and CNNs, enhancing the semantic recognition of characters at different granularity levels.
and improving the generalization capability of the method by achieving information complementarity among different semantic features, thus making the machine semantically understand EMRs and improving the CNER task accuracy.

(JMIR Med Inform 2023;11:e44597) doi:10.2196/44597

KEYWORDS
Chinese clinical named entity recognition; multisemantic features; image feature; Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking; RoBERTa-wwm; convolutional neural network; CNN

Introduction

Background
Abundant medical data have been accumulated since the development of the hospital information system, among which the electronic medical records (EMRs) contain information closely related to patients’ diagnosis and treatment processes [1]. As important records of patients’ medical activities, effective extraction and use of the medical information contained in EMRs could provide clinical decision-making support for doctors and realize personalized medical guidance and health management for patients. It could also help biomedical researchers discover the tacit medical knowledge, thus providing ideas for studies of the association between diseases, the relationship between symptoms, the prediction of diseases and therapies, complication prediction, comorbidity analysis, etc. The medical information would be rapidly extracted from the unstructured EMRs through named entity recognition (NER). NER is a basic task of natural language processing, which will lay the foundation for the construction of medical knowledge graphs, medical knowledge bases, and so on by steps such as medical entity structuring, medical entity standardization, and medical entity relationship extraction. It will also provide fundamental support for practical application scenarios such as medical knowledge retrieval systems, clinical decision support systems, clinical event extraction, and so on [2,3].

Clinical NER (CNER) refers to the recognition of entities such as anatomy, disease, symptoms, clinical examination, medication, surgical procedure, and so on from EMRs [4,5]. Chinese CNER is more difficult than English NER for several reasons. First, Chinese words lack space segmentation and have blurred boundaries. Second, the composition of a Chinese entity is complex and may contain various figures, letters, and abbreviations. Third, Chinese grammar is complicated, and the same word may represent different entity types in different contexts. Therefore, Chinese CNER remains a research focus.

Recently, the features of radicals for Chinese characters have been widely used to improve the efficiency of different Chinese natural language processing tasks [6-8]. Chinese characters, known for thousands of years, are highly developed morpheme scripts that are still used worldwide with unique ideology [9]. Chinese characters include single-component and multiple-component characters. A single-component character cannot be divided, for example, “心” (heart), “手” (hand), and “口” (mouth), and so on; whereas a multiple-component character is composed of basic components, accounting for >90% of Chinese characters [10]. For example, the radical for “肿 (swelling)” and “胀 (swelling)” is “月 (month),” which refers to meat or organs in ancient times. Chinese characters are divided into associative compound characters, indicative characters, pictographic characters, and picto-phonetic characters based on their characteristics. In addition, Chinese characters are also called square characters, as they are square, and there are 8 structures of Chinese characters that are subdivided based on their intrinsic shape and construction. Therefore, Chinese characters contain rich deep semantic information. Applying radicals, phonological codes, shape structures, and other features would help to improve Chinese CNER accuracy.

The contributions of this study are as follows: (1) using pretrained language model (PLM) Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking (RoBERTa-wwm) with a dynamic fusion transformer layer to obtain the semantic features of Chinese characters; (2) using CNNs for extracting the radicals and picto-phonetic features of Chinese characters through the 5-stroke code, Zheng code, phonological code, and stroke code; (3) converting Chinese characters into square images, extracting Chinese character image features from another modality by CNNs, and deeply capturing the pictographic characteristics of Chinese characters; and (4) improving the semantic recognition ability of the model at different levels of granularity, achieving information complementarity between different semantic features, and improving the effect and generalization ability of the model based on multisemantic features.

Related Works

Medical NER

In recent decades, the medical NER is still a research focus. Medical NER research has 3 main development stages as follows: based on dictionaries and rules, based on statistical machine learning, and based on deep learning.

The dictionary-based [11-13] methods need to construct a domain dictionary in advance to achieve medical NER by matching algorithms. The accuracy of this method is relatively higher. However, it may be affected by the large number, strong specialization, and high complexity of Chinese medical entities. In addition, medical terminologies are updated quickly with the rapid development of the medical field, and the lack of new terminologies will also affect medical NER accuracy. The rule-based [14,15] methods need experts in a particular field to formulate the rule templates based on information such as context grammar and structure. However, the rules are poorly universal in different fields. The methods based on dictionaries and rules [16-18] are poorly generalized, time-consuming, and
objective, as much time and labor are required. Therefore, many scholars have gradually applied methods based on statistical machine learning on medical NER. The commonly used methods include maximum entropy [19], support vector machine [20, 21], hidden Markov model [22, 23], and conditional random fields (CRF) [24, 25]. However, these methods rely on large-scale annotation data sets [26] and manual feature selection [13, 27]. Moreover, the quality of the selected features will directly affect the medical NER results.

With the continuous development of deep learning, Cocos et al [28] found that deep learning has advantages over traditional machine learning. It can automatically extract the characteristics of various levels and reduce the subjectivity of artificial feature selection. This thereby improves the result accuracy. The commonly used deep learning models include convolutional neural networks (CNNs) [29], recurrent neural networks [30], long short-term memory (LSTM) [31], Word to Vector (Word2Vec) [32], Bidirectional Encoder Representation from Transformers (BERT) [33], and so on. However, fully extracting the data features by using a neural network alone is challenging. Most scholars took Long LSTM-CRF as the main framework to make up for the medical NER deficiency using a single neural network [34]. The Bidirectional LSTM-CRF (BiLSTM-CRF) [35] model was then developed. This model could better capture contextual information as an important milestone in medical NER and has been widely used in the medical field [36, 37]. To improve the ability to capture details and extract features of medical NER models, many studies added Word2Vec with static representation [38], Global Vectors for Word Representation [39] with static representation, Embeddings from Language Models (ELMo) [40, 41] with dynamic representation, CNN [42], and attention mechanism [43] to the BiLSTM-CRF model. Some studies [44, 45] have shown that the application of the BiLSTM-CRF model combined with the word vector generated by BERT could significantly improve medical NER accuracy. BERT provided a more accurate word representation and achieved better task results than traditional word vector methods. As per the specialty of medicine and the characteristics of Chinese characters involved, the clinical dictionaries, root-level features, parts of speech, radicals, and phonological codes have been added in the BiLSTM-CRF model in some studies [46-51] for improving Chinese CNER performance.

**PLMs Technique**

PLMs are pretrained on a large-scale corpus to obtain prior semantic knowledge from unlabeled text and improve the effectiveness of different downstream tasks. The word vector generated by a bidirectional language model BERT with stacked transformer substructures contains not only the preliminary information from the corpus training but also the encoded contextual information. Some robust versions of BERT have been constructed since BERT was proposed in 2018. For example, the RoBERTa model [52], which replaces the static (MASK) strategy with a dynamic (MASK) strategy, and the words (MASK) in each sequence dynamically change in different epoch trainings. In addition, the RoBERTa model is retrained with bigger batches and longer sequences, and the next-sentence prediction task, which is not related to the downstream task, is canceled during the pretraining. Compared with the BERT model, the RoBERTa model performs better on multiple natural language processing tasks. However, the character-level RoBERTa model does not fit the Chinese natural language processing, as the different segmentation modes between Chinese and English words suffer a limitation of lacking word information. Then, the word-level RoBERTa-wm model [53] was proposed based on Chinese characteristics, which greatly improved the text representation ability in Chinese [54].

**Methods**

**Data Collection**

The Yidu-S4K data set, shared publicly by YiduCloud, is derived from the Chinese EMRs entity recognition task of the China Conference on Knowledge Graph and Semantic Computing 2019 [55]. It contains 1379 EMRs with 6 entity types, including Disease (medically defined disease and diagnoses made by physicians based on etiology, pathophysiology, pathological classification, and clinical staging); Anatomy (anatomical parts of the body where disease, signs, and symptoms occurred); Laboratory (physical or chemical tests performed by the laboratory department in clinical work); Image (imaging [x-ray, computed tomography, magnetic resonance imaging, positron emission tomography-computed tomography, etc], ultrasound, and electrocardiogram); Medicine (specific chemical substances used for disease treatment); and Operation (treatments focused on surgery such as excision and suturing performed by the physician locally on the patient’s body).

Self-annotated EMR data, collected from publicly desensitized Chinese EMR websites [56], contain 2007 EMRs. As per the Terminology of Clinical Medicine issued by the National Health Commission of the People’s Republic of China, we used the BIO (B signifies the beginning of an entity, I signifies that the word is inside an entity, and O signifies that the word is just a regular word outside of an entity) tagging method to pretag 7 entity types in the EMRs, including Disease (same definition as the Yidu-S4K data set); Symptoms (abnormal manifestations as perceived by the sensory organs of patients and physicians); Anatomy (same definition as the Yidu-S4K data set); Examination (includes imaging examinations and laboratory tests mentioned in the Yidu-S4K data set); Instrument (apparatus and mechanical equipment for disease prevention, diagnosis, treatment, health care, and rehabilitation); Medicine (the same definition as the Yidu-S4K data set); and Operation (same definition as the Yidu-S4K data set). Subsequently, 4 medical experts manually checked and corrected the tags. The interclass correlation efficient consistency test revealed that we had good annotation quality.

The ratio of the training set to the test set of the EMRs was 7:3. The Yidu-S4K data set was preprocessed with 1000 EMRs as the training data sets (1000/1379, 72.52%) and 379 EMRs as the test data sets (379/1379, 27.48%). The self-annotated data set was divided by randomization into 1401 EMRs as the training data sets (1401/2007, 69.81%) and 379 EMRs as the testing data sets.
test data sets (606/2007, 30.19%). Table 1 lists the details of the different types of entities in the 2 EMR data sets.

<table>
<thead>
<tr>
<th>Data sets and entity type</th>
<th>Test set, n</th>
<th>Training set, n</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Yidu-S4K</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease</td>
<td>4212</td>
<td>1323</td>
</tr>
<tr>
<td>Anatomy</td>
<td>8426</td>
<td>3094</td>
</tr>
<tr>
<td>Laboratory</td>
<td>1195</td>
<td>590</td>
</tr>
<tr>
<td>Image</td>
<td>969</td>
<td>348</td>
</tr>
<tr>
<td>Medicine</td>
<td>1822</td>
<td>485</td>
</tr>
<tr>
<td>Operation</td>
<td>1029</td>
<td>162</td>
</tr>
<tr>
<td>All entities</td>
<td>17,653</td>
<td>6002</td>
</tr>
</tbody>
</table>

| **Self-annotated**        |             |                 |
| Disease                   | 9470        | 4504            |
| Symptoms                  | 26,334      | 11,065          |
| Anatomy                   | 17,877      | 7588            |
| Examination               | 19,664      | 8746            |
| Instrument                | 1244        | 560             |
| Medicine                  | 5314        | 2566            |
| Operation                 | 2578        | 1133            |
| All entities              | 82,481      | 36,162          |

**Ethical Considerations**
Ethics approval was not required because the patient’s private information was masked by the website.

**Experiments Settings**
In this study, all the experiments were conducted by Python [57] and PyTorch [58]. Table 2 shows the experimental parameters. The experiments used RoBERTa-wmm-ext-large model pretraining data, optimized parameters using Adam W, dropout to prevent overfitting, the batch size of 32, BiLSTM hidden layer dimension of 768, maximum sequence length of 510, RoBERTa-wmm dimension of 768, semantic feature dimension of 124, and image feature dimension of 128. On 2 Chinese CNER data sets, we used the same parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dropout</td>
<td>0.5</td>
</tr>
<tr>
<td>Epoch</td>
<td>Optimization</td>
</tr>
<tr>
<td>Optimization</td>
<td>Adam W</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.0001</td>
</tr>
<tr>
<td>Batch size</td>
<td>32</td>
</tr>
<tr>
<td>BiLSTM(^a) hidden layer</td>
<td>768</td>
</tr>
<tr>
<td>Max_len</td>
<td>510</td>
</tr>
<tr>
<td>RoBERTa-wmm(^b) feature dimension</td>
<td>768</td>
</tr>
<tr>
<td>Semantic feature dimension</td>
<td>124</td>
</tr>
<tr>
<td>Image feature dimension</td>
<td>128</td>
</tr>
</tbody>
</table>

\(^a\)BiLSTM: Bidirectional Long Short-Term Memory.

\(^b\)RoBERTa-wmm: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking.
Evaluation Metrics

The experiments used precision, recall, and $F_1$-score to evaluate the model performance. The formulas for each index are as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (1)
\]
\[
\text{Recall} = \frac{TP}{TP + FN} \quad (2)
\]
\[
F_1\text{-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (3)
\]

where precision is the proportion of positive samples in all samples predicted to be positive; recall is the proportion of positive samples in all positive samples; $F_1$-score is the harmonic mean of precision and recall; true positive (TP) is the number of positive samples predicted to be positive, that is, the number of correctly recognized entities; false positive (FP) is the number of negative samples predicted to be negative, that is, the number of incorrectly recognized other texts as entities; and false negative (FN) is the number of positive samples predicted to be negative, that is, the number of unrecognized entities.

Model Overview

In this study, we proposed a CNER model based on multisemantic features, as shown in Figure 1. First, we used RoBERTa-wwm, the PLM, to obtain the embedded representation at the word level. Dynamic fusion is performed on the semantic representation generated by each transformer layer to make full use of RoBERTa-wwm representation information. Then, the embedded Chinese character fine-grained feature representation, including the 5-stroke code, Zheng code, phonological code, and stroke code, is extracted by 1D CNN, whereas the embedded Chinese character image representation is extracted from another modality by 2D CNN, with the Chinese characters as square images. Finally, the above multisemantic vectors were input into the BiLSTM layer for encoding and were decoded in the CRF layer to predict the tag probability.

Figure 1. The main architecture of our model. 1D CNN: 1D convolutional neural network; 2D CNN: 2D convolutional neural network; B-DIS: beginning of disease entity; CRF: conditional random fields; h: embedding of output character; I-DIS: inside of disease entity; LSTM: long short-term memory; O: other type; RoBERTa-wwm: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking; x: embedding of input character.
Multisemantic Embedding Layer

Overview

Many Chinese characters have retained their original connotations, as they originated from pictographic characters in ancient times. Moreover, the inherent fine-grained character information contained in Chinese characters often implies more additional semantic information. Accordingly, we obtained the 5-stroke code, Zheng code, phonological code, and stroke code information, as shown in Table 3, of the Chinese characters from ZDIC [59] and embedded them in the model. In addition, Chinese characters are squares, and different shapes and structures express different types of information. Characters with similar intrinsic characteristics may have similar meanings. Therefore, we took Chinese characters as graphics and obtained semantic information on Chinese character connotations from another modality. Multisemantics could obtain information comprehensively and learn a better feature representation by making use of information complementarity and eliminating the redundancy among different semantic features compared with a single-semantic feature, resulting in a more generalized model.

As shown in Figure 2, we converted the Chinese character 5-stroke code, Zheng code, phonological code, and stroke code into one-hot vector encoding and interpreted the Chinese characters as 14x14 images. Subsequently, we used a 2-layer CNN deeply extracting the Chinese character multisemantic features. Through the Convolution layer with the ReLU activation function, max pooling layer, and dense layer, we obtained the multisemantic vectors that could be embedded in the BiLSTM layer.

Table 3. Example of Chinese characters' coded information from ZDIC.

<table>
<thead>
<tr>
<th>Character</th>
<th>5-stroke code</th>
<th>Zheng code</th>
<th>Phonological code</th>
<th>Stroke code</th>
</tr>
</thead>
<tbody>
<tr>
<td>呕 (vomit)</td>
<td>kaqy</td>
<td>jhos</td>
<td>du</td>
<td>2,511,345</td>
</tr>
<tr>
<td>吐 (vomit)</td>
<td>kfg</td>
<td>jbvv</td>
<td>tū</td>
<td>251,121</td>
</tr>
<tr>
<td>肿 (swelling)</td>
<td>ekhh</td>
<td>qii</td>
<td>zhêng</td>
<td>35,112,512</td>
</tr>
<tr>
<td>胀 (swelling)</td>
<td>etay</td>
<td>qch</td>
<td>zhàng</td>
<td>35,113,154</td>
</tr>
<tr>
<td>心 (heart)</td>
<td>nny</td>
<td>wz</td>
<td>xīn</td>
<td>4544</td>
</tr>
<tr>
<td>手 (hand)</td>
<td>rtgh</td>
<td>md</td>
<td>shǒu</td>
<td>3112</td>
</tr>
</tbody>
</table>

Figure 2. The process of obtaining Chinese character multisemantic features by convolutional neural network. ReLU: Rectified Linear Unit function; Conv 1: first convolutional layer; Conv 2: second convolutional layer; Max pooling 1: first max pooling layer; Max pooling 2: second max pooling layer; Dense: dense layer.

RoBERTa-wwm With Dynamic Fusion

When RoBERTa-wwm pretrains the corpus, it is segmented on the language technology platform established by the Harbin Institute of Technology based on Wikipedia content in Chinese, which can provide a basis for achieving wwm. As shown in Figure 3, the word “支气管” (bronchi) in the RoBERTa-wwm model is completely masked by random wwm, whereas only single characters can be randomly masked in the BERT model, for example, only 1 character “气” (gases) was masked in the word “支气管 (bronchi).” Thus, the RoBERTa-wwm model can learn the word-level semantic representations in Chinese.

The encoder structure of each transformer layer of the BERT model outputs had different abstract representations of grammar, semantics, and real knowledge in sentences. Studies have confirmed that each layer of the BERT model represents text information differently through 12 natural language processing tasks [60]. As shown in Figure 4, the low transformer mainly learns and encodes surface features; the middle transformer
learns and encodes syntactic features; and the high transformer learns and encodes semantic features.

The transformer structure of the RoBERTa-wwm model is consistent with that of the BERT model. To make full use of the representation information of each transformer layer, we used the RoBERTa-wwm model with dynamic fusion [61]. This helped in assigning the initial weight to the representation vector of 12 transformer layers, determining the weight during training, and weighing the representation vector generated by each layer.

**Figure 3.** Mask process of Bidirectional Encoder Representation from Transformers (BERT) and Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking (RoBERTa-wwm).

**Figure 4.** Coding representation of Transformer with 12 layers of Bidirectional Encoder Representation from Transformers model.

Assume that the text input sequence \(seq = (x_1, x_2, x_3, \ldots, x_n)\), where \(n\) is the total length of the character contained in the sequence; \(x_i\) is the \(i^{th}\) character of the input sequence; and the fusion formula is as follows:

\[
\hat{v}_{RoBERTa-wwm}^{(i)} = \text{Dense}_{unit=512} (x_i) \left( \sum_{c=1}^{12} \alpha_c \times h_c \right), (c \in \{1, 12\}) \tag{4}
\]

\(v_{RoBERTa-wwm}^{(i)}\) is the output representation by the RoBERTa-wwm model with dynamic fusion for the current character \(x_i\); \(h_c\) is the output representation by each transformer layer of the RoBERTa-wwm model, and \(\alpha_c\) is the output weight value assigned to each layer by RoBERTa-wwm.
Fine-Grained Semantic Feature

5-Stroke Code

The 5-stroke code is a typical semantic code, which encodes Chinese characters according to strokes and structures. Currently, it is widely used to code Chinese characters. The expression of the 5-stroke code may inevitably repeat with the phonological code, for example, the 5-stroke code for “亦” (also) is “you,” while the phonological code for “亦” (also) is also “you” [62]. Hence, we combined the 5-stroke code and Zheng code to compensate for the encoding deficiency. We used the 5-stroke code in Zdic.net to vectorize the Chinese characters using the following formulas:

\[ p = f_s (seq) \]
\[ v_{pc} = e_{pc} (pi), (i \in Z \cap i \in [1,n]) \]

where \( f_s \) represents the function that maps the input character sequence into the 5-stroke code and \( v_{pc}^{i} \) represents the 5-stroke code vector corresponding to \( x_i \).

Zheng Code

The Zheng code was created by famous Chinese literature professors as per the strokes and roots of Chinese characters through in-depth research on the patterns and structures of Chinese characters. The early Microsoft operating system in Chinese adopted the Zheng code as the built-in code. This indicates that Zheng code is a scientific coding of Chinese characters. Chinese characters with similar codes may contain related semantic information. Hence, the potential semantic relationship of text may be found by mining the structural information of Chinese characters using Zheng code. The Zheng code was vectorized as the 5-stroke code and has the following formulas:

\[ p = f_z (seq) \]
\[ v_{zc} = e_{zc} (pi), (i \in Z \cap i \in [1,n]) \]

where \( f_z \) represents the function that maps the input character sequence into the Zheng code and \( v_{zc}^{i} \) represents the Zheng code vector corresponding to \( x_i \).

Phonological Code

Over 90% of Chinese characters are picto-phonetic characters [63]. Hence, pronunciation plays an important role in the semantic expressions of Chinese characters. We used the Pinyin toolkit to vectorize the phonological code of Chinese characters using the following formulas:

\[ p = f_p (seq) \]
\[ v_{pc} = e_{pc} (pi), (i \in Z \cap i \in [1,n]) \]

where \( f_p \) represents the function that maps the input character sequence into the phonological code and \( v_{pc}^{i} \) represents the phonological vector corresponding to \( x_i \).

Stroke Code

Chinese characters with similar strokes may have similar meanings. The strokes of each Chinese character were encoded in ZDIC [59], where 1, 2, 3, 4, and 5 represent the horizontal stroke, vertical stroke, left-falling stroke, right-falling stroke, and turning stroke, respectively. We transformed the stroke code into a 5-dimension vector, where each dimension was the corresponding number of strokes. The stroke code was vectorized in the same manner as the 5-stroke code and has the following formulas:

\[ p = f_s (seq) \]
\[ v_{pc} = e_{pc} (pi), (i \in Z \cap i \in [1,n]) \]

where \( f_s \) represents the function that maps the input character sequence into the stroke code and \( v_{pc}^{i} \) represents the stroke code vector corresponding to \( x_i \).

To extract the fine-grained semantic features of Chinese characters deeply, we trained the character features using CNNs. The character features were trained by 2 convolutions with a kernel of 3 and ReLU function as well as max pooling of 2×2, where the number of output channels was the dimension of each feature vector. Finally, the 32-dimension Chinese character vector was obtained through a full connection in the dense layer, as shown in Figure 2.

Image Feature

Chinese characters have been derived from pictographic symbols since ancient times, and characters with similar symbolic appearances have similar image features. However, the fonts of Chinese characters have changed a lot over time. Simplified characters have lost much pictographic information compared with complex characters. Therefore, Cui et al [64] used Chinese character images to extract Chinese character features and achieved better performance. Wu et al [65] tried different character fonts and found that the best result was obtained by using the NotoSansCJKsc-Regular font. On the basis of these findings, we used the Python Imaging Library to convert NotoSansCJKsc-Regular Chinese characters into black-and-white images and extracted image features by 2D CNN in depth as per the following formulas:

\[ e_{i,1} = \text{Max pooling 1} \]
\[ e_{i,2} = \text{Max pooling 2} \]
\[ v_{if} = \text{Dense} (e_{i,2}) \]

where \( K \) is a kernel; \( H \) is the original embedded image matrix; \( Conv \), \( Max pooling \), and \( Max pooling \) are the first convolutions with a kernel of 3 and channel of 8, the first max pooling with the kernel of 2×2, the second convolution with the kernel of 3 and channel of 32, and the second max pooling with the kernel of 2×2, respectively; \( e_{i,1} \) is the result after the first convolution; \( e_{i,2} \) is the result after the second convolution; Dense is the process of realizing the full connection; and \( v_{if} \) is the final 128-dimension Chinese character image vector trained by convolution, as shown in Figure 2.

Finally, the multisemantic features \( v_{i}^{\text{RoBERTa-wmm}}, v_{i}^{f}, v_{i}^{nc}, v_{i}^{pc}, v_{i}^{sc}, \) and \( v_{if} \) were embedded by the array Concat function. The formula used is as follows:

\[ v_{i}^{\text{input}} = \text{Concat} (v_{i}^{\text{RoBERTa-wmm}}, v_{i}^{f}, v_{i}^{nc}, v_{i}^{pc}, v_{i}^{sc}, v_{if}) \]
The role of BiLSTM [66] is essential in NER. As shown in Figure 1, the forward LSTM and backward LSTM are responsible for memorizing the previous and subsequent text information, respectively. By combining the 2, contextual information can be obtained simultaneously, which helps to capture the bidirectional semantic dependency information in the text. The formulas used are as follows:

$$h_{i}^{\text{forward}} = \text{LSTM}^{\text{forward}}(\alpha^{i-1}, x_{i}) \quad (17)$$

$$h_{i}^{\text{backward}} = \text{LSTM}^{\text{backward}}(\alpha^{i+1}, x_{i}) \quad (18)$$

$$h_{i} = [h_{i}^{\text{forward}}, h_{i}^{\text{backward}}] \quad (19)$$

where $\alpha^{i}$ represents the hidden layer state of the current memory cell; $\text{LSTM}^{\text{forward}}$ is the feature representation from front to back; $h_{i}^{\text{forward}}$ is the forward semantic information obtained through the forward LSTM at the $i$-th character position; $h_{i}^{\text{backward}}$ is the backward semantic information obtained through the backward LSTM at the $i$-th character position; and $h_{i}$ represents a combination of hidden states in both.

**CRF Layer**

The BiLSTM can be used to handle contextual relationships. However, it cannot consider the dependencies between tags. Therefore, it is necessary to add a constraint relation for the final predicted label by using the CRF [67] layer to ensure the predicted label rationality. Given an input sequence where $X = \{x_{1}, x_{2}, ..., x_{n}\}$, we assume that the training output label sequence is $Y = \{y_{1}, y_{2}, ..., y_{n}\}$, where $n$ is the number of model labels. The sequence score of the label and the probability of the label sequence $y$ are calculated as follows:

$$P(y|X) = \frac{\sum_{i=1}^{n} (Z_{y_{i}, y_{i+1}} + P_{y_{i+1}, y_{i+1}})}{(Z_{y_{i}, y_{i+1}} + P_{y_{i+1}, y_{i+1}})} \quad (20)$$

where $Z$ is the transfer matrix; $Z_{y_{i}, y_{i+1}}$ is the score of the label transfer from $y_{i}$ to $y_{i+1}$; $P_{y_{i+1}, y_{i+1}}$ is the score of label $y_{i+1}$ corresponding to the $i+1$th character of the input sequence; $Y$ is the set of all possible label sequences. The final label of the output sequence is the set of labels with the highest probability.

Finally, we predicted the best label sequences by using the Viterbi algorithm [68] with the following formula:

$$y^* = \text{argmax}(s(X, y)) \quad (21)$$

**Results**

To get convincing experimental results, we ran each model 5 times and calculated the average precision, average recall, and average $F_1$-score.

**Performance Comparison With Ensemble Models**

To verify the validity of the model, we compared our model with the existing ensemble models BiLSTM-CRF, ELMo-Lattice-LSTM-CRF, ELMo-BiLSTM-CRF, all CNNs, ELMo-encoder from transformer-CRF, and multigranularity semantic dictionary and multimodal tree-NER on Yidu-S4K and self-annotated data sets, and the results are shown in Table 4. The $F_1$-scores of the experimental model on the Yidu-S4K data set were 18.31%, 4.15%, 4.26%, 4.12%, 3.69%, and 2.59% higher than those of the BiLSTM-CRF, all CNNs, ELMo-Lattice-LSTM-CRF, ELMo-BiLSTM-CRF, ELMo-encoder from transformer-CRF, and multigranularity semantic dictionary and multimodal tree-NER models, respectively. On the self-annotated data set, it was 5.14% higher than that of the BiLSTM-CRF. The results showed that the performance of the experimental model is superior to that of the existing model.
Performance Comparison With PLMs Related to BERT

The performance of the PLM, BERT, is a milestone in natural language processing. To verify the BERT robust version’s validity of the RoBERTa-wwm model, we compared our model with the existing ensemble models with the BiLSTM-CRF, BERT-BiLSTM-CRF, and RoBERTa-wwm-BiLSTM-CRF on Yidu-S4K and self-annotated data sets, and the results are shown in Table 5. The $F_1$-scores of the experimental model on the Yidu-S4K data set were 18.31%, 2.99%, and 0.82% higher than those of the BiLSTM-CRF, BERT-BiLSTM-CRF, and RoBERTa-wwm-BiLSTM-CRF models, respectively, and 5.14%, 2.95%, and 1.07% higher on the self-annotated data set, respectively.

Table 5. Performance comparison of PLMs$^a$ on the Yidu-S4K and self-annotated data sets.

<table>
<thead>
<tr>
<th>Data set and model</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>$F_1$-score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Yidu-S4K</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTM-CRF$^b$ [64]</td>
<td>69.43</td>
<td>72.58</td>
<td>70.97</td>
</tr>
<tr>
<td>BERT$^d$-BiLSTM-CRF</td>
<td>89.08</td>
<td>86.90</td>
<td>88.46</td>
</tr>
<tr>
<td>RoBERTa-wwm$^e$-BiLSTM-CRF</td>
<td>90.08</td>
<td>88.22</td>
<td>89.28</td>
</tr>
<tr>
<td>Our model</td>
<td>90.37</td>
<td>88.22</td>
<td>89.28</td>
</tr>
<tr>
<td><strong>Self-annotated</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTM-CRF</td>
<td>81.98</td>
<td>77.10</td>
<td>79.47</td>
</tr>
<tr>
<td>BERT-BiLSTM-CRF</td>
<td>82.48</td>
<td>80.86</td>
<td>81.66</td>
</tr>
<tr>
<td>RoBERTa-wwm-BiLSTM-CRF</td>
<td>84.23</td>
<td>82.86</td>
<td>83.54</td>
</tr>
<tr>
<td>Our model</td>
<td>84.24</td>
<td>84.99</td>
<td>84.61</td>
</tr>
</tbody>
</table>

$^a$PLM: pretrained language model.
$^b$BiLSTM: Bidirectional Long Short-Term Memory.
$^c$CRF: conditional random fields.
$^d$BERT: Bidirectional Encoder Representation from Transformers.
$^e$RoBERTa-wwm: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking.
Performance Comparison of Each Entity

To comprehensively evaluate our model, we calculated the $F_1$-score for each entity type on the Yidu-S4K and self-annotated data sets, as shown in Tables 6 and 7. The $F_1$-score of our model on the Yidu-S4K data set for each of the 6 entity categories, except for the Image entity, increased by 0.2% to 7.6% compared with the data listed in the tables. The $F_1$-score for the Image entity was 0.35% lower than that of the ELMo-BiLSTM-CRF model. However, the $F_1$-scores for the Laboratory entity and Operation entity were 7.6% and 7.54% higher than those of the ELMo-BiLSTM-CRF model, respectively. The overall $F_1$-score was 4.12% higher than that of the ELMo-BiLSTM-CRF model. For the self-annotated data set, our model improved each entity in 7 categories ranging from 0.09% to 14.49% over the listed data, with a greater improvement for Instrument entities.

Table 6. Performance comparison of each entity category on the Yidu-S4K data set.

<table>
<thead>
<tr>
<th>Model</th>
<th>All</th>
<th>Disease</th>
<th>Anatomy</th>
<th>Image</th>
<th>Laboratory</th>
<th>Medicine</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELMo\textsuperscript{a}-BiLSTM\textsuperscript{b}-CRF\textsuperscript{c} [41]</td>
<td>85.16</td>
<td>82.81</td>
<td>85.99</td>
<td>88.01</td>
<td>75.65</td>
<td>94.49</td>
<td>86.79</td>
</tr>
<tr>
<td>BERT\textsuperscript{d}-BiLSTM-CRF</td>
<td>86.29</td>
<td>87.14</td>
<td>86.36</td>
<td>83.43</td>
<td>77.98</td>
<td>89.46</td>
<td>93.11</td>
</tr>
<tr>
<td>BERT-wwm\textsuperscript{e}-BiLSTM-CRF</td>
<td>87.12</td>
<td>86.18</td>
<td>85.47</td>
<td>81.52</td>
<td>79.69</td>
<td>90.14</td>
<td>92.49</td>
</tr>
<tr>
<td>RoBERTa\textsuperscript{f}-wwm-BiLSTM-CRF</td>
<td>88.46</td>
<td>87.71</td>
<td>87.01</td>
<td>86.69</td>
<td>82.36</td>
<td>93.22</td>
<td>92.87</td>
</tr>
<tr>
<td>Our model</td>
<td>89.28</td>
<td>87.91</td>
<td>87.47</td>
<td>87.66</td>
<td>83.25</td>
<td>94.98</td>
<td>94.33</td>
</tr>
</tbody>
</table>

\textsuperscript{a}ELMo: Embeddings from Language Models.  
\textsuperscript{b}BiLSTM: Bidirectional Long Short-Term Memory.  
\textsuperscript{c}CRF: conditional random fields.  
\textsuperscript{d}BERT: Bidirectional Encoder Representation from Transformers.  
\textsuperscript{e}wwm: Whole Word Masking.  
\textsuperscript{f}RoBERTa: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach.

Table 7. Performance comparison of each entity category on the self-annotated data set.

<table>
<thead>
<tr>
<th>Model</th>
<th>All</th>
<th>Disease</th>
<th>Symptoms</th>
<th>Anatomy</th>
<th>Examination</th>
<th>Instrument</th>
<th>Medicine</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT\textsuperscript{a}-BiLSTM\textsuperscript{b}-CRF\textsuperscript{c}</td>
<td>81.66</td>
<td>81.33</td>
<td>85.87</td>
<td>83.86</td>
<td>90.36</td>
<td>60.38</td>
<td>89.72</td>
<td>79.75</td>
</tr>
<tr>
<td>BERT-wwm\textsuperscript{d}-BiLSTM-CRF</td>
<td>81.58</td>
<td>74.91</td>
<td>83.89</td>
<td>81.23</td>
<td>88.84</td>
<td>54.76</td>
<td>85.63</td>
<td>68.49</td>
</tr>
<tr>
<td>RoBERTa\textsuperscript{e}-wwm-BiLSTM-CRF</td>
<td>83.54</td>
<td>81.99</td>
<td>86.69</td>
<td>84.68</td>
<td>91.21</td>
<td>66.01</td>
<td>91.04</td>
<td>81.17</td>
</tr>
<tr>
<td>Our model</td>
<td>84.61</td>
<td>82.34</td>
<td>86.93</td>
<td>85.62</td>
<td>91.30</td>
<td>69.25</td>
<td>91.28</td>
<td>82.49</td>
</tr>
</tbody>
</table>

\textsuperscript{a}BERT: Bidirectional Encoder Representation from Transformers.  
\textsuperscript{b}BiLSTM: Bidirectional Long Short-Term Memory.  
\textsuperscript{c}CRF: conditional random fields.  
\textsuperscript{d}wwm: Whole Word Masking.  
\textsuperscript{e}RoBERTa: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach.

Ablation Analysis

Ablation Experiments for Multisemantic Features

To verify the fine-grained semantic features and image features of Chinese characters, dynamic fusion was effective. We used the RoBERTa-wwm-BiLSTM-CRF model as the baseline to perform ablation experiments for the above contents on 2 EMR data sets, and the results are shown in Figure 5.

The performance of the model was significantly improved with the dynamic fusion of RoBERTa-wwm. After incorporating the semantic features of Chinese characters into the model alone, the overall performance of the model was not as high as that after dynamic fusion. However, the performance on both data sets was superior to that of the baseline. The performance of the model was unstable when image features of Chinese characters were added to the model alone. On the Yidu-S4K data set, the model’s performance was inferior to that of the baseline, whereas on the self-annotated data set, the model’s performance only improved slightly. After adding the semantic and image features of Chinese characters to the model, the performance of the model on the Yidu-S4K data set was superior to that of the baseline. Furthermore, it was better than that of the model with semantic or image features of Chinese characters alone. The performance of the model on the self-annotated data set was superior to that of the baseline and better than that of the model with the image features of Chinese characters alone. When the model combined dynamic fusion with the semantic features of Chinese characters, the performance was greatly improved.
features and image features of Chinese characters, it was found that the performance of the model was significantly improved on the 2 data sets. Dynamic fusion with image features of Chinese characters showed the best comprehensive performance on the Yidu-S4K data set, whereas dynamic fusion with semantic features of Chinese characters achieved the best comprehensive performance on the self-annotated data set. After combining the semantic and image features of the Chinese characters and dynamic fusion, it was noted that the performance of the model was superior to that of the baseline. Because the quality of the self-annotated EMRs is inferior to that of the public Chinese EMRs corpus and the self-annotated data set contains a wider coverage of departments, the comprehensive effect of the self-annotated data set is lower than that of the YiduS4K data set in Figure 5.

**Figure 5.** The results of ablation experiments for mutisemantic features on the Yidu-S4K and self-annotated data sets. BiLSTM: Bidirectional Long Short-Term Memory; CRF: Conditional Random Fields; RoBERTa-wwm: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking.

**Ablation Experiments for Fine-Grained Semantic Features**

The fine-grained semantic features of Chinese characters used in this study included the 5-stroke code, Zheng code, phonological code, and stroke code. To verify the effectiveness of these features, we used the RoBERTa-wwm-BiLSTM-CRF model as the baseline to perform ablation experiments for the 4 features on the 2 EMR data sets, and the results are shown in Figure 6 and Figure 7.
Figure 6. The results of ablation experiments for fine-grained semantic features on the Yidu-S4K data set. BiLSTM: Bidirectional Long Short-Term Memory; CRF: conditional random fields; RoBERTa-wwm: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking.

Figure 7. The results of ablation experiments for fine-grained semantic features on the self-annotated data set. BiLSTM: Bidirectional Long Short-Term Memory; CRF: conditional random fields; RoBERTa-wwm: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking.

The $F_1$-score of the model on the Yidu-S4K data set ranked in the top 2 for the 5-stroke code and Zheng code, whereas the $F_1$-score on the self-annotated data set ranked in the top 2 for the phonological code or Zheng code. The performance of the
model combining 2 features (the combination of 5-stroke code and Zheng code or the combination of phonological code and stroke code) was better than that of the model with only 1 feature, regardless of the data set. On the Yidu-S4K data set, the model combining 5-stroke code+phonological code+stroke code showed the best comprehensive performance, followed by the combinations of 5-stroke code+Zheng code+phonological code, 5-stroke code+Zheng code+stroke code, and Zheng code+phonological code+stroke code. On the self-annotated data set, the model combining the 5-stroke code+Zheng code+phonological code showed the best comprehensive performance, followed by Zheng code+phonological code+stroke code, 5-stroke code+phonological code+stroke code, and 5-stroke code+Zheng code+stroke code. On the Yidu-S4K data set, only the model combining the 5-stroke code+phonological code+stroke code showed a comprehensive performance superior to that of the baseline. However, on the self-annotated data set, the comprehensive performance of all combinations was superior to that of the baseline. The performance of the model combining 3 features was less stable. The model combined 4 features on the Yidu-S4K and self-annotated data sets and achieved the best comprehensive performance among all the combinations.

Error Analysis

From Tables 6 to Table 7, our model improved the entity recognition performance of each entity category to different degrees. However, the entity recognition effect differs for each category. The $F_1$-scores of Disease, Anatomy, Image, Laboratory, Medicine, and Operation entity recognition on the Yidu-S4K data set were 87.91%, 87.47%, 87.66%, 83.25%, 94.98%, and 94.33%, respectively. The $F_1$-scores of Disease, Symptoms, Anatomy, Examination, Instrument, Medicine, and Operation entity recognition on the self-annotated data set were 82.34%, 86.93%, 85.62%, 91.31%, 69.25%, 91.28%, and 82.49%, respectively. On the Yidu-S4K data set, the precision of Laboratory entity recognition was the lowest, followed by the Anatomy entity, Image entity, and Disease entity. On the self-annotated data set, the precision of Instrument entity recognition was the lowest, followed by the Disease entity, Anatomy entity, and Operation entity. We concluded the following 7 main causes of the errors that occurred based on a review of the data set and model prediction results, as shown in Table 8.

We strictly controlled the annotation quality of both data sets. Hence, the probability of causes (1-3) was relatively low. Causes (4-6) were more likely to occur, and cause (7) mainly occurred on some entities that were less common or had fewer training samples.
Table 8. Different types of errors on 2 data sets.

<table>
<thead>
<tr>
<th>Types of errors</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Annotation error</td>
<td>For instance, some Laboratory entities, like “PLT,” “NEUT,” and “CAE,” on the Yidu-S4K data set contained commas, which were correctly recognized as “PLT,” “NEUT,” and “CAE” in our model.</td>
</tr>
<tr>
<td>1. Some manually annotated entities contained punctuation marks unrelated to the entities.</td>
<td>For example, “PET-CT” was manually annotated as a Laboratory entity on the Yidu-S4K data set, but our model correctly predicted as an Image entity.</td>
</tr>
<tr>
<td>2. A few entity categories were confused.</td>
<td>On the Yidu-S4K data set, the character “下 (below)” expressing orientation of “剑突下 (below xiphoid)” was not annotated, and the character “部 (part)” expressing the part of “咽喉部 (hypopharynx)” was also not annotated. Most of the characters expressing specific locations were annotated.</td>
</tr>
<tr>
<td>(2) Inconsistent annotation</td>
<td>The Disease entity “窦性心律 (sinus rhythm)” was missed annotated on the Yidu-S4K data set, and the Medicine entity “生理盐水 (normal saline)” was missed annotated on the self-annotated data set.</td>
</tr>
<tr>
<td>The inconsistent annotation will affect the accuracy of machine learning.</td>
<td>The failure to recognize the non-Chinese character entities, like the Laboratory entity “AFP” on the Yidu-S4K data set and the Examination entity “HCG” on the self-annotated data set, so did such situations as the Medicine entity “VP-16” was recognized as “VP-,” and “50%葡萄糖 (50% glucose)” as “葡萄糖 (glucose)” on the Yidu-S4K data set.</td>
</tr>
<tr>
<td>(3) Missing annotation</td>
<td>The Disease entity “窦性心律 (sinus rhythm)” was missed annotated on the Yidu-S4K data set, and the Medicine entity “生理盐水 (normal saline)” was missed annotated on the self-annotated data set.</td>
</tr>
<tr>
<td>The missing annotated entity will also affect the overall effect of the model.</td>
<td>The failure to recognize the non-Chinese character entities, like the Laboratory entity “AFP” on the Yidu-S4K data set and the Examination entity “HCG” on the self-annotated data set, so did such situations as the Medicine entity “VP-16” was recognized as “VP-,” and “50%葡萄糖 (50% glucose)” as “葡萄糖 (glucose)” on the Yidu-S4K data set.</td>
</tr>
<tr>
<td>(4) Entity with a non-Chinese character symbol</td>
<td>The Disease entity “窦性心律 (sinus rhythm)” was missed annotated on the Yidu-S4K data set, and the Medicine entity “生理盐水 (normal saline)” was missed annotated on the self-annotated data set.</td>
</tr>
<tr>
<td>Figures, letters, and other symbols cannot be extracted with more semantic features than Chinese characters. Hence, it may be difficult to recognize entities with symbols other than Chinese characters in the Chinese corpus.</td>
<td>The failure to recognize the non-Chinese character entities, like the Laboratory entity “AFP” on the Yidu-S4K data set and the Examination entity “HCG” on the self-annotated data set, so did such situations as the Medicine entity “VP-16” was recognized as “VP-,” and “50%葡萄糖 (50% glucose)” as “葡萄糖 (glucose)” on the Yidu-S4K data set.</td>
</tr>
<tr>
<td>(5) Presence of nested entities</td>
<td>The Disease entity “窦性心律 (sinus rhythm)” was missed annotated on the Yidu-S4K data set, and the Medicine entity “生理盐水 (normal saline)” was missed annotated on the self-annotated data set.</td>
</tr>
<tr>
<td>On the Yidu-S4K data set, the Disease entity and Image entity might contain the Anatomy entity.</td>
<td>The failure to recognize the non-Chinese character entities, like the Laboratory entity “AFP” on the Yidu-S4K data set and the Examination entity “HCG” on the self-annotated data set, so did such situations as the Medicine entity “VP-16” was recognized as “VP-,” and “50%葡萄糖 (50% glucose)” as “葡萄糖 (glucose)” on the Yidu-S4K data set.</td>
</tr>
<tr>
<td>On the self-annotated data set, entity nesting is more severe, the Disease entity, Examination entity, and Instrument entity might contain the Anatomy entity, and the Instrument entity might contain the Operation entity.</td>
<td>The failure to recognize the non-Chinese character entities, like the Laboratory entity “AFP” on the Yidu-S4K data set and the Examination entity “HCG” on the self-annotated data set, so did such situations as the Medicine entity “VP-16” was recognized as “VP-,” and “50%葡萄糖 (50% glucose)” as “葡萄糖 (glucose)” on the Yidu-S4K data set.</td>
</tr>
<tr>
<td>(6) More entities with mixed representation</td>
<td>The Disease entity “窦性心律 (sinus rhythm)” was missed annotated on the Yidu-S4K data set, and the Medicine entity “生理盐水 (normal saline)” was missed annotated on the self-annotated data set.</td>
</tr>
<tr>
<td>Entity composition is more complex, mixed representations occur more often.</td>
<td>The failure to recognize the non-Chinese character entities, like the Laboratory entity “AFP” on the Yidu-S4K data set and the Examination entity “HCG” on the self-annotated data set, so did such situations as the Medicine entity “VP-16” was recognized as “VP-,” and “50%葡萄糖 (50% glucose)” as “葡萄糖 (glucose)” on the Yidu-S4K data set.</td>
</tr>
<tr>
<td>(7) Insufficient entity training data</td>
<td>The Disease entity “窦性心律 (sinus rhythm)” was missed annotated on the Yidu-S4K data set, and the Medicine entity “生理盐水 (normal saline)” was missed annotated on the self-annotated data set.</td>
</tr>
<tr>
<td>In the case of insufficient training samples, the machine may provide inadequate training for entities, so that the machine cannot fully learn the features of such entities, failing to recognize many entities.</td>
<td>The failure to recognize the non-Chinese character entities, like the Laboratory entity “AFP” on the Yidu-S4K data set and the Examination entity “HCG” on the self-annotated data set, so did such situations as the Medicine entity “VP-16” was recognized as “VP-,” and “50%葡萄糖 (50% glucose)” as “葡萄糖 (glucose)” on the Yidu-S4K data set.</td>
</tr>
</tbody>
</table>

\(^{(a)}\) PLT: platelet count. 
\(^{(b)}\) NEUT: neutrophil count. 
\(^{(c)}\) CAE: carcinoembryonic antigen. 
\(^{(d)}\) PET-CT: positron emission tomography-computed tomography. 
\(^{(e)}\) AFP: alpha fetoprotein. 
\(^{(f)}\) HCG: human chorionic gonadotropin. 
\(^{(g)}\) VP: etoposide. 
\(^{(h)}\) CD5: a differentiation antigen, cluster of differentiation 5.
Discussion

Principal Findings

In this study, we developed a Chinese CNER method based on multisemantic features. The method extracted the semantic features of text using the RoBERTa-wwm model after dynamic fusion, extracted the fine-grained semantic features of Chinese characters by 1D CNN, and converted Chinese characters into square images to extract the image features of the simplified Chinese characters from another modality by 2D CNN. We conducted a series of experiments to evaluate the model’s performance on the Yidu-S4K data set and self-annotated data set; the results showed that the F1-scores of the proposed model in this study were 89.28% and 84.61% on the 2 data sets, respectively. The model showed a higher and more stable performance in all experiments and could help recognize entities in most categories. Furthermore, its migratory property and adaptability to different data were acceptable. We also demonstrated that multisemantic features were effective through 2 ablation experiences and analyzed the error cases of NER, which might provide a basis for subsequent studies and standardization of the corpus.

Compared with ensemble models, for the BiLSTM-CRF model, the representation information of characters was obtained with the help of a vector look-up table. However, the information obtained by this method was too simple to excavate the text’s semantic meaning or solve problems such as the polysemy of words. Hence, the model did not perform well. Kong et al [69] constructed a multilayer CNN to obtain short-term and long-term contextual information, and the attention mechanism was used to calculate the weight distribution in each hidden layer so that the features of each coding layer could be fully extracted and used to improve the entity recognition performance. However, this model required numerous radical and dictionary features to complete the semantic supplement of the context. Li et al [70] proposed an ELMo-Lattice-LSTM-CRF model. The ELMo word dynamic representation model could learn complicated word features and the context-based changes of these features, while the lattice structure provided extra entity boundaries and other semantic information for CNER of EMRs through the Word2Vec model and dictionaries. Li et al [41] proposed an ELMo-BiLSTM-CRF model that improved the semantic recognition ability of the machine for text. It reduced problems, such as word polysemy, when compared with the BiLSTM-CRF model and reduced the computational complexity of the lattice structure compared with the ELMo-Lattice-LSTM-CRF model. Moreover, this model could fully use contextual information by replacing LSTM with BiLSTM. Wan et al [71] fine-tuned the ELMo model based on EMRs to achieve embedding for domain-specific text and then used a transformer as an encoder to alleviate the long context–dependent problems and finally achieved CNER through CRF decoding. Wang et al [72] proposed a model for NER based on the LSTM-CRF model by storing and merging characters, words, and other features. However, as the text embedding process of this method is more complicated, it is necessary to create dictionaries of characters and words to obtain multi granularity text features at first and then store and merge the obtained features using a tree structure to achieve text embedding. These methods have achieved a few good results, but our proposed method is still competitive and has the best performance among all the models, as shown in Table 4.

Compared with PLMs related to BERT, both the BERT-BiLSTM-CRF and BiLSTM-CRF models could obtain word-level vector representations. However, the word-level vector obtained by BERT contained rich contextual characteristics, including morphology, syntax, semantics, location, and other important semantic information, which can directly improve the task performance by replacing the lattice structure and complicated text representation methods in Table 4, such as dictionaries of characters and words. Compared with BERT, RoBERTa-wwm used more data for pretraining, and the dynamic wwm allows itself to flexibly learn word-level representation information, which compensates for the shortcomings that BERT can only obtain character-level representation. Thus, richer word-based text representation information could be obtained. Combined with the experimental results in Table 4, the RoBERTa-wwm-BiLSTM-CRF model, without introducing features, outperformed the other ensemble models. Therefore, using the PLM RoBERTa-wwm with a whole word mask can effectively improve the Chinese CNER performance, thus avoiding the use of complex text embedding and feature embedding methods.

In addition, 2 ablation experiments showed that different features and means lead to different degrees of improvement in the semantic comprehension ability of the model. Multisemantic features could help the machine to obtain richer semantic information, whereas dynamic fusion could fully recognize and use the representation information so that the model performance could be comprehensively improved. Considering the heterogeneity among data, using 1 method alone or both methods may affect the generalization ability of the model. In this study, the model combining the fine-grained semantic features and image features of Chinese characters and dynamic fusion might not show the best performance. However, it was more universal and could maintain the performance at a relatively high level compared with other experimental models. Furthermore, introducing more feature engineering was conducive to fully mining the semantic information of text annotation with the help of fine-grained semantic information contained in Chinese characters and improving the performance of the model on different data sets through the cross-complementarity of different features in a relatively stable manner.

To reduce the error rate of entity recognition, specifically for human-caused errors, we could try to avoid these problems by further improving the annotation quality. For the data special characteristics or data defects, the errors might be reduced by medical knowledge, medical dictionaries, and some rules, regardless of the lack of training data.

Limitations and Future Work

The limitation of this study was that the ratio of the 6 entity types on the Yidu-S4K data set did not exactly follow 7:3, such that the ratio of the training set to test set for disease entities is approximately 0.7610:0.2390; the ratio of the training set to test...
set for medicine entities is approximately 0.7898:0.2102; and the ratio of the training set to test set for all entities is approximately 0.7463:0.2537. The unbalanced data of different entity types in the training and test sets caused a performance bias. Although the ratio of the training set to the test set of the EMRs was 7:3, we could not ensure that the number of entities of each type in each EMR in the training set and test set remained at a similar ratio.

In the future, we may focus on the recognition of a specific entity type in EMRs to improve the CNER performance. In addition, we will incorporate other prior medical knowledge or assign different weights to the Chinese character semantic features and image features, such as using the attention mechanism to capture important features, to improve the performance of the model.

**Conclusions**

This study proposes a Chinese CNER method to learn a semantics-enriched representation of Chinese character features in EMRs to enhance the specificity and diversity of feature representations. The results showed that the model had state-of-the-art performance on 2 Chinese CNER data sets compared with existing models. We demonstrated that multisemantic features could provide richer and more fine-grained semantic information for Chinese CNER through the cross-complementarity of different semantic features. This enabled the model to learn a better feature representation and improve its generalization ability.

**Acknowledgments**

The authors would like to thank the YiduCloud for providing the Yidu-S4K corpora. This work was supported by the Science and Technology Innovation 2030 “New Generation Artificial Intelligence” major project “Research on the construction and application of knowledge system for Medical Artificial intelligence Services,” China (grant 2020AAA0104901).

**Conflicts of Interest**

None declared.

**References**


Abbreviations

BERT: Bidirectional Encoder Representation from Transformers
BiLSTM: Bidirectional Long Short-Term Memory
CNER: clinical named entity recognition
CNN: convolutional neural network
CRF: conditional random fields
ELMo: Embeddings from Language Models
EMR: electronic medical record
FN: false negative
FP: false positive
LSTM: long short-term memory
NER: named entity recognition
PLM: pretrained language model
RoBERTa: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach
TP: true positive
Word2Vec: Word to Vector
wwm: Whole Word Masking

Edited by G Eysenbach; submitted 01.12.22; peer-reviewed by C Yu, L Heryawan; comments to author 31.01.23; revised version received 18.02.23; accepted 31.03.23; published 10.05.23.

Please cite as:
Wang W, Li X, Ren H, Gao D, Fang A
Chinese Clinical Named Entity Recognition From Electronic Medical Records Based on Multisemantic Features by Using Robustly Optimized Bidirectional Encoder Representation From Transformers Pretraining Approach Whole Word Masking and Convolutional Neural Networks: Model Development and Validation
JMIR Med Inform 2023;11:e44597
URL: https://medinform.jmir.org/2023/1/e44597
doi:10.2196/44597
PMID:37163343

©Weijie Wang, Xiaoying Li, Huiling Ren, Dongping Gao, An Fang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 10.05.2023. 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.
Improving an Electronic Health Record–Based Clinical Prediction Model Under Label Deficiency: Network-Based Generative Adversarial Semisupervised Approach

Runze Li1*, BE; Yu Tian1*, PhD; Zhuyi Shen1, BE; Jin Li2, PhD; Jun Li3, MD; Kefeng Ding3, MD; Jingsong Li1, PhD

1College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China
2Institute for Artificial Intelligence in Medicine, School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, China
3Department of Surgical Oncology, Zhejiang University School of Medicine Second Affiliated Hospital, Hangzhou, China
*these authors contributed equally

Corresponding Author:
Jingsong Li, PhD
College of Biomedical Engineering and Instrument Science
Zhejiang University
Zhou Yiqing Science and Technology Building, 2nd floor
38 Zheda Road
Hangzhou, 310027
China
Phone: 86 571 87951564
Fax: 86 571 87951564
Email: ljs@zju.edu.cn

Abstract

Background: Observational biomedical studies facilitate a new strategy for large-scale electronic health record (EHR) utilization to support precision medicine. However, data label inaccessibility is an increasingly important issue in clinical prediction, despite the use of synthetic and semisupervised learning from data. Little research has aimed to uncover the underlying graphical structure of EHRs.

Objective: A network-based generative adversarial semisupervised method is proposed. The objective is to train clinical prediction models on label-deficient EHRs to achieve comparable learning performance to supervised methods.

Methods: Three public data sets and one colorectal cancer data set gathered from the Second Affiliated Hospital of Zhejiang University were selected as benchmarks. The proposed models were trained on 5% to 25% labeled data and evaluated on classification metrics against conventional semisupervised and supervised methods. The data quality, model security, and memory scalability were also evaluated.

Results: The proposed method for semisupervised classification outperforms related semisupervised methods under the same setup, with the average area under the receiver operating characteristics curve (AUC) reaching 0.945, 0.673, 0.611, and 0.588 for the four data sets, respectively, followed by graph-based semisupervised learning (0.450, 0.454, 0.425, and 0.5676, respectively) and label propagation (0.475, 0.344, 0.440, and 0.477, respectively). The average classification AUCs with 10% labeled data were 0.929, 0.719, 0.652, and 0.650, respectively, comparable to that of the supervised learning methods logistic regression (0.601, 0.670, 0.731, and 0.710, respectively), support vector machines (0.733, 0.720, 0.720, and 0.721, respectively), and random forests (0.982, 0.750, 0.758, and 0.740, respectively). The concerns regarding the secondary use of data and data security are alleviated by realistic data synthesis and robust privacy preservation.

Conclusions: Training clinical prediction models on label-deficient EHRs is indispensable in data-driven research. The proposed method has great potential to exploit the intrinsic structure of EHRs and achieve comparable learning performance to supervised methods.

(JMIR Med Inform 2023;11:e47862) doi:10.2196/47862
Introduction

The recent rise of observational biomedical research, driven by greatly expanding electronic health records (EHRs) and the prevalence of machine learning methods, has drawn great attention [1-4]. Conventional strategies tend to screen out subgroups of interest based on expert supervision or established risk factors. An alternative data-driven paradigm extracts underlying subtypes by comprehensively profiling the longitudinal irregularity, interdimensional heterogeneity, and intrinsic homogeneity of the database, thus progressively facilitating the practice of precision medicine. For instance, the Electronic Medical Records and Genomics (eMERGE) network [5] leverages expertise from multiple institutions and communities to integrate biorepositories and EHRs to support genomic research. Observational research approaches exhibit both potential and challenges for more sophisticated data analysis.

However, the acquisition of realistic data, especially data labels, is still restricted when confronting concerns about system security, patient privacy, and intellectual property protection [6,7]. Excluding data and labels may be ubiquitous during the data collection phase. Long-term studies often lack sufficient time to gather data and have no control over the switching behaviors of patients [8,9], resulting in the loss of accurate outcome measurements.

Restrictions on transferring intellectual property among different institutions hinder the sharing of data, which is expected to be complete. Additionally, some expertise-requiring annotations are tedious and have no guarantee of correctness [10]. Generally, label deficiencies occur frequently when analyzing observational EHR data.

There have been some attempts to address insufficient labeling by realistic synthesized EHR (RS-EHR) generation. One approach to RS-EHRs is knowledge-based [11,12]. Such approaches combine publicly available statistics, clinical practice guidelines, and medical coding dictionaries to improve the fidelity of generated EHRs. However, the models are still restricted to development, testing, and public demonstrations.

Another strategy is data-driven. Generative adversarial networks (GANs) are a new class of methods for obtaining realistic synthesized data [13,14]. The philosophy of GANs is to train two networks, one generating fake samples and the other discriminating fake and real samples, in a min-max game until equilibrium is achieved, indicating that the generated fake samples cannot be distinguished from the real samples. There has been some work on applying state-of-the-art GANs to generate synthesized EHR data sets [15,16]. However, these studies have not fully applied the generated data to augment EHR computational phenotyping and classification. GANs for few-labeled data are still unlikely to recover the whole distribution of labels from the raw data set due to imbalanced labeling. Additionally, there are some arguments that GAN-generated samples are likely to copy real samples exactly, which is a potential violation of privacy [17,18].

Semisupervised learning (SSL) is a set of techniques that are usually adopted to leverage unlabeled data and an underlying data set structure. With a relatively small set of labeled data compared to that needed in supervised learning (SL), SSL can still display decent learning performance. Some previous studies used SSL to phenotype EHR databases [19,20]. These studies achieved excellent performance in EHR-based risk prediction, but the feature dimensions were restricted to discrete medical codes. GANs were adopted to boost the SSL [21], but as mentioned above, the generator was trained to eventually remember exact copies of the samples for the limited span of an EHR data set in a discrete and high-dimensional space, which therefore raised privacy concerns. SSL is a powerful tool for label-deficient circumstances but needs specifications for observational research.

Network analysis is a solution to both obstacles. Encoding the similarities among patients into their connections protects their identities. The input of the analysis is only the network structure and embedded vectors. Network analysis is the basis for manifold learning, which has an advantage in approximating the data structure in a high-dimensional space. Many manifold-based methods have prevailed in intuitively visualizing and phenotyping coordinated data sets [22-24]. Additionally, there have been quite a few attempts to extend deep learning to irregular data structures, such as graphs. Several studies have shown great performance in representative learning with SSL [25-28], and endeavors have been made to apply GANs to graphs [29-31].

However, few studies have considered exploiting the inherent network structure of an EHR database in SSL tasks. GANs on networks have not been fully investigated in terms of privacy preservation. Additionally, under various label-deficient situations, the performance remains to be evaluated. It is very promising to scale SSL and GANs to the graph structure extracted from an EHR database and to thereby acquire a new perspective on EHR data sets.

For this paper, the main contributions are as follows: (1) This study tries to address limitations due to existing label deficiency in observational EHR analytical research by extending the network analysis pipeline to EHRs. A boosting learning model is proposed by applying GAN-boosted SSL to network data extracted from label-deficient coordinated EHRs. (2) Experiments are conducted on 3 public data sets as well as one from the First Affiliated Hospital of Zhejiang University, and they are evaluated by prediction metrics that are compared to conventional learning methods. The proposed method shows superior performance over conventional semisupervised methods and indicates comparable performance with supervised learning methods when data are fully labeled. (3) To ensure the utility of the proposed model, further evaluations of data quality, nondisclosure, and memory space consumption are performed.
The proposed method shows higher data fidelity, lower precision metrics against compromised attack, and less graphics processing unit (GPU) memory consumption over conventional semisupervised methods.

Methods

Data Set Structure Conversion to a Graph

Graph structure definition and semisupervised learning on graph formulation are shown in Multimedia Appendix 1A [31-37]. According to the well-accepted assumption that a manifold is locally Euclidean in topological space, it is plausible to represent a data set X with a graph G. However, this conversion rule should be scrutinized. First, it depends on the number of edges |E| that comprise the edge set. |E| should be restricted to a range that avoids disconnected components and short circuits that obscure structural information. Second, the neighborhood searching strategy should be scalable to feature value scales and effective in practice. Third, the local density variance should be preserved during conversion, which means that the weights of edges should not be binary.

To circumvent this problem, the k-nearest neighbors (k-NN) method was selected to convert the original data set into a graph measure space. As its name indicates, the k nearest points in the Euclidean space of point x are identified as its neighbors, N_k(x). Each edge weight w_ij is refined with the Jaccard coefficient:
\[ w_{ij} = \frac{\text{card}(N_i \cap N_j)}{\text{card}(N_i \cup N_j)} \]

The Jaccard coefficient addresses the unified weight problem brought by k-NN searching and restricts the weights to [0,1], which scales the local densities as node degrees: \( \text{deg}(v_i) = \sum_{j \in N_k(v_i)} w_{ij} \) Additionally, when the lower bound is reached, the edge is removed from the graph, and eventually, nodes with degree zero will be considered noise and therefore removed. The final graph serves as one of the inputs of the GAN.

GANs for Graphs and Their Modified Losses

In this study, we focus on generating vectorized fake samples by the use of both the graph structure and coordinated features. The coordinated features of the graph structure are acquired by feeding the Jaccard graph into large-scale information network embedding [38], explicitly setting the output dimension as half of d.

The fundamentals of GAN [32,33] are presented in Multimedia Appendix 1C [31-37]. Nevertheless, it is important to take into account the unconventional loss of semisupervised adversarial learning, as insufficient labels do not effectively minimize the current adversarial learning loss. The generator is trained to produce samples that bridge the density gap between samples from distinct classes. In the case of binary classification tasks, the 2 classes are “true” and “false.” By expanding the density gap between labeled true samples, labeled false samples, and generated density gap samples, the adjusted discriminator loss can enhance semisupervised learning performance. The refined discriminator loss \( L_D \) for SSL purposes comprises semisupervised loss, entropy loss, and class distance. (1) Semisupervised loss: there are two terms; the first is the supervised loss calculated by cross-entropy between the labels and prediction. The second emphasizes the loss due to incorrect classification by SSL. \( \lambda_0 \) is a hyperparameter that balances these 2 terms.
\[
\text{loss}_{\text{semi}} = \text{loss}_{\text{sup}} + \lambda_0 \text{loss}_{\text{un}}
\]
\[
= -E_{(y_i | x_i, y_i < M)} \log P(y_i | x_i) + E_{y_i \sim G(z)} \log P(M | x_i) \quad (2)
\]
(2) Entropy regularization [39,40]: this term calculates the entropy of a distribution over all labels M to enhance the certainty of the prediction.

(3) Cluster distance loss [31]: this term tends to enlarge the density gap so that samples from different classes are separate. \( h^G(x) \) is the last-layer output of the discriminator.

The final loss term for the discriminator generator is
\[
\text{loss}_D = \text{loss}_{D_{\text{Gwp}}} + \text{loss}_{\text{semi}} + \text{loss}_{\text{ent}} + \text{loss}_{\text{pt}} \quad (5)
\]
The loss of generator \( L_G \) is also modified by adding the term (4). The final loss term for the training generator is
\[
\text{loss}_G = \text{loss}_{D_{\text{Gwp}}} + \text{loss}_{\text{pt}} \quad (6)
\]
The network structure is illustrated in Figure 1. During the training phase, real embeddings and fake inputs of the same size are fed into the network separately in batches, with the goal of optimizing the aforementioned losses of the discriminator and generator. During 1 training epoch, batches of real labeled data, real unlabeled data, and fake data are fed into the network to calculate different loss terms for optimization. Batch normalization is conducted. After training, the discriminator loss is expected to be stable and could be exploited as a classifier for testing samples and predictions. The generator is suitable for measuring data quality and preserving privacy.
Data Sets and Experimental Setup

EHR data sets were obtained from public resources, including University of California Irvine Machine Learning Repository Type 2 Diabetes 30-Day Readmission (UCI-T2D) [41]; Surveillance, Epidemiology, and End Results Ovarian Cancer (SEER-OVC) [42]; and Surveillance, Epidemiology, and End Results Colorectal Cancer (SEER-CRC) [42]. The dimensional information is summarized in Table 1. Another colorectal cancer data set from the Second Affiliated Hospital Zhejiang University School of Medicine (SAHZU-CRC) was selected to investigate feasibility in practical situations. These data sets were selected because they are long-term follow-ups, the labels of which take much time and effort to obtain and are likely to be missing due to regulations on data collection. The selected features included basic demographics, medication, clinical codes, stage codes, laboratory variables, and dispositions. A basic description of the data sets and preprocessing is provided in Multimedia Appendix 1B.

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Records</th>
<th>Categorical variables</th>
<th>Numerical variables</th>
<th>Preprocessed dimensions</th>
<th>Labeling standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of California Irvine Machine Learning Repository Type 2 Diabetes 30-Day Readmission</td>
<td>61,675</td>
<td>44</td>
<td>8</td>
<td>57</td>
<td>Readmission in 30 days</td>
</tr>
<tr>
<td>Surveillance, Epidemiology, and End Results Ovarian Cancer</td>
<td>10,038</td>
<td>18</td>
<td>3</td>
<td>34</td>
<td>Survival over 5 years</td>
</tr>
<tr>
<td>Surveillance, Epidemiology, and End Results Colorectal Cancer</td>
<td>40,014</td>
<td>7</td>
<td>2</td>
<td>14</td>
<td>Survival over 5 years</td>
</tr>
<tr>
<td>Second Affiliated Hospital of Zhejiang University</td>
<td>1244</td>
<td>8</td>
<td>2</td>
<td>14</td>
<td>Survival over 5 years</td>
</tr>
</tbody>
</table>

Ethical Considerations

This study did not involve any human or animal experiments. The UCI-T2D, SEER-OVC, and SEER-CRC data sets are public, and we complied with their ethical requirements. We also used a colorectal cancer–specific disease cohort of the Second Affiliated Hospital Zhejiang University School of Medicine; this was approved by the Human Research Ethics Committee of Zhejiang University in August 2017 (2017-067).

Results

SSL-based Classification of EHR Data

In the aforementioned experiments, the proposed method for semisupervised classification outperformed related semisupervised methods by a decent margin. Basic graph
semisupervised methods (ie, GSSL) are limited in classification performance, mostly due to their assumption that edges encode only the similarity of nodes. The spectral methods (LP and GCNs) do not perform well, perhaps because their low-order approximation may smooth the frontiers in the graph. Neither of these 2 methods consider the local properties of the input graph, and under some circumstances, they classify the majority of nodes into 1 class. Additionally, at a 10% to 15% label rate, the proposed method achieves the best performance on SEER-OVC, SEER-CRC, and SAHZU-CRC (Table 2). The AUCs declined as label rates continued to rise. GCNs, as the state-of-the-art semisupervised deep learning method, had somewhat better results for a data set with a size that can be handled by a GPU, but still exhibited worse performance than the proposed method, presumably due to oversmoothing the graphs and having less refined loss.

In regard to supervised learning, as shown by the bars in Figure 2A, even with a label rate of 10%, SSL on a graph with a GAN performed comparably to the supervised learning methods. As the portion of labeled data increased, the learning performance progressively increased, which is a consequence of the abundant information of the label distribution over the constructed graph. However, as the label rate continued to rise, the performance decreased because of mode collapse and overfitting. As the error bars show, with 10% labeled data, the standard deviations of the proposed model are slightly larger, as shown in Figures 2A and 2C, indicating a limitation of our proposed method; it only applies to certain low label-rate circumstances. When the labels are sufficient, more robust SL methods are better. However, some poorly trained and undertuned SL methods show far worse metrics in testing. Additionally, as the vector dimensions, shown in Table 1, decreased somewhat, the learning performance showed a significant decrease. This is perhaps a consequence of the lack of dimension diversity for similarity encoding and the local graph structure.

Table 2. Summary of the results of the classification AUCs for semisupervised learning methods under progressively increasing label rates. The learning performance of the graph convolutional network on the large data sets—that is, data sets other than Second Affiliated Hospital of Zhejiang University Colorectal Cancer—is unavailable due to memory limits.

<table>
<thead>
<tr>
<th>Label Rate</th>
<th>University of California Irvine Machine Learning Repository Type 2 Diabetes 30-Day Readmission</th>
<th>Surveillance, Epidemiology, and End Results Ovarian Cancer</th>
<th>Surveillance, Epidemiology, and End Results Colorectal Cancer</th>
<th>Second Affiliated Hospital of Zhejiang University Colorectal Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>5% AUC</td>
<td>GSSL b</td>
<td>0.450</td>
<td>0.512</td>
<td>0.650</td>
</tr>
<tr>
<td></td>
<td>10% AUC</td>
<td>0.472</td>
<td>0.537</td>
<td>0.677</td>
</tr>
<tr>
<td></td>
<td>15% AUC</td>
<td>0.523</td>
<td>0.537</td>
<td>0.677</td>
</tr>
<tr>
<td></td>
<td>20% AUC</td>
<td>0.542</td>
<td>0.591</td>
<td>0.678</td>
</tr>
<tr>
<td></td>
<td>25% AUC</td>
<td>0.602</td>
<td>0.591</td>
<td>0.650</td>
</tr>
<tr>
<td>10% AUC</td>
<td>LP c</td>
<td>0.475</td>
<td>0.364</td>
<td>0.512</td>
</tr>
<tr>
<td></td>
<td>15% AUC</td>
<td>0.564</td>
<td>0.462</td>
<td>0.512</td>
</tr>
<tr>
<td></td>
<td>20% AUC</td>
<td>0.585</td>
<td>0.478</td>
<td>0.540</td>
</tr>
<tr>
<td></td>
<td>25% AUC</td>
<td>0.566</td>
<td>0.491</td>
<td>0.513</td>
</tr>
<tr>
<td>15% AUC</td>
<td>Proposed</td>
<td>0.929</td>
<td>0.719</td>
<td>0.640</td>
</tr>
<tr>
<td></td>
<td>10% AUC</td>
<td>0.979</td>
<td>0.677</td>
<td>0.678</td>
</tr>
<tr>
<td></td>
<td>15% AUC</td>
<td>0.964</td>
<td>0.677</td>
<td>0.678</td>
</tr>
<tr>
<td></td>
<td>20% AUC</td>
<td>0.930</td>
<td>0.678</td>
<td>0.650</td>
</tr>
<tr>
<td></td>
<td>25% AUC</td>
<td>0.924</td>
<td>0.650</td>
<td>0.650</td>
</tr>
<tr>
<td>20% AUC</td>
<td>GSSL b</td>
<td>0.525</td>
<td>0.527</td>
<td>0.640</td>
</tr>
<tr>
<td></td>
<td>10% AUC</td>
<td>0.547</td>
<td>0.527</td>
<td>0.640</td>
</tr>
<tr>
<td></td>
<td>15% AUC</td>
<td>0.564</td>
<td>0.512</td>
<td>0.640</td>
</tr>
<tr>
<td></td>
<td>20% AUC</td>
<td>0.585</td>
<td>0.512</td>
<td>0.640</td>
</tr>
<tr>
<td></td>
<td>25% AUC</td>
<td>0.578</td>
<td>0.581</td>
<td>0.581</td>
</tr>
<tr>
<td>25% AUC</td>
<td>LP c</td>
<td>0.540</td>
<td>0.532</td>
<td>0.652</td>
</tr>
<tr>
<td></td>
<td>10% AUC</td>
<td>0.540</td>
<td>0.532</td>
<td>0.652</td>
</tr>
<tr>
<td></td>
<td>15% AUC</td>
<td>0.540</td>
<td>0.512</td>
<td>0.640</td>
</tr>
<tr>
<td></td>
<td>20% AUC</td>
<td>0.540</td>
<td>0.512</td>
<td>0.640</td>
</tr>
<tr>
<td></td>
<td>25% AUC</td>
<td>0.540</td>
<td>0.512</td>
<td>0.640</td>
</tr>
<tr>
<td>5% AUC</td>
<td>Proposed</td>
<td>0.595</td>
<td>0.595</td>
<td>0.595</td>
</tr>
<tr>
<td></td>
<td>10% AUC</td>
<td>0.595</td>
<td>0.595</td>
<td>0.595</td>
</tr>
<tr>
<td></td>
<td>15% AUC</td>
<td>0.595</td>
<td>0.595</td>
<td>0.595</td>
</tr>
<tr>
<td></td>
<td>20% AUC</td>
<td>0.595</td>
<td>0.595</td>
<td>0.595</td>
</tr>
<tr>
<td></td>
<td>25% AUC</td>
<td>0.595</td>
<td>0.595</td>
<td>0.595</td>
</tr>
</tbody>
</table>

aAUC: area under the receiver operating characteristics curve.
bGSSL: graph-base semisupervised learning.
cLP: label propagation.
dGCN: graph convolutional network.
Figure 2. Summary of the results of the classification of the proposed method versus those of the conventional supervised learning methods. (A) AUC; (B) accuracy; (C) recall. The proposed method was evaluated under progressively increasing label rates. The supervised learning models were trained on fully labeled data. The error bars indicate the SD for each metric. AUC: area under the receiver operating characteristics curve; LR: logistic regression; RF: random forest; SAHZU-CRC: Second Affiliated Hospital of Zhejiang University Colorectal Cancer; SEER-CRC: Surveillance, Epidemiology, and End Results Colorectal Cancer; SEER-OVC: Surveillance, Epidemiology, and End Results Ovarian Cancer; SVM: support vector machine; UCI-T2D: University of California Irvine Machine Learning Repository Type 2 Diabetes 30-Day Readmission.

Boosting Semisupervised Learning by Generating a Density Gap

In this section, we visualize the final layer of discriminator D in the proposed method by feeding it real samples from UCI-T2D and their generated counterparts. By embedding the output layer at different iteration steps with t-distributed stochastic neighbor embedding [22], the progression of the density gap from the generated samples, described in equation 4, is verified.

In Figure 3, we can see that at the starting epochs, the generated samples are mixed with the real samples, and the different classes are not divided. During training, D gradually learns a nonlinear map to project the fake samples and real samples into distinct clusters, while G learns to generate samples to take over the central area and isolate the clusters of different classes. This process has 2 advantages. First, the fake samples from the generator are unlikely to be copies of the original data, avoiding the direct disclosure of private information. Second, the samples on the borders of different classes are more correctly divided, which not only improves the accuracy of classification but also reveals the underlying training strategy of splitting one large class into several smaller classes to obtain a better classification.
Figure 3. The progressive generation of density gaps in high-dimensional space and its visualization. (A) 0 epochs; (B) 40 epochs; (C) 80 epochs; (D) 120 epochs. The generated samples ultimately span the gap between a real and a false sample. The line chart (E) indicates the function of the pull-away term and how its optimization affects the generator and discriminator.

Fidelity Evaluation of the Generated Data
Frontier nodes are nodes at the borders of different clusters in a graph. The definition is given in Multimedia Appendix 1D. It is possible that a trained model is exploited directly for secondary purposes, such as fundamental profiling or developmental usage during the primary phase of data sharing [12]. We calculated the dimensionwise probability (DWPro) and dimensionwise prediction (DWPRe) proposed by Choi et al [15] to evaluate the fidelity of the generator in our proposed model. DWPro is a basic statistical confirmation of the distributions of real data that are appropriately learned by the generator in the model. A training set R and synthetic sample set S of the same sample size are compared using the Bernoulli success probability pk of each dimension k. DWPRe measures the extent to which the internal relations of every feature are captured. One dimension k is selected, and the rest of the features are used as training data. R and S are used to train the LR classifiers. Then, the dimension k is regarded as the label column for testing. It is a rational assumption that a smaller margin between the predictions of 2 models implies a better synthetic quality. The $F_1$-score is selected as the metric for comparison.

Figure 4 shows that all 4 data sets were depicted well from a featurewise perspective, and over half of the dots are near the diagonal line. In Figure 4C, the consistency of each feature indicates high synthetic quality. Figure 5 shows a mildly diminished learning quality considering interdimensional fidelity. However, half of the features are still likely to be inferred from the remaining columns and the same proportion of features. Considering that the generated frontier is still different from the directly generated datapoints [15,16], the fidelity is acceptable for some secondary uses.
Figure 4. Dimensionwise probability of 4 selected data sets: (A) University of California Irvine Machine Learning Repository Type 2 Diabetes 30-Day Readmission; (B) Surveillance, Epidemiology, and End Results Ovarian Cancer; (C) Surveillance, Epidemiology, and End Results Colorectal Cancer; (D) Second Affiliated Hospital of Zhejiang University Colorectal Cancer. The x-axis is the Bernoulli success probability for the features of the real data sets, while the y-axis is the corresponding value from the synthetic data. Each blue dot represents a feature of the data set. The red diagonal line indicates an identical Bernoulli success probability of both the real and generated data sets, and ideal fidelity is learned featurewise by the generator.

Figure 5. Dimensionwise prediction of 4 selected data sets: (A) University of California Irvine Machine Learning Repository Type 2 Diabetes 30-Day Readmission; (B) Surveillance, Epidemiology, and End Results Ovarian Cancer; (C) Surveillance, Epidemiology, and End Results Colorectal Cancer; (D) Second Affiliated Hospital of Zhejiang University Colorectal Cancer. The x-axis is the $F_1$-score of models trained on the real data sets, while the y-axis is the corresponding values from the synthetic data. Each blue dot represents a feature of the data set. The red diagonal line indicates that the $F_1$-score was identical for the models trained and tested on the real and generated data sets, and ideal interdimensional fidelity was learned by the generator.
Evaluation of the Disclosure Risk of the Generated Data

The generator in our proposed model may be exploited to generate data points similar to the original data sets, posing threats to patient privacy. We need to ensure that the frontier nodes generated by the proposed model can be protected from attackers with compromised data. Therefore, a quantitative evaluation of presence and attribute disclosure was conducted on the SAHZU-CRC data set. Of the real samples N, 1% were randomly sampled, and among the 11 dimensions (the numerical dimensions are left out and the nominal columns are collapsed into 2 for simplification), a progressively increasing number of features, denoted as r, were assumed to be known by the attacker. Then, the attacker could exploit the knowledge of the data (1% × N × r) to conduct k-NN searches of the synthetic data, and the other unknown feature values were estimated according to those of the k-NN. Finally, the unknown features were compared to the real features to gain precision and accuracy. The calculation was repeated 100 times with 1% of the real records chosen at random.

Under these circumstances, the sensitivity indicates when the attacker has 1% × N × r of the disclosed data and all the synthetic data and how many records of all the positive features the attacker can correctly estimate with a 1-NN attack. The precision indicates how many features among all the features estimated positive by the attack were on average accurate. For instance, in Figure 6A, if an attacker with 1% of the records (12 of 1244 records for SAHZU-CRC) and 5 features from the real data conducts 1-NN on the synthetic SAHZU-CRC data generated by the proposed method, the positive estimation of the remaining unknown features of the real data will be 12.5% correct on average (0.125 precision), and of all the positive predictions, 15.8% will be correct (0.158 sensitivity).

In summary, the precision and sensitivity of attacks on synthetic data is relatively low, 0.158 at best when r is 11. The most effective attack setting is 1-NN. It is difficult to estimate more information from our frontier nodes due to the modification of the network losses. Substitution of both the generator and discriminator learning strategies boosts the model performance on classification with label deficiency and provides synthetic samples capable of preventing disclosure.

Figure 6. Privacy preservation evaluation. When increasing the number of known features, the attack achieves (A) precision and (B) sensitivity with 1% compromised records from the Second Affiliated Hospital of Zhejiang University data set.

In summary, the precision and sensitivity of attacks on synthetic data are relatively low, 0.158 at best when r is 11. The most effective attack setting is 1-NN. It is difficult to estimate more information from our frontier nodes due to the modification of the network losses. Substitution of both the generator and discriminator learning strategies boosts the model performance on classification with label deficiency and provides synthetic samples capable of preventing disclosure.

Scalability of the Memory Consumption of Batch-Based Training

Because GPUs have been used in deep learning–based computation, we further examined whether the proposed method could achieve practical memory consumption. The aforementioned semisupervised methods are compared with our proposed method on memory consumption for 4 data sets. For the algorithms that do not need GPUs, their central processing unit consumption is measured.

For small data sets (eg, SAHZU-CRC), our proposed method takes up more space because of its complex network structure (Table 3). However, as the size expands, our proposed method shows the least and most stable space consumption, because minibatch training is independent of the number of samples (for SEER-CRC and SEER-OVC). Conventional network-based SSL methods tend to train on full batches. When the data set is large enough, there is a huge obstacle to storing the data in memory. Stable memory consumption implies a scalable model for training and prediction on diverse data sizes. The GCN, as a transductive SSL method, is unable to be directly scaled to larger data sets despite its excellent representative ability.
Discussion

Principal Results

The proposed model fully utilizes the inner graphical structure of EHRs and provides cost-effective prediction metrics. The density gap derived from the modified network loss enables different class labels to be better distinguished. Under label-deficient circumstances, the proposed model achieves a comparable performance to that of conventional supervised learning methods where all of the training data are labeled. Specifically, with only 10% labeled data, the performance of popular supervised machine learning methods is approached, which implies there is a broad set of situations in which this model could be considered for prediction tasks. Following the same setting of label rates for the purpose of comparison, the conventional SSL methods show poor data representation ability. The learning performance, compared to that of our proposed method, shows worse stability and scalability. With the increasing label rate, the conventional SSL models display either poor performance on classification due to label deficiency or extreme cases where the classifier puts every sample into 1 class as a consequence of overfitting. Additionally, the memory cost is worth noting. Most semisupervised methods have a tendency to copy the whole graph structure into memory [19,27,43], which brings a very large burden of computational resources considering that the EHRs absorb increasingly more data.

Extracting the frontier of generated samples that shows high performance in DWPro and DWPre has potential in applying some special frontier nodes as sample data for secondary usage, in the same way as related work applies GANs to RS-EHR generation. According to related studies [15,16], generating data with adequate quality is crucial in cross-organization data sharing. The quality of the data determines the model performance on realistic data sets. Additionally, for diverse developmental needs, the more realistic the generated data are compared to the real samples, the more persuasiveness and fidelity the researching systems will acquire. The generator in our model fulfills this demand by generating similar samples to the original data after the training phase.

To reveal the hidden clinical and physiological characteristics of certain groups, EHRs are among the most reliable information sources. Nonetheless, administrative regulations and the protection of patient privacy have decreased the accessibility of EHRs for a variety of reasons and made downstream analysis inconvenient. Our method first addresses privacy considerations by transforming the data set into a k-NN graph where the similarities between different patients are re-encoded while the identifying information is hidden. Second, the vector from the embedded graph is fed into our model for further analysis. Under practical scenarios, authorization to share and use the original data will not be a necessity. Additionally, even when conventional attacks attempt to reidentify personal information from the publicly generated samples, the k-precision and k-sensitivity metrics indicate that it is quite safe if the attacker holds only a small fraction of the knowledge of the real data and conducts the most powerful 1-NN attack. Furthermore, the density gaps avoid the usual case where GANs would otherwise be trained to copy the real input, thereby shielding the patient information from another possible method of disclosure.

Limitations

The limitations of this study are still worth noting. The evaluation of how the proposed model can improve data quality and predict performance on the actual label collection phase has yet to be considered. Additionally, we excluded all patient duplicates to conduct a prediction method without considering any temporal information. Further investigation of the temporal trajectories of the same patients may reveal more of the inner mechanisms of disease progression, and localization methods of temporal and spatial structure in many other fields may address the same problem [44,45]. Additionally, the proposed model only applies to some label-rate setups, and performance diminishes as more labels become available. The thresholds for switching between the different algorithms (SSL and SL) remain to be studied. Finally, to be more protective of patient privacy and intellectual property, our future explorations include graph generation and attention mechanisms [28,29,34,46]. A whole generated graph can be taken into consideration. With the power of GANs, the underlying structure of large-scale EHRs could be preserved while achieving full anonymity.

Conclusions

EHR-based systems and observational studies with conventional learning strategies are facing diverse challenges as data and...
label inaccessibility increase. Training on few labeled data is a pivotal task and needs substantial resources. Uncovering the underlying graphical structure of EHRs brings a motivating perspective and informative prerequisites to analyzing patient data. As a downstream analysis method, GAN-boosted SSL uses a graphical structure and greatly improves learning quality in label-deficient situations. GANs with refined loss also meet the demands of deidentification and decent data fidelity under multiple-source data-sharing circumstances. This combination achieves impressive performance on prediction metrics, data quality, and protection from compromising attackers over various data sets, while popular machine learning methods encounter obstacles to sufficient training. This study indicates the potential of discovering the structural features that underlie the data instead of merely feeding models coordinated data sets and using unlabeled data when label deficiency occurs.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (82172069), the Major Scientific Project of Zhejiang Lab (2020ND8AD01), and the Fundamental Research Funds for the Central Universities (226-2023-00050). We owe thanks to the staff of the National Cancer Institute and each member involved in the Surveillance, Epidemiology and End Results program.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Graph structure definition and semisupervised learning on graph formularization.

References


Abbreviations

AUC: area under the receiver operating characteristics curve
DWPRe: dimensionwise precision
DWPRo: dimensionwise probability
EHR: electronic health record
eMERGE: Electronic Medical Records and Genomics
GAN: generative adversarial network
GCN: graph convolutional network
GPU: graphics processing unit
GSSL: graph-based semisupervised learning
k-NN: k-nearest neighbors
LP: label propagation
LR: logistic regression
RF: random forest
RS-EHR: realistic synthesized electronic health record
SAHZU-CRC: Second Affiliated Hospital of Zhejiang University Colorectal Cancer
SEER-CRC: Surveillance, Epidemiology, and End Results Colorectal Cancer
SEER-OVC: Surveillance, Epidemiology, and End Results Ovarian Cancer
SL: supervised learning
SSL: semisupervised learning
SVM: support vector machine
UCI-T2D: University of California Irvine Machine Learning Repository Type 2 Diabetes 30-Day Readmission
Correction: Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study

Lama Alsudias\textsuperscript{1,2}, BSc, MSc, PhD; Paul Rayson\textsuperscript{2}, BSc, PhD

\textsuperscript{1}Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
\textsuperscript{2}School of Computing and Communications, Lancaster University, Lancaster, United Kingdom

Corresponding Author:
Lama Alsudias, BSc, MSc, PhD
Information Technology Department
College of Computer and Information Sciences
King Saud University
Prince Turki Bin Abdulaziz Al Awwal Road
Riyadh, 12371
Saudi Arabia
Phone: 966 118051044
Email: lalsudias@ksu.edu.sa

Related Article:
Correction of: https://medinform.jmir.org/2021/9/e27670
doi:10.2196/45742

In “Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study” (JMIR Med Inform 2021;9(9):e27670) the authors made one addition to the Acknowledgments section and changes to the Corresponding Author address and degree list.

The Acknowledgments has been changed from:

\textit{The authors thank Nouran Khalaf, who is a PhD student at Leeds University (mhk@leeds.ac.uk), for her help in labeling the tweets. LA wishes to thank King Saud University for funding her PhD study.}

To:

\textit{The authors thank Nouran Khalaf, who is a PhD student at Leeds University (mhk@leeds.ac.uk), for her help in labeling the tweets. This research project was supported by a grant from the “Research Center of College of Computer and Information Sciences”, Deanship of Scientific Research, King Saud University.}

Additionally, the Corresponding Author Address and degree list has been changed from:

Lama Alsudias, BSc, MSc
School of Computing and Communications
Lancaster University
InfoLab21
Lancaster
GB
Phone: 44 1524 510357
Email:lalsudias@lancaster.ac.uk

To:

Lama Alsudias, BSc, MSc, PhD
Information Technology Department
College of Computer and Information Sciences
King Saud University
Prince Turki Bin Abdulaziz Al Awwal Road
Riyadh, 12371
Saudi Arabia
Phone:: 966 118051044
Email: lalsudias@ksu.edu.sa

The correction will appear in the online version of the paper on the JMIR Publications website on February 3, 2023 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
“To Err Is Evolution”: We Need the Implementation Report to Learn

Caroline Perrin Franck1,2, PhD; Antoine Geissbuhler1,2,3, MD; Christian Lovis1,3, MD

1Campus Biotech, University of Geneva, Geneva, Switzerland
2Geneva Digital Health Hub, Geneva, Switzerland
3Geneva University Hospitals, Geneva, Switzerland

Corresponding Author:
Caroline Perrin Franck, PhD
Campus Biotech
University of Geneva
Chemin des Mines 9
Geneva, 1202
Switzerland
Phone: 41 787997725
Email: caroline.perrin@unige.ch

Abstract

JMIR Medical Informatics is pleased to offer implementation reports as a new article type. Implementation reports present real-world accounts of the implementation of health technologies and clinical interventions. This new article type is intended to promote the rapid documentation and dissemination of the perspectives and experiences of those involved in implementing digital health interventions and assessing the effectiveness of digital health projects.

(JMIR Med Inform 2023;11:e47695) doi:10.2196/47695

KEYWORDS
implementation science; knowledge management; knowledge sharing; digital health; implementation report

Introduction

The accelerating adoption of digital health, combined with evolving terminology and differing definitions, has created a paradox: there is an exponential increase in knowledge and information, but finding relevant data on implementation processes, and in particular errors or failures, remains a major challenge. This lack of effective documentation of implementation knowledge makes it difficult to reliably study and understand global trends in digital health project failures, exacerbated by a bias toward publishing mostly positive studies [1]. As a result, similar and avoidable mistakes are often repeated.

Digitalization offers numerous opportunities to improve the efficiency and equity of health systems. Yet, many digital health implementations stagnate in the pilot phase or fail to sustain or demonstrate impact. This not only wastes valuable resources but also further fragments already complex health systems, potentially leading to adverse health outcomes [2].

Recurrent errors contribute to inefficient implementation and scale-up, and it is crucial to consider that while new technologies offer immense potential benefits, they can also introduce risks or unintended consequences that have a direct impact on patient outcomes. For example, child mortality increased significantly at one site following the implementation of a commercial computerized physician order entry (CPOE) system [3]. This increase in mortality was primarily due to delayed administration of critical medications [4], amplified by a policy change shortly before implementation, suggesting that policy changes should be avoided during or in close proximity to a CPOE implementation process [5].

However, unintended consequences can also be positive, as demonstrated by the implementation of a telemedicine service in rural Nepal [6]. The service attracted a higher proportion of female patients, possibly due to cultural factors or minimal disruption to their daily lives, which suggests that telemedicine may improve access to health care for female patients [6].

Learning from past implementations is critical. Particularly from an ethical and human rights perspective, as in the absence of established implementation norms and best practices, implementers need to define their own standards for responsible and effective digital health implementations.

The Potential Transformative Impact of Implementation Reports

Sharing and connecting fragmented knowledge across institutional boundaries can revolutionize an industry. This is
evidenced by the aviation sector, a safety-critical industry like health care, where such collaboration has led to transformative results [7]:

Back in the 1930s, flying was really dangerous and passengers were scared away by the many accidents. Flight authorities across the world had understood the potential of commercial passenger air traffic, but they also realized flying had to become safer before most people would dare to try it. In 1944 they all met in Chicago to agree on common rules and signed a contract with a very important Annex 13: a common form for incidents reports, which they agreed to share, so they could all learn from each other’s mistakes. Since then, every crash or incident involving a commercial passenger airplane has been investigated and reported; risk factors have been systematically identified; and improved safety procedures have been adopted worldwide.

The aviation industry has set an exemplary precedent for how sharing mistakes can improve safety and build trust. However, much like the aviation industry in the 1930s, the digital health sector is still in the early stages: the potential to strengthen health care systems is clear, but the bigger picture and potential implications and safety risks are not yet fully understood.

Conflicts of Interest
None declared.

References

Abbreviations
CPOE: computerized physician order entry
Original Paper

Unique Device Identification–Based Linkage of Hierarchically Accessible Data Domains in Prospective Surgical Hospital Data Ecosystems: User-Centered Design Approach

Karol Kozak¹, Prof Dr; André Seidel², PhD; Nataliia Matvieieva³, MA; Constanze Neupetsch²,³, MA; Uwe Teicher², PhD; Gordon Lemme³, MSc; Anas Ben Achour³, MSc; Martin Barth¹, MA; Steffen Ihlenfeldt²,³, Prof Dr; Welf-Guntram Drossel¹,³, Prof Dr

¹Center for Evidence-Based Healthcare, Technische Universität Dresden, Dresden, Germany
²Fraunhofer Institute for Machine Tools and Forming Technology IWU, Dresden, Germany
³Professorship for Adaptronics and Lightweight Design in Production, Technische Universität Chemnitz, Chemnitz, Germany
⁴Fraunhofer Institute for Ceramic Technologies and Systems IKTS, Dresden, Germany
⁵Chair of Machine Tools Development and Adaptive Controls, Technische Universität Dresden, Dresden, Germany

Corresponding Author:
Uwe Teicher, PhD
Fraunhofer Institute for Machine Tools and Forming Technology IWU
Pforzheimer Straße 7a
Dresden, 01187
Germany
Phone: 49 351 4772 ext 2620
Email: uwe.teicher@iwu.fraunhofer.de

Abstract

Background: The electronic health record (EHR) targets systematized collection of patient-specific, electronically stored health data. The EHR is an evolving concept driven by ongoing developments and open or unclear legal issues concerning medical technologies, cross-domain data integration, and unclear access roles. Consequently, an interdisciplinary discourse based on representative pilot scenarios is required to connect previously unconnected domains.

Objective: We address cross-domain data integration including access control using the specific example of a unique device identification (UDI)–expanded hip implant. In fact, the integration of technical focus data into the hospital information system (HIS) is considered based on surgically relevant information. Moreover, the acquisition of social focus data based on mobile health (mHealth) is addressed, covering data integration and networking with therapeutic intervention and acute diagnostics data.

Methods: In addition to the additive manufacturing of a hip implant with the integration of a UDI, we built a database that combines database technology and a wrapper layer known from extract, transform, load systems and brings it into a SQL database, WEB application programming interface (API) layer (back end), interface layer (rest API), and front end. It also provides semantic integration through connection mechanisms between data elements.

Results: A hip implant is approached by design, production, and verification while linking operation-relevant specifics like implant-bone fit by merging patient-specific image material (computed tomography, magnetic resonance imaging, or a biomodel) and the digital implant twin for well-founded selection pairing. This decision-facilitating linkage, which improves surgical planning, relates to patient-specific postoperative influencing factors during the healing phase. A unique product identification approach is presented, allowing a postoperative read-out with state-of-the-art hospital technology while enabling future access scenarios for patient and implant data. The latter was considered from the manufacturing perspective using the process manufacturing chain for a (patient-specific) implant to identify quality-relevant data for later access. In addition, sensor concepts were identified to use to monitor the patient-implant interaction during the healing phase using wearables, for example. A data aggregation and integration concept for heterogeneous data sources from the considered focus domains is also presented. Finally, a hierarchical data access concept is shown, protecting sensitive patient data from misuse using existing scenarios.

Conclusions: Personalized medicine requires cross-domain linkage of data, which, in turn, require an appropriate data infrastructure and adequate hierarchical data access solutions in a shared and federated data space. The hip implant is used as an example for the usefulness of cross-domain data linkage since it bundles social, medical, and technical aspects of the implantation.
It is necessary to open existing databases using interfaces for secure integration of data from end devices and to assure availability through suitable access models while guaranteeing long-term, independent data persistence. A suitable strategy requires the combination of technical solutions from the areas of identity and trust, federated data storage, cryptographic procedures, and software engineering as well as organizational changes.

**KEYWORDS**
electronic health record; unique device identification; cyber-physical production systems; mHealth; data integration ecosystem; hierarchical data access; shell embedded role model

**Introduction**

Unique device identification (UDI) is a system used to identify devices within the health care supply chain based on a consistent, standardized, and unambiguous machine-readable identifier to keep track of the postmarketing performance of medical devices [1]. The performance of a hip implant, for example, cannot be evaluated without considering individual factors of the recipient [2] and the conditions of the therapeutic intervention [3]. Consequently, patient, medicine, and product have to be linked and monitored to enable a well-founded evaluation. Regardless of the still-missing legal framework conditions [4] and the existing ethical and political questions [5], digitalization of the health system is advancing [6], which is beneficial for a more holistic assessment. Either way, part of this development is the cross-domain linkage of data [7] that at least can be resolved on a patient-by-patient basis to prepare for intelligent data analysis. This requires going beyond analysis objectives to an appropriate data infrastructure, which enables different data domains to be linked while providing adequate hierarchical data access concepts. Consequently, this paper approached a framework for cross-domain cooperation and intelligent data analysis in a specific application scenario embedded in prospective digital hospital ecosystems. Linking unconnected domains inside of safe frameworks with clarified data sovereignty to enable a holistic approach to personalized treatment benefits health care, lowers risks of mistreatments, and functions as a catalyst for the optimization of medical products. This concept, based on a pilot scenario, can function as a basis to clarify unclear legal issues.

**Methods**

**Framework Conditions for Digitization in Individualized Medicine**

**Intelligent Data Analysis Framework**

An EHR is the systematized collection of electronically stored patient and population health data in a digital format. These records can be shared across different health care settings [8]. Records are provided through network-connected, enterprise-wide information systems or other information networks and exchanges. EHRs can include a range of data such as individual risk assessments, health monitoring, acute diagnostics, and therapeutic interventions while enabling comprehensive, intelligent data analysis, which, according to Hahn et al [9], is considered the information cycle of personalized medicine (Figure 1).
These data are stored as a patient digital twin in a centralized hospital information system (HIS) in the form of text, images (DICOM and other types), and scans. EHRs enable patients and hospitals to manage their health information in public (eg, hospital) and private environments as a personal health record (PHR). The information contained in EHRs is highly sensitive. Unintended exposure of these data threatens an intimate part of a patient’s private sphere and may lead to undesirable consequences. The EHR is a communication tool that supports clinical decision-making, coordination of services (illness type, care type), evaluation of the quality and efficacy of care, research, legal protection, education, and accreditation, and regulatory processes. It is the business record of the health care system, documented in the normal course of its activities. Patients routinely review EHRs and keep PHR in their own digital archive or in patient portals (eg, at health insurance companies such as “TK-Safe,” “Vivy,” or “AOK-Gesundheitsnetzwerk” in Germany), given the patient is the owner of the EHR and PHR. The physician, practice, or organization is the owner of the physical medical record because it is its business record and property, and the patient owns the information in the medical record. Although the record belongs to the care facility or doctor, it is truly also the patient's information. EHRs should be released to other stakeholders only with the patient's permission or as allowed by law or via studies; public registry and ministries. PHRs are already on the market. The purpose of a PHR is to maintain good health and target outcomes; this could include daily vital signs (eg, blood pressure, heart rate), number of walking steps, amount of exercise, and calorie intake. In addition to these, information for medical use might be considered, such as blood type, allergies, pre-existing diseases, medicines the user is taking, emergency contact information, and information about the user’s medical institution.

Synergy Potential in Information Linkage Using the Example of a Hip Implant

Neugebauer [10] stated that the digital transformation is promoted by the interaction of technologies that were previously perceived as independent of each other. Hahn et al [9] noted, in this context, that networking of the individual sectors and the structured use of integrated information are still pending. Moreover, Hahn et al [9] concluded that sustainable success can only be expected if the interlinking of technological and biomedical research on the one hand and clinical implementation and product development on the other hand are permanently guaranteed. Either way, this paper approached data-driven interdisciplinary research from an application-oriented perspective in an incident-based scenario (Figure 2). This example illustrates selected dependencies between social parameters, medical factors and technical aspects important for surgery, and healing, which are currently not linked sufficiently. In fact, it is a common practice to laboriously obtain this information on a case-by-case basis, which requires appropriate lead time before the operation; this is a major disadvantage in the case of emergency medical treatment. Consequently, this paper addressed this shortcoming and developed a possible solution scenario showing how this information can be linked at the EHR level. Moreover, we’ve shown how this information can be made accessible based on implant-inherent features while introducing a role model for access regulation and data protection.

Figure 2. Incident-based networking of social aspects, medical factors, and technical aspects. AM: additive manufacturing; CAD: computer-aided design; CAM: computer-aided manufacturing; CNC: computer numerical control; CT: computed tomography; MRI: magnetic resonance imaging.
Data-Driven Networking of Information

Regulatory Demands

In the medical technology industry, quality assurance is a particular focus due to the stringent regulations. With a changeover period, the European Medical Device Regulation (MDR) EU 2017/745 came into effect in May 2021 and replaced the European directives on medical devices. The regulation obligates manufacturers to mark medical devices that are marketed in the European Union with unique codes. The main objective for the introduction of these codes is to increase patient safety. Unique product identification prevents confusion of medical devices and makes counterfeiting more difficult. The markings are implemented by the UDI system. The UDI enables tracking of medical devices, for example, from the last step of postprocessing in manufacturing where the marking is applied to the component (eg, by laser engraving). Nevertheless, the medical products can only be tracked from this point through the logistics process to the hospital. Figure 3 illustrates selected phases of a medical device life cycle covering design, manufacturing, postprocessing, logistics, and the union with the patient. Unfortunately, doubtless identification of an implant after implantation is impossible with the UDI system, which largely excludes proof of quality and originality after implantation. Across industries, product piracy is a major problem, and the estimated economic damage has been increasing over the years [11]. In addition to the generally valid comments on product piracy and its consequences, other aspects have to be considered in the field of medical technology. After the publication of the “Implant Files” in 2018 by a group of investigative journalists, the problem of defective medical devices became a sociopolitical issue [12].

Component-Inherent Identifier–Based Data Access

Inherent markings of implants could be a solution to prevent counterfeiting while satisfying the regulatory requirements for a machine-readable marking in the form of a barcode or data matrix. In addition, traceability can be expanded back to the design process when the inherent markings are integrated. Moreover, this simultaneously enables inclusion of the manufacturing process into the traceability chain as well (see Figure 3). Either way, the greatest potential for innovation is seen in the ability to clearly identify the implant after surgery, preferably via noninvasive technologies already available in the hospital or technologies that can be provided without great technological effort and financial investment. Either way, the inherent feature could act as a key to access distributed information (see Figure 2) after signal processing and appropriate decoding (Figure 4).
morphological and mathematical operations; once decoded, they enable further linkage with data or information stored in a database (Figure 4).

Technical Focus Data

The starting point is the construction of a 3D model of a hip implant (Figure 5) based on anthropometric data or even patient-specific information in which the coding is also integrated. Considering typical hip implant sizes, geometric complexities, and quantities, 3D printing (additive manufacturing [AM]) is increasingly developing as a competitive manufacturing method [14]. Suitable AM procedures are primarily selective laser melting (SLM) [15,16] or electron beam melting [17,18]. In addition, layer-wise build-up and achievable resolution are very beneficial for the integration of inherent features during manufacturing (see Figure 3). Using these manufacturing processes, the 3D data model needs to be converted into a facet model first (mesh) [19]. Then, the parts are positioned in the build chamber, supported, and sliced using standard AM software [20]. Based on material selection criteria such as biocompatibility, the Young modulus, strength, and fatigue strength, a (certified) raw material is selected [21-23] that is particularly tailored to the AM process requirements [24]. Material selection criteria like particle size, particle size distribution, and morphology or chemical properties are continuously checked and monitored [25,26]. The SLM process data, for example, have to be qualified for applications like implants and are subdivided into predefined parameters (considered static) and parameters to be controlled by in-process sensing (considered dynamic) [27]. Either way, essential parameters are monitored and archived [28-30]. The same applies to processing data from postheat or pressure treatment [31] as well as from destructive and nondestructive material testing [32]. Moreover, mechanical postmachining, performed to adapt the implant to the recipient needs, generates data [33]. An example of this is the description of the geometric interface to the patient (bone-implant interface) including parameters such as surface roughness. The result, however, is an extensive description of the implant and the implant creation process as well as additional information such as corresponding implant tools (Figure 6). Obviously, some of this information is of interest to surgeons (medical focus), whereas information about the implant recipient and the use of the implant (social focus) are of interest to the manufacturer (Figure 2). This means, regardless of the individual legal framework, the possibility of controlled linkage of information is seen as desirable.

Figure 5. Example of the component (part) identification using a Pharmacode integrated into the hip prosthesis and data extraction from the identifier in the implanted state by computed tomography. CAD: computer-aided design; LMS: laboratory management system.
**Social Focus Data**

The social background of an implant-receiving patient (Figure 2) includes individual characteristics and conditions such as age, gender, lifestyle, and constitution type, which is of crucial importance for the formation of diseases, their duration, and treatment. The continuous monitoring or even documentation of this background can be ensured by a variety of technologies from the field of mobile health (mHealth), which allow broad mapping of dynamic data sets such as lifestyle and physical activities [34]. mHealth is an aspect of eHealth, although there is no universal definition of mHealth [1]. However, there is consensus that mHealth can be understood as medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring devices, personal digital assistants, and other wireless devices [35]. This means that mHealth can be understood as “the use of mobile communications for health information and services” as patient-individual behavior without direct involvement of the health service provider [36]. Here, mHealth is seen as the technical prerequisites to monitor and document the social focus data during healing (Figure 7) or even beyond (see Figure 2).
In addition to the fact that mobile communication and audiovisual interaction are the decisive enablers for mHealth, there is also the aspect of powerful sensor-based hardware and high flexibility in software development for smartphones and, to a limited extent, wearables [37]. In contrast to smartphones, wearables can be designed very specifically and can therefore be used for special applications (eg, sleep monitoring) [38]. Widespread sensor systems for smartphones include [39-41] light sensor technology (ambient light, camera system in combination with lighting), proximity sensors, acceleration sensor technology, rotation sensor technology (gyroscope), electromagnetic sensor technology, digital compass (magnetometer), acoustic sensor technology via microphone, and sensor technology for location tracking (GPS).

In contrast to smartphones, wearables are more specific for a particular application and therefore have higher specificity for the integrated sensors, so that headbands can, for example, derive targeted electroencephalography [42]. This means that targeted data collection is possible with the help of sensor technology, internal processing by specific software (eg, apps), visualization based on these data, and mostly wireless communication to third parties in the form of a uniform data image. For example, the following parameters can be acquired by means of wearables and mobile devices: heart rate and pulse oxymetry readings with photoplethysmography [43] and systems to monitor activity and sleep [11]. Hence, mHealth is seen as an enabler that contributes to rehabilitation by providing valuable data about the rehabilitation measures and patient-specific activities (Figure 6) that can be stored in the EHR. In addition, realistic load scenarios can be determined that could contribute to the further optimization of hip implants (Figure 2)—to promote healing [44] while targeting shorter hospital stays and lower treatment costs, for example [45]. In addition, mHealth can help achieve a consistent database (Figure 2) to evaluate the optimal intensity, frequency, and effects of rehabilitation from a wide variety of patients over a longer period of time as these data are currently not available or insufficient [46]. Either way, linking social focus data and production data could enable significant improvements with regard to determining the actual wearing of the implant, for example, based on characteristics such as posture, weight, and movement profiles, which are summarized here as social focus data (Figure 2).

Medical Focus Data
Both routine diagnostics and revision surgery require information about implants that were placed decades ago. Specific identification using only medical imaging is currently not possible and requires access to the documentation by the initial treating physician or hospital (Figure 2). For elective as well as acute medical interventions, this documentation is not available or only available with enormous effort. With inherent markings in the implant, it is feasible to obtain information relevant for revision surgery or routine diagnostics even after
the insertion of an implant into the human body [47] (Figure 4). To assure valid and reliable surgical planning, medical data from past treatments, especially the surgical processes, follow-up examinations, and rehabilitation measures, are needed. Insufficient information about the implant and medical focus data (Figure 2) could lead to complications during revision surgery; therefore, considerable efforts are being made to obtain this information, which significantly extends the preoperative time in the hospital and results in additional costs. In fact, an example of the necessary information about the particular implant concerns the appropriate revision instruments and the existing implant components [48-50]. Moreover, the research effort prior to the operation is continuously increasing, linked to the increasing number of operations and growing variety of implant types, sizes, variants, and material combinations. In addition to knowledge on the implant system used, information on the initial implantation process, which cannot be seen in medical images (CT, x-ray), is highly relevant for a gentle and successful revision of a hip implant. In fact, insufficient information clearly increases the risk of complications for the patient. Here, it shall be emphasized that the use of unsuitable revision instruments causes the risk of an enlargement of the wound surface resulting from an invasive procedure. This, in turn, increases the risk of infection and bleeding or even periprosthetic fracture. In addition, the inevitable prolongation of the surgery time can lead to a higher anesthetic risk, increased risk of thrombosis, and unstable cardiovascular function. Consequently, there is the obligation to report “incidents” in connection with medical devices to the German Federal Institute for Drugs and Medical Devices (BfArM), which includes an indication of the cause. Hence, unique inherent identification (Figure 4) and (partial) networking of information (Figure 2) in a transparent and retraceable database seem promising to assure a higher quality of health care [11], if unauthorized access is avoided.

Results

Database and Data Access

Data Integration From Heterogeneous Domains

Data integration is a crucial issue in the environment of heterogeneous patient-production data sources (Figure 2). First, there are heterogeneous data types and formats located in different databases, which implies that solving data integration challenges is a prerequisite for gaining useful information and knowledge based on appropriate analytical methods. Although concepts for databases to process heterogeneous data sets exist [51], the Laboratory Management System 4.0 (LMS 4.0) was developed specifically for the purpose of taking into account nonmedical stakeholders.

Exemplary implementation and application of this database structure were successfully demonstrated for the case of operations on the lip-jaw-palate region [52]. LMS 4.0 enables requesting data from different locations (eg, surgery, the HIS, or an implant producer) as a routine using web user interfaces. Using LMS 4.0, the surgeon collects magnetic resonance imaging results, for example, from HIS, checks patient data stored there, and has access to the integrated technical focus data (Figure 8) while planning the operational approach.

Figure 8. Partial networking of distributed data and information from different domains. AM: additive manufacturing; API: application programming interface; CAD: computer-aided design; CAM: computer-aided manufacturing; CNC: computer numerical control; CT: computed tomography; HIS: hospital information system; MRI: magnetic resonance imaging.

Based on integrated patient and production data, which can be expanded to an implant database that covers several manufacturers, surgical planning is simplified in order to determine which bone implant is the most suitable in the particular case, for example. Moreover, LMS 4.0 generates a dashboard and report that help the operating staff prepare for
the surgery with the selected implants through the automatic output of the associated tools and instruments. Moreover, the surgeon can use this report to prove his or her preparations for the procedure and comprehensively explain the operation to the patient. LMS 4.0 presents an architecture that implements data integration in the hospital from the production, surgery preparation, and patient data. LMS 4.0 integrates databases without any changes to the individual databases (SQL database, software back end, application programming interfaces [APIs], front end) nor any need to maintain another database. The solution combines database technology and a wrapper layer known from extract, transform, load (ETL) systems and brings it to the SQL database, WEB API (back end) layer, interface layer (REST API), and front end. It also provides semantic integration through a connection mechanism between data elements. The solution allows for integration of patient, surgery, and production data in one technological framework: data management platform and implementation of analytical methods in one end user environment. The patient data (see Figure 2) are transferred, securely, to a HIS. Medical data storage in LMS 4.0 offers a highly scalable clinic web storage service that uses cumulative digital objects (eg, patient, surgery, implant) rather than blocks or files. Object storage typically stores data, along with metadata that identify and describe the content. For metadata management and automated quality control and data fusion (ETL processes), a data consistency model (LMS I4.0 metamodel) is used to enable eventual consistency for updates or deletes to existing objects.

**Role Concept for Secure Data Access**

On the basis of the information domains and the links and interfaces shown in Figure 2, it becomes obvious that the database structure and the information technology (IT) system design in the back end (see Figure 8) have to accommodate different user roles to protect secure data access to sensitive patient data. Consequently, a role model was developed that takes into account both different users or user groups (eg, patients, medical staff, manufacturers of medical products, and first aid providers) as well as special situations (eg, emergency access). Deviating from static access models (role-based access control) as well as traditional shell models in this case, pure login information was linked with additional contextual information (attributed-based access control) in order to allow hierarchical access control. Either way, the principle is shown in Figure 9, indicating the 4 basic roles embedded in the defined shell model.

**Figure 9.** Principle of hierarchical data access based on a shell embedded role model.

Nevertheless, the person category is subdivided into 4 roles: “first responder,” “producer of medical products,” “medical staff,” and “user.” Users can view all data via a terminal device after registration, for example, with a digital health card equipped with a radio-frequency identification transponder (radio-frequency identification), and allowed to only make entries in a dedicated area of the social focus data section (Figure 2). The write permissions include data integration from fitness watches, training performance in rehab or sports facilities, and self-collected nutritional data, while interfaces to cell phone apps, for example, are available to substitute manual input. On the other hand, the medical staff can read medically relevant data and has the right to make entries in the medical focus data (Figure 2), which shows who made the entry. This corresponds to an entry in the EHR stored in the HIS (Figure 7). Producers of medical products only have access to the medical device data area, summarized as production (Figure 7). Technical focus data (Figure 2; eg, appropriate revision instruments as described in the Medical Focus Data section) are stored here, while the patient and implant are linked in the medical focus database (Figures 2 and 7). This information is requested from the manufacturer via a modified procurement process, which ensures that the agreed data are available before the invoice for the implant is paid. Consequently, all relevant product information is stored, enabling the simplification of follow-up treatments, support for minimally invasive interventions, and excluding of medical interaction. The information transfer to the manufacturer (see the Technical Focus Data section), on the other hand, can be enabled using a data integration center with a data use and access committee for research inquiries, as is currently being developed by Prokosch et al [53]. The last role in the person category is the first responder (Figure 8), which is introduced to explain the dynamic access approach. The first responder occurs in case of an accident or emergency when lifesaving measures, for example, are necessary. For example, access is granted for a certain period if several predefined factors that were detected using a fitness watch or other smart device take effect at the same time (eg, oxygen saturation in the blood, blood...
pressure, or other health-endangering characteristics). However, these attributes are securely transferred to the LAB 4.0 database management system (see the Data Integration From Heterogeneous Domains section) to obtain the necessary information depending on the authorization or to allow the addition of data. For this purpose, a standardized, well-defined interface is used to realize the data exchange and integrate smart devices for pure information retrieval as well as to develop software extensions that can be used to store the data in the database while complying with access restrictions. This enables the creation of a digital ecosystem for different participants to provide patients with optimal and, above all, digital, end-to-end health care while providing adaptive access regulations meeting authenticity requirements and assuring authenticity and appropriate access tracking. A method of secure patient-centered management of EHR data, though it can also be further processed in a deidentified format for statistical purposes, has been demonstrated with blockchain technology using cancer care as an example [54].

**Data Integration Scenario**

Using the hospital database LMS 4.0 (Figure 8), all individual elements of the data ecosystem are presented and explained in reference to **Figure 10.** In fact, the data ecosystem is divided into the following 4 levels, which are distinguished for functional structuring: “data storage,” “data harmonization,” ”interfaces,” and ”data input/data output.” The ”data storage“ level contains different relational databases, which, again, contain medically relevant data (medical device [see the Technical Focus Data section], patient record [see the Medical Focus Data section], and health [see the Social Focus Data section]). Data preprocessing is performed at the ”data harmonization“ level (see Figures 8 and 10), which means that incoming data are adapted to the requirements of the LMS 4.0 database structure (Figure 8) and sorted, while outgoing (anonymized) data (eg, via the data integration center) are transferred via defined data exchange procedures with strictly recorded accesses. In the underlying, but closely related, ”interfaces“ level (Figure 10), interfaces are established by extensible middleware to communicate with the ”hospital database“ (see Figure 8). This enables the integration of data users from different domains and querying data from the database. The 3 levels ”interfaces,” ”data harmonization,” and ”data storage“ are subjected to the CIA (confidentiality, integrity, availability) triad and ensure the functionality of the system. The ”data input/data output“ level connects the ”hospital database“ with the environment. For example, manufacturers of medical devices can transfer product-specific data into the database to make the data available to hospital staff (see Figures 3 and 10). Likewise, patients can store their vital signs from wearables, for example, in this database to support long-term examinations or enable access in medical emergencies through attribute-based access control (see the Role Concept for Secure Data Access section). At the same time, patients can see their EHR, read digital doctor's notes, or view exam results. Physicians have interfaces to both connect medical exam machines to the database and write data; medical staff can also store information, and the data can be viewed hospital-wide and processed with appropriate IT systems.
Discussion

Principal Findings

The connection of different data lakes, beginning with implant design including the entire manufacturing to the medical treatment process as well as the tracking of lifetime characteristics inside of a data integration ecosystem, shown in this paper can disrupt today’s health care system, leading to a cost-efficient, personalized system. The strict hierarchical data access concept based on a shell-embedded role model can be used to handle highly sensitive data and as a template to help clarify legal issues.

The overriding goal is to use digitization to improve the networking of interdisciplinary domains and to create secure interfaces for exchange as a prerequisite for intelligent data analysis. For this purpose, a representative hip implant application scenario was chosen due to an existing network of social, medical, and technical domains. Moreover, the interaction resulted from a skill- and experience-based union of implant and recipient results in an individual constellation that is subject to change over time. A key element in the resulting constellation is the UDI based on an inherent feature, which can be read out noninvasively after implantation. The permanently readable feature acts as a key to technical focus data, which represent testable or documented properties that are made available by the provider and are of direct or downstream interest.

Consequently, we showed how the technical focus data can be integrated into existing data ecosystems. This, however, was only approached at the hospital level, which is explained by the unclear legal framework and the missing data infrastructure for a broader context. Nevertheless, the consolidation of distributed databases in a single technology solution is a scalable concept that can be transferred from a single hospital to a global solution. Another important aspect is that the introduced hierarchical data access is based on a shell-embedded role model and staggered user rights. Here, the attribute-based access control shall be emphasized because this represents nonrigid boundary
conditions in preparation for future regulations. The selected user profiles and the granted rights, on the other hand, are only examples that are up for discussion and need to be specified and challenged in further research. However, the data integration scenario distinguishes 4 levels of action layering data storage, data harmonization, interfaces, and the data input/data output layer, which harmonizes the application scenario and digital ecosystem. Nevertheless, future research must show how real benefit can be created through data linkage and how this can be monetized. Balancing the personal rights of the individual while achieving sustainable technological innovation is seen as the central challenge, which must be faced in a global context.

Conclusions

Personalized medicine requires cross-domain linkage of data, which, in turn, requires an appropriate data infrastructure and adequate hierarchical data access solutions.

Hip implant is a prime example of the usefulness of cross-domain linkage of data because it bundles social factors of the individual patient, medical aspects in the context of the implantation, and technical aspects of the implant.

UDI in terms of inherent identifiers can be the key to (selective) long-term data access especially if the postoperative readout is guaranteed.

Acknowledgments

“Unique Device Identification Based Linkage of Hierarchically Accessible Data Domains in Prospective Surgical Hospital Data Ecosystems” is part of the Fraunhofer Light-house Project “futureAM - Next Generation Additive Manufacturing” funded internally by the Fraunhofer-Gesellschaft e.V. Partial sponsorship was received from the Else Kröner Fresenius Center for Digital Health of the TU Dresden. The article processing charge was funded by Fraunhofer.

Authors’ Contributions

KK, AS, and UT conceptualized the study. AS designed the methodology and wrote the initial manuscript draft. KK and GL developed the software. KK, NM, CN, and GL performed the investigation. SI and WGD provided the resources. KK, AS, NM, CN, UT, GL, ABA, SI, and WGD reviewed and edited the manuscript. AS, UT, KK, NM, CN, ABA, and GL created the visualizations. AS and SI supervised the study. KK and AS provided project administration. SI and WGD acquired the funding. All authors read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

References


Abbreviations

AM: additive manufacturing
API: application programming interface
CIA: confidentiality, integrity, availability
CT: computed tomography
EC: eddy current
ETL: extract, transform, load
HIS: hospital information system
IT: information technology
LMS: Laboratory Management System
MDR: Medical Device Regulation
mHealth: mobile health
PHR: personal health record
SLM: selective laser melting
UDI: unique device identification
US: ultrasound

Edited by C Lovis; submitted 02.08.22; peer-reviewed by N Karnik; comments to author 10.11.22; revised version received 23.11.22; accepted 24.11.22; published 27.01.23.

Please cite as:
Unique Device Identification–Based Linkage of Hierarchically Accessible Data Domains in Prospective Surgical Hospital Data Ecosystems: User-Centered Design Approach
JMIR Med Inform 2023;11:e41614
URL: https://medinform.jmir.org/2023/1/e41614
doi:10.2196/41614
PMID:36705946

©Karol Kozak, André Seidel, Nataliia Matvieieva, Constanze Neupetsch, Uwe Teicher, Gordon Lemme, Anas Ben Achour, Martin Barth, Steffen Ihlenfeldt, Welf-Guntram Drossel. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 27.01.2023. 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.
Establishment of a Public Mental Health Database for Research Purposes in the Ferrara Province: Development and Preliminary Evaluation Study

Maria Ferrara\textsuperscript{1,2,3}, MD, PhD; Elisabetta Gentili\textsuperscript{4}, MA; Martino Belvederi Murri\textsuperscript{1,2}, MD; Riccardo Zese\textsuperscript{5}, PhD; Marco Alberti\textsuperscript{6}, PhD; Giorgia Franchini\textsuperscript{5}, PhD; Ilaria Domenicano\textsuperscript{1}, PhD; Federica Folesani\textsuperscript{1,2}, MD; Cristina Sorio\textsuperscript{2}, MPH; Lorenzo Benini\textsuperscript{2}, BA; Paola Carozza\textsuperscript{2}, MD; Julian Little\textsuperscript{8}, PhD; Luigi Grassi\textsuperscript{1,2}, MD

\textbf{Corresponding Author:} Maria Ferrara, MD, PhD

\textbf{Abstract}

\textbf{Background:} The immediate use of data exported from electronic health records (EHRs) for research is often limited by the necessity to transform data elements into an actual data set.

\textbf{Objective:} This paper describes the methodology for establishing a data set that originated from an EHR registry that included clinical, health service, and sociodemographic information.

\textbf{Methods:} The Extract, Transform, Load process was applied to raw data collected at the Integrated Department of Mental Health and Pathological Addictions in Ferrara, Italy, from 1925 to February 18, 2021, to build the new, anonymized Ferrara-Psychiatry (FEPSY) database. Information collected before the first EHR was implemented (ie, in 1991) was excluded. An unsupervised cluster analysis was performed to identify patient subgroups to support the proof of concept.

\textbf{Results:} The FEPSY database included 3,861,432 records on 46,222 patients. Since 1991, each year, a median of 1404 (IQR 1117.5-1757.7) patients had newly accessed care, and a median of 7300 (IQR 6109.5-9397.5) patients were actively receiving care. Among 38,022 patients with a mental disorder, 2 clusters were identified; the first predominantly included male patients who were aged 25 to 34 years at first presentation and were living with their parents, and the second predominantly included female patients who were aged 35 to 44 years and were living with their own families.

\textbf{Conclusions:} The process for building the FEPSY database proved to be robust and replicable with similar health care data, even when they were not originally conceived for research purposes. The FEPSY database will enable future in-depth analyses regarding the epidemiology and social determinants of mental disorders, access to mental health care, and resource utilization.

\textit{(JMIR Med Inform 2023;11:e45523)} doi:10.2196/45523

\textbf{KEYWORDS}

mental health; psychosis; epidemiology; electronic health registry; health care; machine learning; medical health records; electronic health records; clinical database; support; mental disorder; social determinants; mental health care; resource utilization

\textbf{Introduction}

Electronic health records (EHRs) assemble and enable access to large volumes of clinical and sociodemographic data that are routinely collected by local health authorities. EHRs offer a unique opportunity to conduct research on various topics, including, among others, the patterns of health care resource use and factors that influence the course and outcomes of mental disorders in large, representative samples \cite{1,2}. EHRs can be linked to data related to census and geolocalization information \cite{3}; such investigations span the epidemiology of mental disorders, hospitalization rates, morbidity, and mortality.

The breadth and nature of information represented in the sample of EHRs in the mental health sector make such information...
particularly suitable for using artificial intelligence (AI) and machine learning techniques, in addition to traditional methods (eg, linear regression models), in order to increase the potential for research on social and clinical factors [4].

Applications that use AI take advantage of AI's ability to process large amounts of data in order to extract information or identify underlying patterns of relationships that conventional methods may overlook [5]. AI may be particularly suitable for the investigation of large amounts of clinical data, thanks to (1) the flexibility and scalability of AI techniques, which are higher than those of traditional methods, and (2) the ability of AI to consider all of the available predictors (ie, not only a subset), which makes AI and, in particular, machine learning suitable for performing tasks such as classification, prediction, and resource optimization [5,6].

Indeed, in recent years, the use of AI techniques in mental health care research have rapidly increased, including its use to identify a disease at its earliest stages, predict illness onset in vulnerable individuals, study illness progression, optimize treatment, and discover novel therapeutic agents [7,8].

As of yet, there are few examples (mainly from the United States) of how data collected from EHRs can be successfully adapted for analysis with AI. For example, Hughes et al [9] analyzed clinical variables in the EHRs of 81,630 adults from 2 academic medical centers in Boston, Massachusetts (spanning 10 years) and identified predictors of treatment response for major depressive disorders.

Xu et al [10] compiled a data set of 11,275 patients from 5 large medical centers across New York City by using EHR data collected between 2008 and 2017; they used machine learning methods to identify markers of depression phenotypes to inform clinical decisions about patients' care. Pradier et al [11] analyzed a data set of 67,807 individuals to predict the risk of receiving a misdiagnosis of bipolar disorder among individuals with depression, using only information retrieved from EHRs. Perlis et al [12] applied natural language processing to classify the mood states of 127,504 patients, using data from an EHR.

In order to fully exploit the potential of EHRs for mental health research however, important issues need to be considered. One preliminary, controversial issue is whether the use of EHR data should be restricted to the purpose for which they were collected [13]. Indeed, privacy constraints, data security, and overall ethics regulation must be taken into account when considering whether to use EHRs for research purposes [6]. Nonetheless, nowadays, medical data that were originally collected for purposes other than research are being used to study health phenomena in many different fields, including mental health, substance use, noncommunicable diseases (eg, cancer), and health behaviors (eg, cancer screening) [14]. A further challenge is that data may not be homogeneous or may not be collected systematically, and most data are not derived from structured scales or questionnaires. The adaptation of the EHR represents the first necessary step to planning research projects that include models for predicting health resource utilization, identifying predictors of diagnostic accuracy, and differentiating between remission and chronicity, as done in other fields such as oncology [15].

Given this premise, the aim of this paper is to describe (1) the challenges and pitfalls that were encountered in the process of adapting EHR data derived from the public mental health agency in Ferrara, Italy, for research purposes and (2) the development of a data set that is suitable for analysis via AI and traditional techniques. In order to test the feasibility of using these data in analyses and the robustness of analyses based on such data, a clustering analysis was also performed, and preliminary results are presented herein.

## Methods

### Ethics Approval

Ethical approval was obtained by the Area Vasta Emilia Centro Ethical Committee on December 12, 2019 (protocol number: 197/2018/Oss/AUSLF). This study conforms to the principles expressed in the Declaration of Helsinki.

### Setting

In Italy, mental health care is provided by departments of mental health [16-19]. The levels of care within each department of mental health include community-based mental health centers, hospital psychiatric inpatient units, and rehabilitation or residential facilities. Each community-based mental health center serves as a hub of psychiatric care for geographically defined catchment areas with 50,000 to 150,000 inhabitants [20]. In Ferrara, Northern Italy, the Integrated Department of Mental Health and Pathological Addiction covers an area of 2630 km², with a catchment of 342,061 inhabitants as of 2020 [21].

### Data Collection and EHRs

Data were collected in 2 periods that were distinct in terms of the methodology used, the psychiatric services delivered, and the level of digitalization. Data related to the first period, which began in 1925 and ended in 1990, were gathered mostly in a psychiatric asylum, during a time when digital health was not fully developed or adopted.

In 1991, the first structured EHR (ie, SIPER [Sistema Informativo Psychiatrico dell'Emilia-Romagna]) was introduced and implemented locally by the Local Health Trust of Ferrara for Mental Health in Adults. Different software programs were adopted during the years following the implementation of SIPER, and each new software program replaced the previous one by importing already existing data and adding new features (and thus information), as detailed in Textbox 1.
Data Preparation

The first goal was the creation of a new, fully deidentified database with data available in EFESO (Newteam SRL)—a necessary step for complying with privacy constraints.

In order to remove all protected health information (PHI), source data needed to be modified. This could not be done by directly editing data in EFESO, since the source could not be altered directly. Thus, the new research database—the Ferrara-Psychiatry (FEPSY) database—was built via the Extract, Transform, Load process, which is a 3-phase process [22] in which data are first extracted from 1 source or multiple and possibly different sources (eg, databases, flat or formatted files, and web pages). Afterward, the extracted data are stored in a staging area, where they undergo transformation, such as filtering, cleaning, summarization, and normalization. Finally, the transformed data are loaded into the destination storage. For example, one type of transformation was record exclusion. We excluded records of patients that could not be unequivocally identified by the tax code—a unique 16-digit alphanumeric code that identifies a person in Italian public administration forms.

While assessing the suitability of the FEPSY database for research purposes, we noted that historical information dating back up to 1925 had also been maintained in EFESO. We understood that such data were manually imported into the electronic databases that preceded SIPER (year 1991); however, because we could not confirm the procedures, scope, and quality of this historical data import, we decided to document the existence of these data but exclude them from analysis.

More details about the FEPSY database can be found in Multimedia Appendix 1. In Table S1 in Multimedia Appendix 1, for each table in the FEPSY database, the total number of rows (number and percentage of records retained in the FEPSY database) is reported, alongside the number of records in the corresponding original EFESO table from which data were extracted (number of records in the corresponding EFESO table). As detailed in Table S1 in Multimedia Appendix 1, of the 4,264,954 records, 3,861,432 (90.54%) were kept. These records included detailed information about the patient, their illness, and the treatments provided.

In Table S2 in Multimedia Appendix 1, 2 types of anomalies for each table in the FEPSY database are described; one is date inconsistency (eg, when the closing date precedes the opening date of the medical chart), and the other is a date anomaly that was generated by the automatized mechanism that was introduced by EFESO when migrating data from IPPOCRATE (GPI SRL; August 26, 2008).

Clustering

Once the anonymized database was built, a clustering analysis was performed to investigate the data set quality. A clustering algorithm is an unsupervised machine learning technique that is used to group objects, so that objects of the same group (or cluster) are very similar to one another and objects of different groups are very dissimilar. To decide the degree of similarity (or dissimilarity) between 2 objects, various distance measures can be used, such as the Euclidean distance between (normalized) numerical representations of the objects. We tested the hypothesis that the patients modeled in the data set could be divided into homogeneous clusters. The k-means algorithm computes a numerical distance between objects to determine which cluster they belong. However, in our case, data were categorical. In 1995, Ralambondrainy [23] introduced an approach that enables the use of the k-means algorithm with categorical data. In this approach, nominal attributes are converted into binary attributes—one for each value that the attribute can take—so that they can be considered as numerical attributes by the algorithm.

We performed the clustering analysis with the WEKA (Waikato Environment for Knowledge Analysis; University of Waikato) data mining tool, which provides an implementation of the k-means algorithm (ie, SimpleKMeans) and can handle categorical data [24]. SimpleKMeans can also handle missing values by replacing them with the mean or mode.

For this preliminary clustering analysis, we included only the patients who had at least one recorded diagnosis of a mental disorder (ie, International Classification of Diseases, Ninth Revision [ICD-9] codes 290-319) [25]. Patients were excluded if they had nonpsychiatric diagnoses (ICD-9 codes V01-V91; 2707/46,222, 5.86%) or had never received an ICD-9 diagnosis (5493/46,222, 11.88%).
resulting subset for the clustering analysis included 38,022 patients.

We considered sociodemographic variables, such as biological sex, age at first visit, nationality, marital status, living situation, education, occupational status, birthplace (district), and the catchment area (district) providing care (determined by domicile postal code or by residence postal code when the domicile was missing).

Results

Sociodemographic Characteristics

The sample included 46,222 individuals, whose sociodemographic characteristics are detailed in Table 1.
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Female patients (n=28,109)</th>
<th>Male patients (n=18,113)</th>
<th>All patients (N=46,222)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age at first visit (years), mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;18, n (%)</td>
<td>249 (0.89)</td>
<td>164 (0.91)</td>
<td>413 (0.89)</td>
</tr>
<tr>
<td>18-24, n (%)</td>
<td>2173 (7.73)</td>
<td>1842 (10.17)</td>
<td>4015 (8.69)</td>
</tr>
<tr>
<td>25-34, n (%)</td>
<td>4221 (15.02)</td>
<td>2913 (16.08)</td>
<td>7134 (15.43)</td>
</tr>
<tr>
<td>35-44, n (%)</td>
<td>5141 (18.29)</td>
<td>3308 (18.26)</td>
<td>8449 (18.28)</td>
</tr>
<tr>
<td>45-54, n (%)</td>
<td>4761 (16.94)</td>
<td>3119 (17.22)</td>
<td>7880 (17.05)</td>
</tr>
<tr>
<td>55-64, n (%)</td>
<td>4142 (14.74)</td>
<td>2423 (13.38)</td>
<td>6565 (14.20)</td>
</tr>
<tr>
<td>65-74, n (%)</td>
<td>3780 (13.45)</td>
<td>2180 (12.04)</td>
<td>5960 (12.89)</td>
</tr>
<tr>
<td>≥75, n (%)</td>
<td>3634 (12.93)</td>
<td>2160 (11.93)</td>
<td>5794 (12.54)</td>
</tr>
<tr>
<td>Missing data, n (%)</td>
<td>8 (0.03)</td>
<td>4 (0.02)</td>
<td>12 (0.03)</td>
</tr>
<tr>
<td><strong>Nationality, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>26,486 (94.23)</td>
<td>17,167 (94.78)</td>
<td>43,653 (94.44)</td>
</tr>
<tr>
<td>Foreign</td>
<td>1580 (5.62)</td>
<td>920 (5.08)</td>
<td>2500 (5.41)</td>
</tr>
<tr>
<td>Missing data</td>
<td>43 (0.15)</td>
<td>26 (0.14)</td>
<td>69 (0.15)</td>
</tr>
<tr>
<td><strong>Birthplace (district), n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outside Ferrara province</td>
<td>7525 (26.77)</td>
<td>4825 (26.64)</td>
<td>12,350 (26.72)</td>
</tr>
<tr>
<td>Ferrara</td>
<td>7299 (25.97)</td>
<td>5034 (27.79)</td>
<td>12,333 (26.68)</td>
</tr>
<tr>
<td>Codigoro</td>
<td>3414 (12.15)</td>
<td>2167 (11.96)</td>
<td>5581 (12.07)</td>
</tr>
<tr>
<td>Portomaggiore</td>
<td>2803 (9.97)</td>
<td>1795 (9.91)</td>
<td>4598 (9.95)</td>
</tr>
<tr>
<td>Copparo</td>
<td>2439 (8.68)</td>
<td>1466 (8.09)</td>
<td>3905 (8.45)</td>
</tr>
<tr>
<td>Cento</td>
<td>2262 (8.05)</td>
<td>1463 (8.08)</td>
<td>3725 (8.06)</td>
</tr>
<tr>
<td>Outside Italy</td>
<td>2253 (8.02)</td>
<td>1258 (6.95)</td>
<td>3511 (7.60)</td>
</tr>
<tr>
<td>Missing data</td>
<td>114 (0.41)</td>
<td>105 (0.58)</td>
<td>219 (0.47)</td>
</tr>
<tr>
<td><strong>Marital status, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married or partnered</td>
<td>10,748 (38.24)</td>
<td>6001 (33.13)</td>
<td>16,749 (36.24)</td>
</tr>
<tr>
<td>Single</td>
<td>5551 (19.75)</td>
<td>5721 (31.59)</td>
<td>11,272 (24.39)</td>
</tr>
<tr>
<td>Separated, divorced, or widowed</td>
<td>5587 (19.88)</td>
<td>1744 (9.63)</td>
<td>7331 (15.86)</td>
</tr>
<tr>
<td>Missing data</td>
<td>6223 (22.14)</td>
<td>4647 (25.66)</td>
<td>10,870 (23.52)</td>
</tr>
<tr>
<td><strong>Living situation, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Living with acquired family (partner and children)</td>
<td>12,147 (43.21)</td>
<td>6148 (33.94)</td>
<td>18,295 (39.58)</td>
</tr>
<tr>
<td>Living with parents</td>
<td>3079 (10.95)</td>
<td>3338 (18.43)</td>
<td>6417 (13.88)</td>
</tr>
<tr>
<td>Alone</td>
<td>3194 (11.36)</td>
<td>1768 (9.76)</td>
<td>4962 (10.74)</td>
</tr>
<tr>
<td>Living with other family members</td>
<td>1649 (5.87)</td>
<td>781 (4.31)</td>
<td>2430 (5.26)</td>
</tr>
<tr>
<td>Living with others (eg, roommates)</td>
<td>612 (2.18)</td>
<td>403 (2.22)</td>
<td>1015 (2.20)</td>
</tr>
<tr>
<td>Community housing facilities</td>
<td>184 (0.65)</td>
<td>267 (1.47)</td>
<td>451 (0.98)</td>
</tr>
<tr>
<td>Other</td>
<td>193 (0.69)</td>
<td>220 (1.21)</td>
<td>413 (0.89)</td>
</tr>
<tr>
<td>Safe house</td>
<td>181 (0.64)</td>
<td>199 (1.10)</td>
<td>380 (0.82)</td>
</tr>
<tr>
<td>Retirement home</td>
<td>226 (0.8)</td>
<td>146 (0.81)</td>
<td>372 (0.80)</td>
</tr>
<tr>
<td>Characteristic</td>
<td>Female patients (n=28,109)</td>
<td>Male patients (n=18,113)</td>
<td>All patients (N=46,222)</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td>---------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td><strong>Prison</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing data</td>
<td>6643 (23.63)</td>
<td>4827 (26.65)</td>
<td>11,470 (24.82)</td>
</tr>
<tr>
<td><strong>Education, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illiterate</td>
<td>2941 (10.46)</td>
<td>1540 (8.50)</td>
<td>4481 (9.69)</td>
</tr>
<tr>
<td>Literate (without formal degree)</td>
<td>3044 (10.83)</td>
<td>2312 (12.76)</td>
<td>5356 (11.59)</td>
</tr>
<tr>
<td>Primary school</td>
<td>3674 (13.07)</td>
<td>2167 (11.96)</td>
<td>5841 (12.64)</td>
</tr>
<tr>
<td>Middle school</td>
<td>3247 (11.55)</td>
<td>2644 (14.60)</td>
<td>5891 (12.75)</td>
</tr>
<tr>
<td>High school</td>
<td>3946 (14.04)</td>
<td>2378 (13.13)</td>
<td>6324 (13.68)</td>
</tr>
<tr>
<td>College or university</td>
<td>1294 (4.60)</td>
<td>584 (3.22)</td>
<td>1878 (4.06)</td>
</tr>
<tr>
<td>Missing data</td>
<td>9963 (35.44)</td>
<td>6488 (35.82)</td>
<td>16,451 (35.59)</td>
</tr>
<tr>
<td><strong>Occupational status, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>3873 (13.78)</td>
<td>2735 (15.10)</td>
<td>6608 (14.30)</td>
</tr>
<tr>
<td>Retired</td>
<td>2949 (10.49)</td>
<td>1833 (10.12)</td>
<td>4782 (10.35)</td>
</tr>
<tr>
<td>Unemployed</td>
<td>1531 (5.45)</td>
<td>1291 (7.13)</td>
<td>2822 (6.11)</td>
</tr>
<tr>
<td>Disability</td>
<td>612 (2.18)</td>
<td>670 (3.70)</td>
<td>1282 (2.77)</td>
</tr>
<tr>
<td>Other</td>
<td>760 (2.70)</td>
<td>491 (2.71)</td>
<td>1251 (2.71)</td>
</tr>
<tr>
<td>Homemaker</td>
<td>944 (3.36)</td>
<td>1 (0.01)</td>
<td>945 (2.04)</td>
</tr>
<tr>
<td>Student</td>
<td>512 (1.82)</td>
<td>334 (1.84)</td>
<td>846 (1.83)</td>
</tr>
<tr>
<td>Unknown</td>
<td>16,928 (60.22)</td>
<td>10,758 (59.39)</td>
<td>27,686 (59.90)</td>
</tr>
<tr>
<td><strong>Catchment area (district), n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ferrara</td>
<td>10,964 (39.01)</td>
<td>6784 (37.45)</td>
<td>17,748 (38.40)</td>
</tr>
<tr>
<td>Codigoro</td>
<td>4186 (14.89)</td>
<td>2931 (16.18)</td>
<td>7117 (15.40)</td>
</tr>
<tr>
<td>Portomaggiore</td>
<td>3425 (12.18)</td>
<td>2193 (12.11)</td>
<td>5618 (12.15)</td>
</tr>
<tr>
<td>Cento</td>
<td>3380 (12.02)</td>
<td>2097 (11.58)</td>
<td>5477 (11.85)</td>
</tr>
<tr>
<td>Copparo</td>
<td>2754 (9.80)</td>
<td>1671 (9.23)</td>
<td>4425 (9.57)</td>
</tr>
<tr>
<td>Unknown</td>
<td>3400 (12.10)</td>
<td>2437 (13.45)</td>
<td>5837 (12.63)</td>
</tr>
</tbody>
</table>

*FEPSY: Ferrara-Psychiatry.

**Extract, Transform, Load Process**

**Built-In Tables**

Built-in tables from the EFESO relational database, which are detailed in Textbox 2, were included in the FEPSY database.
Textbox 2. Built-in tables included in the FEPSY (Ferrara-Psychiatry) database.

1. **Table Patients** included individual personal data, such as name, place and date of birth, biological sex (male or female), home address, living condition, education, marital status, occupation, and other sociodemographic characteristics.

2. **Table Medical Records** contained 1 or more medical records for each patient, with information such as the date of admission, date and type of discharge, primary diagnosis of a mental disorder, and facility providing care.

3. **Table Diagnoses** included 1 or more diagnoses that were assigned to each individual. Diagnoses were classified according to the *International Classification of Diseases, Ninth Revision (ICD-9)* categorical system; therefore, every diagnosis included the associated ICD-9 code, description, group, and chapter [24]. Diagnoses recorded before the introduction of the ICD-9 were recorded in SIPER (Sistema Informativo Psichiatrico dell'Emilia-Romagna), using standardized conversion criteria [24].

4. **Table Products** referred to the different types of medical services, such as consultations or hospitalizations. A product had a start date and end date, and it may have contained 1 or more medical services.

5. **Table Medical Services** stored every service that each individual had received or undergone, such as consultations, first visits, the administration of pharmacological treatment, social skill–oriented activities, structured diagnostic assessments, and mandatory medical treatments, as well as the facility providing care.

6. **Table Medication Prescription** and **Medication Administration** referred to the prescription and administration of pharmacological treatment, type of medication and dosage, start and stop dates, and responsible facility.

7. **Table Psychometric Tests** included every test administered to each patient and the test types, dates, questions, and scores.

8. **Table Projects** listed the treatment plans for each patient. There were individual and group projects, and within a project, there could have been 1 or more products and medical services.

9. **Table Facilities** contained all of the facilities of the Health Trust of Ferrara for Mental Health in Adults, such as hospitals, day care centers, and clinics, along with their types and locations.

**Extract**

Data were extracted from EFESO by using an automated procedure that executed an SQL select query. This query selected all relevant fields of a table and other useful information from linked support tables, such as the descriptions of the codes. The result of the query was stored in a Pandas DataFrame (The Pandas Development Team) [26,27], which can be easily manipulated in the next phase.

**Transform**

**Record Exclusion**

Data imported before 1991 were excluded (as detailed in the Data Preparation section). Some fields and records were removed [22] to ensure data consistency, because there were duplicate or erroneous records (Table S1 in Multimedia Appendix 1). These were (1) fields containing unreliable information, (2) fields that were present but not in use (their values were always null), and (3) all records marked as “deleted” (i.e., wrong records that were not to be used) and all records in other tables referencing the “deleted” ones. We decided to remove 36 individuals that had unique fiscal codes but duplicate patient IDs—corresponding to 0.15% (72/48,001) of the records in the total data set—since it was not possible to determine which of the two entries was the correct one. When a patient was first included in the database, a unique identifier—the patient ID—was assigned. The combination of the tax code and the patient ID allowed for the unique identification of a patient in the database. We also excluded patients for whom a record was opened earlier than the birth date (16/48,001, 0.03%), patients with no medical records (603/48,001, 1.26%), and patients marked as “deleted” (77/48,001, 0.16%). Overall, 1.6% (768/48,001) of the total records, which related to 732 patients, were removed from the source table.

**Anonymization**

Anonymization was necessary in order to use the extracted data for research projects and was performed on tables Patients and Medical Records. First, the extracted records were shuffled. Afterward, the original patient and medical record IDs were replaced with a universally unique identifier (UUID), which is a 128-bit string that is usually represented as a sequence of 32 hexadecimal digits [28]. These new random, unique identifiers were generated with the `uuid4` function of the Python `uuid` package (Python Software Foundation) [29] and used as the new primary key. In order to maintain the referential integrity (i.e., the primary key of one table is a foreign key in another table, meaning that they are related), the old IDs were replaced with the new ones within every table in which they appeared. Furthermore, all PHI were excluded from table Patients; these data included first names, last names, days and months of birth, tax codes, home addresses, phone numbers, and note fields that could potentially include personal data (e.g., relatives’ names). For the same reason, text note fields were also excluded from other tables, when present.

**Field Transformation**

Transformation was necessary for date fields. EFESO stored dates in “datetime” format, that is, “dd/mm/yyyy hh:mm.” However, previous EHRs stored only the date, without the hour information. Furthermore, even when specified, the hour information is not always reliable. For this reason, the date fields were split into 2 fields—one for the date and one for the time.

**Missing Values**

Missing values were assessed to avoid the introduction of bias. In specific analyses, the level and pattern of missingness will be assessed for each variable included and dealt with accordingly.
Load
Data extracted from EFESO were loaded in the FEPSY database—the newly created MySQL (Oracle Corporation) relational database—by using the same automated procedure that was used to extract them. For each built-in table, an insert query, which took the values from the same DataFrame of the select query, was executed.

Analysis of the Extracted Data
The data included in the final composite FEPSY data set were those collected from 1991 to February 2021. Since 1991, each year, a median of 1404 (IQR 1117.5-1757.7) individuals had newly accessed care, and a median of 7300 (IQR 6109.5-9397.5) individuals were actively receiving care, as represented in Figure 1. Figure 2 shows the number of patients treated per year in total and by sex. The sudden decrease observed in 2009 was due to an automated closing procedure that was introduced in 2008 and retained from then on. When migrating from IPPOCRATE to EFESO, all medical records, products, and diagnoses that had not been updated in 365 days were assumed to be closed or terminated, and the missing closing date was replaced with the date of the migration to EFESO, that is, August 26, 2008.

Figure 1. New admissions per year in total and by sex (upper panel). Timeline of the electronic health records adopted by the health care agency in the Ferrara province (lower panel). CURE: Cartella Unificata Regionale Elettronica; GESAP: Gestione attività Psichiatrica; SIPER: Sistema Informativo Psichiatrico dell’Emilia-Romagna.
As described in Table 2, the most frequent diagnoses at first admission were depression and anxiety disorder. During the 30-year time span, more than half (32,230/46,222, 69.73%) of the patients had only 1 chart open, and only 5184 patients had at least one psychiatric hospitalization.
Main clinical characteristics of the sample (N=46,222; years 1991-2021).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Female patients (n=28,109, 68.81%)</th>
<th>Male patients (n=18,113, 39.19%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at first visit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years, mean (SD)</td>
<td>50.46 (18.82)</td>
<td>48.72 (19.02)</td>
</tr>
<tr>
<td>Years, median (range)</td>
<td>49.0 (0-109)</td>
<td>47.0 (2-98)</td>
</tr>
<tr>
<td>Number of charts/patient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value, mean (SD)</td>
<td>1.62 (1.55)</td>
<td>1.64 (2.38)</td>
</tr>
<tr>
<td>Value, median (range)</td>
<td>1.0 (1-63)</td>
<td>1.0 (1-132)</td>
</tr>
<tr>
<td>Patients with at least one hospitalization, n</td>
<td>2680</td>
<td>2504</td>
</tr>
<tr>
<td>Number of hospitalizations/patient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value, mean (SD)</td>
<td>0.33 (2.36)</td>
<td>0.46 (2.35)</td>
</tr>
<tr>
<td>Value, median (range)</td>
<td>0 (0-143)</td>
<td>0 (0-102)</td>
</tr>
<tr>
<td>Patients with at least one compulsory admission, n (%)</td>
<td>415 (1.48)</td>
<td>485 (2.68)</td>
</tr>
<tr>
<td>Duration of hospitalization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Days, mean (SD)</td>
<td>5.08 (40.89)</td>
<td>7.53 (62.39)</td>
</tr>
<tr>
<td>Days, median (range)</td>
<td>0 (0-2661)</td>
<td>0 (0-4090)</td>
</tr>
<tr>
<td>First recorded mental disorder diagnosis*, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety disorders</td>
<td>6884 (24.49)</td>
<td>3725 (20.57)</td>
</tr>
<tr>
<td>Dementia and other organic disorders</td>
<td>2092 (7.44)</td>
<td>1601 (8.84)</td>
</tr>
<tr>
<td>Depression</td>
<td>7648 (27.21)</td>
<td>3335 (18.41)</td>
</tr>
<tr>
<td>Drug and substance use or abuse</td>
<td>415 (1.48)</td>
<td>861 (4.75)</td>
</tr>
<tr>
<td>Eating disorders</td>
<td>241 (0.86)</td>
<td>18 (0.10)</td>
</tr>
<tr>
<td>Intellectual disability</td>
<td>468 (1.66)</td>
<td>636 (3.51)</td>
</tr>
<tr>
<td>Mania and bipolar disorders</td>
<td>713 (2.54)</td>
<td>460 (2.54)</td>
</tr>
<tr>
<td>Personality disorders</td>
<td>1287 (4.58)</td>
<td>1186 (6.55)</td>
</tr>
<tr>
<td>Schizophrenia and other nonorganic psychoses</td>
<td>1468 (5.22)</td>
<td>1515 (8.36)</td>
</tr>
<tr>
<td>Other mental disorders</td>
<td>2339 (8.32)</td>
<td>1130 (6.24)</td>
</tr>
<tr>
<td>No formal mental disorder diagnosis</td>
<td>4554 (16.20)</td>
<td>3646 (20.13)</td>
</tr>
</tbody>
</table>

*Mental disorder diagnoses: International Classification of Diseases, Ninth Revision codes 290.xx-319.xx.

**Clustering Results**

This analysis, which was carried out on the subset of 38,022 individuals who had at least one mental disorder diagnosis, identified 2 distinct clusters (Figure 3). One is represented by single male patients who were born in Ferrara, those who were living with parents, and those whose age at first visit was between 25 and 34 years; the other is represented by married female patients who were living with their own acquired families, those who were born outside the province of Ferrara, and those whose age at first visit was 35 to 44 years. The following sociodemographic features were similar in the two clusters: Italian nationality, individuals with a high school degree, employed individuals, and individuals who were receiving treatment in the Ferrara catchment area.
Discussion

Principal Findings

This study describes the process of adapting one of the longest running EHRs of public mental health care for research purposes. The FEPSY data set covers a catchment area with 342,061 inhabitants (as of 2020) and includes a total of 46,222 unique individuals who had access to mental health services over a span of 30 years (1991-2021). The FEPSY database is suitable for descriptive, predictive, and inferential analyses via conventional analysis and AI techniques, as demonstrated by the preliminary findings of the clustering analysis. To our knowledge, our database is the first of its kind in Italy. In Europe, longitudinal and prospective registries have long been in use. For example, large data sets were extracted from the Danish National Patient Registry [30] and the Danish Psychiatric Central Research Register [31]. The first data set contained data on 8,085,603 patients, which were collected from 1977 to 2012. The second data set included data on a total of 747,176 patients, which were collected from 1970 to 2010. In both cases, the register contained dates of the onset and end of any treatment, diagnoses, types of referrals, and places of treatment, thereby allowing for the possibility to perform health registry-based research [32], considering a total population of approximately 27 million.

The main finding of this study is that data that were not originally conceived for research were successfully extracted from EHR software and loaded into a new anonymized database.
This step is of foremost importance, as the data originated from an information system that was changed and updated multiple times and was not designed to allow for exploratory investigations in a structured manner. Thus, this new data set may represent the ideal setting to build, test, and refine an analytical methodology for extracting data and preparing these data for research purposes. This methodology could also be applied to other clinical data sets, such as data sets from other medical disciplines (eg, oncology), with characteristics that are similar to those of the FEPSY data set [33]. Additionally, it will be of foremost importance to validate the methodological approach and findings of upcoming research originating from the FEPSY data set by proactively seeking collaboration with other research groups, in order to enable the replication of findings from the territory of Ferrara and the use of the FEPSY data set to replicate findings from research involving other registries.

This work also allows for collaborations in terms of learning health networks, which use comparable data that originate from EHRs to support clinical decisions, improve the delivery of efficient and effective medical care, and help with the integration of research in health care [34-36].

Our results will also pave the way for an in-depth study on the use of health care resources; the results will be used to develop a system that is capable of planning the use of such resources. Such a system would optimize the use of health care resources while maintaining or possibly improving the quality of treatment. For example, in Italy, Donisi et al [37] predicted the cost of community mental health care by using clinical and sociodemographic information originating from the Psychiatric Case Register in the Verona Health District. This allowed for the linking of social deprivation to psychiatric service utilization [38] and shed a light on possible contributors to social isolation in an already vulnerable population. Our clustering analysis, which was conducted on the FEPSY data set to test its feasibility and robustness, identified 2 clusters; women appeared to access mental health services later in life and were typically married, in comparison to men. These findings were consistent with the literature [39-41] and supported the ability of the extracted data to detect known patterns, even though the results should be interpreted with caution, given the large amount of missing sociodemographic data. Clustering analyses can be useful for building prediction models and planning a department’s resource allocation, as they provide relevant information on patients at presentation and on illness trajectory [42].

The process described in this paper faced 5 major challenges that we mitigated, as follows. First, the source data included several built-in structural informatic elements (so-called tables) that had to be screened and deleted in order to get to the core data. Second, the anonymization step was of absolute importance, and in order to both comply with privacy constraints and be able to preserve the integrity of the data, the study team decided to keep only subelements of certain data items (eg, for the birth date, only the year was kept, and for residence, only the postal code was kept). Third, in order to establish which records were correct, an iterative comparison of the FEPSY database and the local and regional database was performed by a third party who had access to PHI. Fourth, records that were deemed unreliable were excluded (eg, clinical procedures referring to nonexistent medical records). Furthermore, the data extracted from EFESO originated from different software and thus possibly generated some errors. In the FEPSY data set, these errors seem to be limited to records, and the proportion of records with errors was very low (49,854/3,861,432, 1.29%). In the end, we decided to exclude data that were collected before 1991 and duplicate patient records. Fifth, missing data challenges were also addressed, especially in the clustering process. For this purpose, the WEKA data mining tool was used.

**Strengths of This Study**

The quality and completeness of the collected and cleaned data, as well as the large number of records stored in the FEPSY database, resulted in the definition of a data set that is particularly suitable for automatic analysis and has appealing characteristics for research, such as a long period of data availability, great diversity in the sociodemographic factors of the patients represented, and a history of treatments and drugs administered. This could possibly represent a strong foundation for many different studies of mental illness and resource use, favoring comparisons between Italy and other countries regarding the delivery and quality of community and hospital psychiatric care [43-45]. Furthermore, the newly created database does not include sensitive information, even though this information can be retrieved by using an external supporting table created ad hoc, which could link the FEPSY data set with other data sets (eg, hospital data and tumor registries) for future research.

The novelty of this project is represented by its interdisciplinary nature (psychiatry, public health, epidemiology, sociology, mathematics, computer science, and AI), the potential versatility of the methods that can be used with the FEPSY database, and the versatility of the systems that could be created via analyses involving the FEPSY database. To our knowledge, this study is the first attempt to retrospectively build a single data set that includes more than 30 years’ worth of data on mental health services in a specific area. Such a data set would also allow for longitudinal analyses, such as those that have already been performed with the Nordic registry (a prospective registry) and, more recently, the South London and Maudsley National Health Service [46] and the Camden & Islington Research Database [47].

We believe that historical data can add value to subsequent analyses, because they allow researchers to understand how mental health services have evolved over the past decades and the extent to which phenotypical presentations of different diseases have changed over time. In light of these considerations, factors that should be taken into account are (1) potential cohort and time effects, such as historical events (eg, the Great Recession in 2008); (2) changes in legal and medical approaches to mental health; and (3) changes in the classification of mental disorders [48].

**Limitations**

Our findings must be interpreted in the light of some limitations. First, the sample size is limited by the geographical catchment. A larger catchment or a more densely populated region would...
probably have a larger volume of treated individuals and thus have more data, which would facilitate machine learning analyses. However, we believe that even if the sample is limited by the geographical catchment, the diverse socioeconomic distribution in Ferrara is a strength that mitigates this limitation, providing insight into the possible moderator or mediator roles of socioeconomic variables that are considered social determinants of mental health. Second, another potential limitation is the missing data for some sociodemographic attributes, which may reduce the statistical power of a study [49] or affect the accuracy of machine learning algorithms [50]. In order to overcome this issue, missing values can be handled with multiple imputation methods or replaced with the mean or the mode (ie, for quantitative or qualitative data, respectively). Moreover, sociodemographic information can be drawn from external and publicly available sources, such as the Italian National Institute of Statistics [51], which includes the census of the population as well as social, economic, and environmental surveys and analyses. Lastly, there is the risk of introducing bias while building prediction models, especially when using supervised machine learning techniques, due to small sample sizes and the poor handling of missing data and overfitting [52]. With regard to the sample size, our sample appears to be sufficiently large for risk prediction analyses. Overfitting can be addressed with penalized models [53].

Future Directions
This work sets a starting point for future investigations, which can be described as follows: (1) identifying patients who have a higher severity index or chronicity level and those who require a greater use of health resources; (2) identifying and validating, by means of machine learning models, demographic, clinical, and social predictors of clinically relevant outcomes that are useful for an ad hoc programming of resources (eg, sex, gender, or social deprivation [38,39,54]); (3) further optimizing and tailoring the analysis methods, so that they can also be applied to other data sets (eg, the local mental health registries for child and adolescent neuropsychiatry and for drug addiction services); (4) interacting with international learning health networks [55-57]; and (5) linking the FEPSY data set with external data sources, such as census data, tumor registries [58], death registries, and criminal justice data [59]. As a result of the increasing digitalization of medical records, it was possible to gather years of mental health history for every patient. This will enable for the conduct of symbolic and subsymbolic analyses on time series via automatic methodologies. Classical supervised and unsupervised machine learning and deep learning techniques will be evaluated. In order to explore the relationship between sociodemographic characteristics and specific diagnostic questions (eg, the incidence and prevalence of psychosis), a supervised framework will be deployed, in which binary labels (eg, “psychosis” or “no-psychosis”) or multiple classification labels (eg, “ICD-9 diagnosis”) will be associated with the patient. The main problem to overcome will be the imbalance of the data set, that is, when there is an unequal distribution of classes in the data set. In such instances, a standard machine learning technique, such as a support vector machine or random forest [60,61], will be applied. Moreover, each patient could potentially be considered as a distinct time series by including the temporal dimension of the treatment and by applying recurrent neural networks [62,63]. By doing so, the prediction of the new onset of a disease and the subsequent use of health resources will be the focus, in order to plan and optimize health care resources.

Conclusions
The process described in this study resulted in the building of a data set that included the information of 46,222 individuals who had access to psychiatric services in the Ferrara province over the course of almost 30 years. The preliminary findings from the clustering analysis confirmed the quality of the newly established database. The process we implemented proved to be a solid method that can be replicated with similar data sets, even if they were not originally compiled for research purposes.

Acknowledgments
This work was supported by the “Fondo per l’Incentivazione alla Ricerca (FIR),” granted by the University of Ferrara in 2021 to MF, with the project titled “Intelligenza Artificiale per Predizione Diagnostica, di Carico Assistenziale, e di Esito: 40 Anni di accessi Presso i Centri di Salute Mentale della Provincia di Ferrara (AI4MentalHealth).”

Authors’ Contributions
MF designed the study, wrote the first draft of the manuscript, and supervised the project. EG processed the experimental data, performed the analysis, and designed the figures. MBM, RZ, MA, GF, ID, FF, and LG were involved in planning and interpreting the results. CS, LB, and JL aided in interpreting the results. All authors discussed the results, commented on the manuscript, and approved the final version.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Characteristics of the records within the tables of the Ferrara-Psychiatry (FEPSY) database.
[DOCX File, 17 KB - medinform_v11i1e45523_app1.docx ]
References


https://medinform.jmir.org/2023/1/e45523

XSL-FO RenderX


56. STEP Learning Collaborative. URL: www.etectralypsychosisnetwork.org/ [accessed 2023-07-21]


Abbreviations

AI: artificial intelligence
EHR: electronic health record
FEPSY: Ferrara-Psychiatry
ICD-9: International Classification of Diseases, Ninth Revision
PHI: protected health information
SIPER: Sistema Informativo Psichiatrico dell'Emilia-Romagna
UUID: universally unique identifier
WEKA: Waikato Environment for Knowledge Analysis

https://medinform.jmir.org/2023/1/e45523

JMIR Med Inform 2023 | vol. 11 | e45523 | p.635
(page number not for citation purposes)
A Standardized Clinical Data Harmonization Pipeline for Scalable AI Application Deployment (FHIR-DHP): Validation and Usability Study

Elena Williams¹, MSc; Manuel Kienast¹, MSc; Evelyn Medawar¹, MSc; Janis Reinelt¹, MD; Alberto Merola¹, PhD; Sophie Anne Ines Klopfenstein², MD; Anne Rike Flint², MSc; Patrick Heeren², BSc; Akira-Sebastian Poncette², MD, PD; Felix Balzer², Prof Dr; Julian Beimes³, MSc; Paul von Bünau³, PhD; Jonas Chromik³, MSc; Bert Arnrich³, Prof Dr; Nico Scherf⁴, PhD; Sebastian Niehaus¹, MSc

¹AICURA Medical GmbH, Berlin, Germany
²Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Berlin, Germany
³idalab GmbH, Berlin, Germany
⁴Digital Health – Connected Healthcare, Hasso Plattner Institute, University of Potsdam, Potsdam, Germany
⁵Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Corresponding Author:
Evelyn Medawar, MSc
AICURA Medical GmbH
Bessemerstr 22
Berlin, 12103
Germany
Phone: 49 173 9449677
Email: evelyn.medawar@aicura-medical.com

Abstract

Background: Increasing digitalization in the medical domain gives rise to large amounts of health care data, which has the potential to expand clinical knowledge and transform patient care if leveraged through artificial intelligence (AI). Yet, big data and AI oftentimes cannot unlock their full potential at scale, owing to nonstandardized data formats, lack of technical and semantic data interoperability, and limited cooperation between stakeholders in the health care system. Despite the existence of standardized data formats for the medical domain, such as Fast Healthcare Interoperability Resources (FHIR), their prevalence and usability for AI remain limited.

Objective: In this paper, we developed a data harmonization pipeline (DHP) for clinical data sets relying on the common FHIR data standard.

Methods: We validated the performance and usability of our FHIR-DHP with data from the Medical Information Mart for Intensive Care IV database.

Results: We present the FHIR-DHP workflow in respect of the transformation of “raw” hospital records into a harmonized, AI-friendly data representation. The pipeline consists of the following 5 key preprocessing steps: querying of data from hospital database, FHIR mapping, syntactic validation, transfer of harmonized data into the patient-model database, and export of data in an AI-friendly format for further medical applications. A detailed example of FHIR-DHP execution was presented for clinical diagnoses records.

Conclusions: Our approach enables the scalable and needs-driven data modeling of large and heterogenous clinical data sets. The FHIR-DHP is a pivotal step toward increasing cooperation, interoperability, and quality of patient care in the clinical routine and for medical research.

(JMIR Med Inform 2023;11:e43847) doi:10.2196/43847
KEYWORDS
data interoperability; fast healthcare interoperability resources; FHIR; data standardization pipeline; medical information mart for intensive care; MIMIC IV; artificial intelligence; AI application; AI; deployment; data; usability; care unit; diagnosis; cooperation; patient care; care; medical research

Introduction

The increasing digitalization of health care creates vast amounts of clinical data that are collected and stored in an Electronic Health Record (EHR). Patient information from all medical domains is captured in diverse sets of data recorded in stand-alone systems. With the prevalent use of EHRs in health care organizations, there is abundant opportunity for the additional application of EHR data in clinical and translational research. For instance, such data can be used to develop artificial intelligence (AI) algorithms, which have the potential to transform patient care and medical research. Resource-intensive and inefficient clinical workflows could be optimized by the analysis of historical data with AI applications [1,2]. In particular, the time-consuming and financially costly process of identifying and enrolling the right patients into a clinical trial manually can be reduced significantly by automation [3,4]. However, the exchange of medical data remains limited due to the lack of data interoperability between health care providers, owing to outdated IT infrastructure, inconsistencies in data formats, poor data quality, inadequate data exchange solutions, and data silos [5,6]. To achieve data interoperability, the following steps must be incorporated: (1) integration of isolated data silos, (2) safe exchange of data, and (3) effective use of the available data [7]. Each of these operations includes database schema matching [8] and schema mapping [9], which allow translation of the relationships between the source database and the target data standard.

Employing a harmonized data format will facilitate the exchange of medical data, enabling wide-ranging data-driven collaborations within the private and public health care sectors. Data interoperability requires EHR data to be structured in a common format and in standardized terminologies. Standardization is often performed by adopting the Health Level 7 Fast Healthcare Interoperability Resources (FHIR) model [10], which is supported by numerous health care institutions and vendors of clinical information systems [11]. FHIR is an international industry standard that integrates diverse sets of data in well-defined exchangeable segments of information, which are known as FHIR resources. Therefore, FHIR facilitates interoperability between health care organizations and allows third-party developers to provide medical applications that can be easily integrated into the existing systems. FHIR enables the harmonization of data and thus allows standardized data processing as well as the rollout of AI applications across different clinics and hospitals regardless of which information system they use. Consequently, FHIR forms an important component for the scalable development and deployment of AI in clinics and hospitals.

However, to apply AI, the input data need to be adapted to the AI algorithms. The conventional AI frameworks such as Tensorflow [12] and Pytorch [13] require data to take a tensor form, which is a vector or matrix of n-dimensions that represents various types of data (eg, tabular, time series, image, and text). Since the FHIR format has a multilayered nested structure, a use case–specific data preprocessing is needed. For instance, depending on the AI application and the chosen data source, a custom data preprocessing pipeline should be designed leading to diminished AI scalability. Prior research addressed this problem in different forms but focused on individual applications, thereby constraining the purpose of FHIR to be applicable regardless of the use case [11]. There have been a few attempts to flatten the hierarchical FHIR structure and transform it into NDJSON-based data format [14] or tabular format saved in CSV files [15]. Such formats are more AI-friendly as they represent the data in a more accessible and standardized form for an application of common AI frameworks. Nonetheless, the NDJSON-based FHIR data transformation approach [14] does not provide data selection criteria and filtering capabilities [16]. The approach presented in [15] requires expert knowledge of FHIRPath query language. Moreover, FHIR-based data preprocessing pipelines have been implemented in different contexts, for instance, as electronic data capture [17], as a natural language processing tool [12], and as a standardization protocol based on the Resource Description Framework [6]. Despite the immense benefit they offer regarding processing EHR data, existing approaches are limited to specific use cases or require considerable data preparation to perform standardization. Furthermore, their final output is not easily accessible by common data preprocessing tools and thus hinders the application of AI.

In this paper, we address the challenge of data interoperability in the health care sector by proposing an FHIR data harmonization pipeline (DHP) that provides EHR data in an AI-friendly format. The newly developed FHIR-DHP represents a data workflow solution that includes the aforementioned operations, such as data exchange, mapping, and export. Data privacy is a delicate topic in health care and is of great ethical concern [18]. Given the degree of automation, FHIR-DHP should allow the preprocessing of unseen data in an isolated hospital environment, which makes harmonization privacy preserving.

Methods

Ethical Considerations

The authors did not seek an ethics review board assessment due to the methodology of the study, which included open datasets and data preprocessing pipelines only.

FHIR-DHP Architecture Development

In our work, we propose a generic solution to harmonize hospital EHR data. The FHIR-DHP was designed based on the extract-transform-load framework [19], in which the data are pulled out (ie, queried) from diverse sources, processed into the desired format, and loaded into a data warehouse, namely the
"patient-model" database (DB). As the hospital database contains highly sensitive patient data, it is located behind the hospital’s security infrastructure and is completely isolated from outside access. Therefore, an edge-computation solution was designed, bringing the FHIR-DHP into the hospital’s own infrastructure. The edge-computation solution represents a set of frameworks that perform data querying, preprocessing, storage, and export. In this setting, direct access to the sensitive data is not required to run the standardization pipeline. The queries to the data are defined beforehand based on the database documentation.

To bring the data into a harmonized form, we used an FHIR data model, which is applied by mapping the relationships between the source database and the desired data standard. The FHIR standard is straightforward to implement because it provides a choice of JSON, XML, or resource description format for data representation. The mapping pipeline was developed in the Python programming language to translate queried hospital data into matching FHIR concepts and save the resulting resources in JSON format. The semantics of features from the source database and FHIR concepts are examined using available database and FHIR documentation. The conversion to FHIR was designed to only support a core release 4 standard of the FHIR format to allow generic data preprocessing.

To prevent errors in the remote data standardization scenario, the syntactic validation of FHIR resources is necessary. For instance, the conversion of data types can sometimes lead to erroneous values, especially with date features. Automatic syntactic validation allows the logging of occurred errors and the improvement of harmonization pipeline when working with unseen data. When syntactic validation is completed, FHIR resources should be transferred to the data warehouse to allow the fast and easy retrieval of standardized data for AI applications.

In the final stage of data export, we designed the output that provides the benefits of the original FHIR format with a high level of clinical detail that is also easily accessible for computational tools. We wanted to restructure the data representation in a way that supports effortless data selection and filtering capabilities and would not require a knowledge of FHIRPath query language. Consequently, this output format would enable the smooth conversion of data into a "tensor" format required by conventional AI frameworks.

**FHIR-DHP Validation**

To demonstrate and evaluate how the FHIR-DHP works, we used the openly available Medical Information Mart for Intensive Care IV (MIMIC IV) database [20]. MIMIC IV includes patient data from the intensive care units at a tertiary academic medical center in Boston, MA, United States. We selected a wide range of tables from MIMIC IV, which cover most of the events occurring during the hospital stay as well as core patient details, information about admissions, and hospital transfers (further referred to as core tables). The event tables include laboratory results, diagnoses, prescriptions, and other details, as shown in Table 1. In addition, MIMIC IV includes the so-called reference tables containing matching dictionaries with medical terms that are used in the hospital records.

<table>
<thead>
<tr>
<th>Selected core and event MIMIC IV tables</th>
<th>Selected MIMIC IV reference tables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>_a</td>
</tr>
<tr>
<td>Admissions</td>
<td>_</td>
</tr>
<tr>
<td>Transfers</td>
<td>_</td>
</tr>
<tr>
<td>Chartevents</td>
<td>d_items</td>
</tr>
<tr>
<td>Labevents</td>
<td>d_labitems</td>
</tr>
<tr>
<td>Procedureevents</td>
<td>d_items</td>
</tr>
<tr>
<td>Prescriptions</td>
<td>_</td>
</tr>
<tr>
<td>Inputevents</td>
<td>d_items</td>
</tr>
<tr>
<td>Microbiologyevents</td>
<td>_</td>
</tr>
<tr>
<td>Outputevents</td>
<td>d_items</td>
</tr>
<tr>
<td>Procedures_icd</td>
<td>d_icd_procedures</td>
</tr>
<tr>
<td>Diagnoses_icd</td>
<td>d_icd_diagnoses</td>
</tr>
</tbody>
</table>

|a Not available. |

The selected tables were mapped to FHIR standard. Automatic semantic validation is unfeasible, so 2 of the authors manually validated the mapping semantics independently of each other. There are many tools that perform automatic syntactic validation, such as the Python-based package “fhir.resources” used herein [21]. To evaluate the exporting of data from the patient-model DB, we retrieved the diagnosis records.

**Results**

**FHIR-DHP Architecture**

The approach presented here represents a scalable protocol for harmonizing hospital EHR data sets based on 5 stages from data query to data export in a standardized format.
Querying Data From the Hospital Database

To connect the FHIR-DHP pipeline to the hospital DB, a communication server is employed. This server runs all necessary queries to retrieve the patient data. The query execution can be run at regular intervals as well as in batches of patients, so as not to overload the data pipeline. Furthermore, the queries prestructure the data according to their semantic relations before proceeding to data mapping.

Mapping Data to FHIR

FHIR allows describing data formats and elements that are recorded as “resources” and an application programming interface for exchanging EHRs. To perform the mappings, semantics of features from the source database and FHIR concepts are explored as well as the relationships between the data tables. Consequently, the mappings between the database tables and FHIR resources are defined. Features where a matching FHIR concept is not found are excluded. The resulting FHIR resources are then saved in JSON format.

Syntactic Validation of FHIR Mappings

During validation, mapped data are ensured to have the correct data types as well as the syntactic format where the hierarchy is maintained, and entries follow FHIR standard specifications. All mappings are validated first during the development stage to identify structural errors and data type inconsistencies. A validation algorithm is incorporated into the pipeline to confirm the correctness of the transformed data in the remote data standardization scenario.

Transferring FHIR Resources to Patient-Model DB

The DB of choice for the patient model is Postgres [22], which is an open-source relational DB management system featuring SQL compliance and storage of JSON documents. The database for the FHIR resources is used to harmonize the locally available data only once to allow the further application of various medical AI-based solutions. The data are stored according to FHIR resource type where each resource is saved in a separate JSON structure.

Exporting Data Into Custom JSON format

To export the data from the patient-model DB, the selection is performed by outlining the tables and features of interest in a configuration file, which is then used to determine which harmonized data should be queried. FHIRPath queries were written to retrieve all elements from FHIR resources adhering to specific formatting rules in respect of the predefined key-value structure and to place the extracted elements into the custom JSON file. Such transformation flattens the hierarchical structure of FHIR resources and makes the data more accessible for common data preprocessing tools. The final flattened output does not require expert knowledge of FHIRPath query language and supports effortless data selection and filtering. The resulting file also allows the uncomplicated conversion of data into a “tensor” format required by conventional AI frameworks and fast data selection based on the following 4 keys: feature_name, table_name, value, and metadata.

In Figure 1, we demonstrate how the FHIR-DHP recodes nested FHIR syntax to more accessible features in an AI-friendly format. Example FHIR concepts from an observation resource are given in Figure 1a, where the code’s entity “text” defines the record or measurement label. The entity “text” is often duplicated in the item “display.” However, depending on the coding system, this “display” item can change, whereas “text” always stays the same and is therefore used as a feature name. The information from the FHIR resource is grouped into the 4 concept keys of feature name (eg, “Blood pressure”), value (eg, “114”), table name (eg, “observation”), and metadata (Figure 1b). For a given FHIR resource type, the metadata may include concepts such as dates, references, coding system details, and resource ID, among other things. As an output, feature names together with a corresponding value and available metadata are provided in a custom JSON structure (Figure 1c). The defined format allows uncomplicated data selection and aggregation based on resource type (eg, “table_name”), feature name, and value. Additional information in a standardized format can be easily accessed from the metadata key and allows further data manipulation.
Figure 1. Conceptual overview for an exemplary Fast Healthcare Interoperability Resources (FHIR) structure and hospital record, which are transformed from FHIR standard to an artificial intelligence (AI)–friendly format.

FHIR-DHP Validation

The MIMIC IV data were queried accordingly to the defined FHIR mappings. The core and event MIMIC IV tables were merged with reference tables to contain a complete description of the hospital records. As a result, the data were grouped and restructured into the information blocks required in FHIR standard. Manual independent validation of the mapping semantics resulted in slight discrepancies, which were subsequently resolved to adhere closely to the FHIR standard. The automatic syntactic validation allowed the prompt verification of standardization operations.

Table 2 shows to which FHIR resources the MIMIC IV tables were mapped. The largest proportion of tables (4 out of 12 tables) were mapped to the Observation FHIR resource type, which included lab, microbiology, output, and charted events.
collected throughout the patient’s stay. The information on admissions and transfers was translated into the Encounter FHIR resource (2 out of 12 tables). Procedure events and International Classification of Diseases codes (2 out of 12 tables) were stored in the Procedure FHIR resource. Given that the prescriptions table contains medication requests (1 out of 12 tables) and the input events table holds records of medication administration (1 out of 12 tables), these tables were mapped to the corresponding FHIR resource types. Finally, the Condition FHIR resource was used to map the table with the patients’ diagnosis details (1 out of 12 tables).

In Table 3, we demonstrate how the mapping of the MIMIC IV “diagnoses_icd” table to Condition FHIR resource was conducted. Multiple columns of the “diagnoses_icd” table such as “icd_code”, “icd_version,” and “long_title” were mapped to the FHIR “condition.code” concept, which has a nested structure and provides keys to store the exact International Classification of Diseases code, the version of the coding system, and the code title. The full diagnosis title was mapped both to the “display” and “text” entities.

Figure 2 shows an example of how queried diagnoses records are harmonized to an AI-friendly format. The standardization follows the FHIR-DHP stages described above. At first, the raw data from tables “diagnoses_icd” and “d_icd_diagnoses” are queried (Figure 2a) and merged accordingly to the defined FHIR mappings. Then, the features are renamed as defined in Table 3 for the FHIR condition resource, and the required entities such as “resourceType” and “id” are created (Figure 2b). Finally, the values are placed into a nested FHIR structure (Figure 2c), and subsequently, the data are transformed into a JSON format (Figure 2d), which can be automatically validated (Figure 2e) and saved in the patient-model DB. When the resource is not approved in terms of its syntactic quality (eg, data type, nested structure, or cardinality), an error is raised, which prevents the further saving of this resource in the patient-model DB (Figure 2e). Otherwise, the resource is transferred into a storage (Figure 2f), and the requested data are exported in a custom AI-friendly JSON format (Figure 2g).

We provide an example of a further 2-step transformation of harmonized diagnosis data to a “tensor” format in Multimedia Appendix 1 [12,23].

Table 3. Overview of the mappings performed on the selected Medical Information Mart for Intensive Care (MIMIC) database (DB) tables to Fast Healthcare Interoperability Resources (FHIR) types.

<table>
<thead>
<tr>
<th>MIMIC IV DB</th>
<th>FHIR resource type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>Patient</td>
</tr>
<tr>
<td>Admissions</td>
<td>Encounter</td>
</tr>
<tr>
<td>Transfers</td>
<td>Encounter</td>
</tr>
<tr>
<td>Chartevents</td>
<td>Observation</td>
</tr>
<tr>
<td>Labevents</td>
<td>Observation</td>
</tr>
<tr>
<td>Procedureevents</td>
<td>Procedure</td>
</tr>
<tr>
<td>Prescriptions</td>
<td>MedicationRequest</td>
</tr>
<tr>
<td>Inputevents</td>
<td>MedicationAdministration</td>
</tr>
<tr>
<td>Microbiologyevents</td>
<td>Observation</td>
</tr>
<tr>
<td>Outputevents</td>
<td>Observation</td>
</tr>
<tr>
<td>Procedure_icd</td>
<td>Procedure</td>
</tr>
<tr>
<td>Diagnoses_icd</td>
<td>Condition</td>
</tr>
</tbody>
</table>

Table 3. Mapping of “diagnoses_icd” table to Condition Fast Healthcare Interoperability Resources (FHIR) resource.

<table>
<thead>
<tr>
<th>MIMIC format</th>
<th>FHIR resource format</th>
</tr>
</thead>
<tbody>
<tr>
<td>mimic.diagnoses_icd.subject_id</td>
<td>fhir.condition.subject</td>
</tr>
<tr>
<td>mimic.diagnoses_icd.hadm_id</td>
<td>fhir.condition.encounter</td>
</tr>
<tr>
<td>mimic.diagnoses_icd.icd_code</td>
<td>fhir.condition.code_code</td>
</tr>
<tr>
<td>mimic.diagnoses_icd.icd_version</td>
<td>fhir.condition.code_version</td>
</tr>
<tr>
<td>mimic.diagnoses_icd.long_title</td>
<td>fhir.condition.code_display</td>
</tr>
<tr>
<td>mimic.diagnoses_icd.long_title</td>
<td>fhir.condition.code_text</td>
</tr>
</tbody>
</table>

*MIMIC: Medical Information Mart for Intensive Care.
**Figure 2.** Flowchart showing an example diagnosis data being processed through the 5 stages in Fast Healthcare Interoperability Resources (FHIR) data harmonization pipeline (DHP). The first stage (a) includes querying of the diagnoses records, at the second stage (b-c) the data are mapped to FHIR standard, and the third stage carries out the syntactic resource validation. (f) If the FHIR resource is successfully validated, it is being transferred into the patient-model database (DB), and then (g) exported in a custom artificial intelligence (AI)-friendly JSON format.

**Discussion**

**Principal Findings**

The harmonization of EHR data is a crucial step toward increasing cooperation, interoperability, and quality of patient care in the clinical routine and medical research. To drive the harmonization of medical data forward, we developed the FHIR-DHP and evaluated it on key MIMIC IV tables. A detailed example of data standardization was presented for clinical diagnosis records from the MIMIC IV database. The FHIR-DHP allows the querying of health data in an isolated environment by employing an edge-computation solution and a communication server, which retrieve patient data and

https://medinform.jmir.org/2023/1/e43847
prestructure it for further mapping to the FHIR standard. A validation step ensures syntactic compliance and initiates the transfer of formatted data to the patient-model DB. The data export provides FHIR resources in a custom JSON file format.

Owing to the FHIR format’s multilayered nested structure, its accessibility for AI algorithms is low as it requires transformation into a format compatible with common data preprocessing tools. Thus far, a number of studies have attempted to solve this problem. However, the final output of these studies has not supported data selection criteria and filtering capabilities [14] and requires expert knowledge of FHIRPath query language [15]. In this study, we introduce a custom JSON format that represents a higher level of abstraction to support easier data selection based on the following 4 keys: feature_name, table_name, value, and metadata. Moreover, the newly developed JSON structure fits the expected data format of common data preprocessing frameworks, which are designed to work efficiently with tabular data. As a result, the output presented facilitates the generic and fast deployment of AI and patient cohort identification algorithms.

In comparison to [17,24], the details of FHIR-DHP execution inside the hospital environment in respect of protecting data privacy are discussed. This step, though crucial, is often omitted and left out of the published standardization protocols. The edge-computation solution sets up the FHIR-DHP in a privacy-preserving way where the preprocessing of the patient-related data is performed inside the hospital and is completely isolated from outside access. The so-called federated learning (FL) framework [25] can be integrated into the FHIR-DHP workflow to run algorithms locally, using data from the on-premises database in the respective hospitals and to merge model parameters centrally in the cloud without any patient data leaving the hospital. The FL framework requires data to be in a consistent format across various hospital systems. The developed pipeline achieves such a format and enables the scaling of AI applications.

Thus far, there are only 2 studies attempting to perform the mapping of an MIMIC IV database [26,27]. In [26], the mapping was performed on fewer tables than our approach (8 versus 12 tables). The FHIR mappings from [27] have been recently released and were not yet widely validated. Similar to the approach taken in [17,24,26], FHIR-DHP includes the verification of the performed FHIR mapping, which is essential to ensure the validity of data transformation and to adhere to FHIR version updates. Moreover, in comparison to [17,24,26], FHIR-DHP represents a generic approach to standardize EHR data and can be applied to various hospital database systems.

With the introduction of the FHIR-DHP into the hospital environment, a number of patient-stay parameters can be potentially optimized using AI-based algorithms. For example, the length of stay as well as mortality could be reduced [28], and patients suitable for trial treatment could be automatically and efficiently identified [29]. In consequence, the financial impact on medical providers in respect of personnel time and resources would decrease considerably. The FHIR-DHP aims to bring health care closer to digital transformation and thus toward “Healthcare 4.0” [30] by making EHR data usable “from bedside-to-bench.” By inverting the idea of translational research, in contrast to “from bench-to-bedside,” accessing the full potential of medical big data with AI will further inform and advance basic research.

Limitations
There are several limitations that we would like to emphasize. FHIR-DHP only works with a core standard of the FHIR format. Those core FHIR resource types have a bounded set of concepts that present a constraint to mapping accuracy. Although the standard resources can be expanded using a profiling technique or FHIR extensions, the use of those would make the FHIR-DHP less generic. Hence, we implemented the mapping using only the standard FHIR resources and omitted some of the MIMIC IV data features that did not have a matching concept in FHIR. Additionally, the FHIR mapping step is subject to the extent of the detail of the database documentation used to infer the semantic and syntactic properties of the data. A solution for an automatic concept recognition can potentially solve this problem. The existing approach in [6] is limited to a small number of FHIR resources and requires an extensive data preparation. Further experiments in this direction could alleviate the concept-matching problem and the requirement for a detailed database description. Moreover, the validation and robustness of FHIR-DHP needs to be tested on other EHR data sets to evaluate its generic setup. In addition, to validate the FHIR-DHP compatibility with machine learning pipelines, further experiments are needed.

Future Prospects
The proposed FHIR-DHP pipeline highlights the therein featured essential data standardization stages and holds the potential to becoming an interoperable harmonization system with an AI-friendly data format. FHIR-DHP enables interoperability and cooperation between clinical institutions and a rapid patient cohort identification for clinical trials; it also unlocks the potential of big medical data.

Conclusions
We provide a comprehensive approach to transforming unstandardized EHR data into a harmonized multilayered nested FHIR format and then to a more readable and more efficient AI-friendly JSON structure. We developed a 5-stage data harmonization pipeline, which includes validation checks. The AI-friendly format of hospital data allows the generic and fast integration of both AI and patient cohort identification algorithms. Harmonized and standardized health care data are of great value to advancing efficiency in big data processing, cooperation, and multicenter data exchange in the clinical sector, boosting medical research, patient care, and clinical trial cohort identification. The next steps would include validating our approach in a hospital environment and applying a privacy-preserving FL framework to make use of advanced AI deployment.
Acknowledgments
This work was partially funded by the German Federal Ministry of Education and Research under Grant 16SV8559.

Availability of Data and Materials
The MIMIC IV database used in this study is openly available to credentialed users who sign the “Data Use Agreement” at PhysioNet website [20]. The code is not publicly available due to privacy, but a demo is available from the corresponding author on request.

Authors’ Contributions
EW, SN, MK, JR, and AM were responsible for the study conception; EW and MK took part in data analysis; EW, SN, and EM created the figures; EW, MK, AM, and SN were responsible for methods. EW, EM, JR, and SAIK wrote the draft; BA, JB, PVB, JC, ARF, ASP, and NS reviewed and revised the work.

Conflicts of Interest
FB reports grants from the German Federal Ministry of Education and Research, German Federal Ministry of Health, Berlin Institute of Health, personal fees from Elsevier Publishing, grants from Hans Böckler Foundation, other from Robert Koch Institute, grants from Einstein Foundation, grants from Berlin University Alliance, personal fees from Medtronic and personal fees from GE Healthcare.

Multimedia Appendix 1
Transformation of data saved in custom JSON to tensor format.

[DOCX File, 998 KB - medinform_v11i1e43847_app1.docx ]

References


Abbreviations

AI: artificial intelligence
DB: database
DHP: data harmonization pipeline
EHR: Electronic Health Record
FHIR: Fast Healthcare Interoperability Resources
FL: federated learning
MIMIC: Medical Information Mart for Intensive Care
A SNOMED CT Mapping Guideline for the Local Terms Used to Document Clinical Findings and Procedures in Electronic Medical Records in South Korea: Methodological Study

Sumi Sung1, DPhil; Hyeoun-Ae Park2, DPhil; Hyesil Jung3, DPhil; Hannah Kang4, DPhil

1Biomedical Research Institute, Seoul National University Hospital, Seoul, Republic of Korea
2College of Nursing, Seoul National University, Seoul, Republic of Korea
3Department of Nursing, Inha University, Incheon, Republic of Korea
4Kakao Healthcare Corp, Seongnam-si, Gyeonggi-do, Republic of Korea

Corresponding Author:
Hyeoun-Ae Park, DPhil
College of Nursing
Seoul National University
103 Daehak-ro
Jongno-gu
Seoul, 03080
Republic of Korea
Phone: 82 2 740 8827
Fax: 82 2 765 4103
Email: hapark@snu.ac.kr

Abstract

Background: South Korea joined SNOMED International as the 39th member country. To ensure semantic interoperability, South Korea introduced SNOMED CT (Systemized Nomenclature of Medicine–Clinical Terms) in 2020. However, there is no methodology to map local Korean terms to SNOMED CT. Instead, this is performed sporadically and independently at each local medical institution. The quality of the mapping, therefore, cannot be guaranteed.

Objective: This study aimed to develop and introduce a guideline to map local Korean terms to the SNOMED CT used to document clinical findings and procedures in electronic health records at health care institutions in South Korea.

Methods: The guidelines were developed from December 2020 to December 2022. An extensive literature review was conducted. The overall structures and contents of the guidelines with diverse use cases were developed by referencing the existing SNOMED CT mapping guidelines, previous studies related to SNOMED CT mapping, and the experiences of the committee members. The developed guidelines were validated by a guideline review panel.

Results: The SNOMED CT mapping guidelines developed in this study recommended the following 9 steps: define the purpose and scope of the map, extract terms, preprocess source terms, preprocess source terms using clinical context, select a search term, use search strategies to find SNOMED CT concepts using a browser, classify mapping correlations, validate the map, and build the final map format.

Conclusions: The guidelines developed in this study can support the standardized mapping of local Korean terms into SNOMED CT. Mapping specialists can use this guideline to improve the mapping quality performed at individual local medical institutions.

(JMIR Med Inform 2023;11:e46127) doi:10.2196/46127

KEYWORDS

semantic interoperability; Systematized Nomenclature of Medicine–Clinical Terms; mapping guideline; local terms; mapping; guideline; SNOMED; nomenclature; interoperable; interoperability; terminology; medical term; health term; terminologies; ontologies
Introduction

South Korea is a leader in global information and communication technology. According to a 2021 Organization for Economic Co-operation and Development (OECD) survey of national health data infrastructure and governance, South Korea ranked second among OECD countries on data availability, maturity, and use [1]. However, data are not used fully, owing to a lack of interoperability and data security problems [2]. Interoperability is the ability of different information systems, devices, and applications to access, exchange, integrate, and cooperatively use data in a coordinated manner [3]. This occurs within and across organizational, regional, and national boundaries. Interoperability provides timely and seamless use of information and helps to globally improve the health of individuals and populations [3]. There are 4 levels of interoperability: foundational, structural, semantic, and organizational. Among them, the key strategy for ensuring semantic interoperability is the use of standard terminology that allows concepts to be represented unambiguously between the senders and receivers of information [4].

To achieve semantic interoperability, interface or local terms extracted from natural language written by a health care provider can be stored as reference terminology, such as in the SNOMED CT (Systemized Nomenclature of Medicine–Clinical Terms; SNOMED International). The stored terms with reference terminology can then be used in classification systems such as the International Classification of Diseases (ICD) for statistical purposes [2]. Health care providers in South Korea write medical records in natural language rather than using standard interface terms. Korean Standard Classification of Disease (KCD) codes are used for mortality and morbidity reports, and electronic data interchange (EDI) codes are used for national health insurance claims. Therefore, to fully use health care data, mapping terms extracted from phrases written in natural languages or using interface terms in medical records, disease classification codes, and national health insurance claim codes to reference terminologies are required.

Various efforts have mapped the terms used to document clinical findings and procedures in electronic medical records (EMRs), classification systems such as ICD 10th revision, and existing health care terminologies such as Logical Observation Identifiers Names and Codes (LOINC) and International Classification of Nursing Practice (ICNP) to the SNOMED CT in other countries [5-11]. The South Korean government, aiming to encourage the use of standard terminology in health care institutions, joined SNOMED International as the 39th member country and introduced SNOMED CT in 2020 to ensure interoperability. Subsequently, various efforts have mapped terms used to document clinical findings and procedures in EMRs or national health checkup questionnaires, classification systems including KCD-7, and EDI codes to the SNOMED CT [12-17]. Furthermore, these results are used for research purposes in the Common Data Model [17,18]. Individual medical institutions have attempted to map their terms to SNOMED CT. However, the mapping quality cannot be guaranteed due to its sporadic and independent map development. SNOMED International introduced Snap2, a tool to support mapping. However, since the tool is based on English source terms, it is difficult to apply to the Korean terms used in South Korea. This study, therefore, aimed to develop a guideline to ensure high-quality mapping of terms used to document clinical findings and procedures in the EMRs of local institutions in South Korea to SNOMED CT. The guideline focuses on a process of defining a relationship between concepts used in EMRs and the concepts of SNOMED CT [18]. The guideline does not include organizing a mapping team or reviewing existing maps. In addition, this guideline’s scope is limited to mapping to the SNOMED CT concept and excludes mapping to SNOMED CT post-coordinated expression.

Methods

The process of developing the mapping guideline was led by a mapping guideline development committee. The committee consisted of 3 mapping experts. All committee members had SNOMED CT mapping experiences spanning more than 5 years and have conducted various national projects, such as SNOMED CT mapping of KCD-7 and EDI codes supported by the Korea Health Information Service under the Ministry of Health and Welfare in South Korea. The development of the guidelines was conducted from December 2020 to December 2022.

To develop the mapping guidelines, the committee members first developed an overall structure based on a review of the existing guidelines and their own mapping and teaching experiences. An extensive literature review was performed in PubMed, MEDLINE, and Google Scholar. The existing mapping guidelines and previous studies [19-24] were reviewed according to these criteria: (1) scope and purpose, (2) involvement of the stakeholders, (3) rigor of development, (4) clarity of presentation, and (5) applicability [25]. After the review, the existing mapping guidelines and previous studies could not be used due to the following reasons: (1) not matched in scope or purpose because of mapping SNOMED CT concepts to other classification systems such as ICD-10 or ICD-9-CM [20,21], (2) limited to automatic mapping only without any information about mapping rules [21,22], (3) lack of currency in SNOMED CT, published more than 10 years ago [20,21], and (4) no detailed information on the mapping process [23,24]. As a result, version 2.0 of the SNOMED CT-AU mapping guideline, developed by the National Electronic Health Transition Authority of Australia, also known as the Australia Digital Health Agency [19], was chosen as the framework of the guideline. The SNOMED CT-AU mapping guideline clearly describes the preprocessing process, classification and validation of mapping results, and final map structures. It matches the scope and purpose of our SNOMED CT mapping guideline and uses the most recent version of SNOMED CT as a target code among the existing guidelines. In addition, it includes rigorous mapping examples. The following sections were adopted from SNOMED CT-AU: define the purpose and scope of the map, preprocessing source terms, mapping patterns, validation, and structure of the map.
Based on the committee members’ experiences, the most difficult aspects of mapping between local terms and SNOMED CT were extracting and understanding the source terms, developing search terms, and searching the target SNOMED CT. However, the SNOMED CT-AU mapping guideline does not describe how to extract and understand source terms from EMRs, how to develop search terms, or how to search for target concepts using the SNOMED International browser. It is difficult for beginners to apply the existing SNOMED CT mapping guidelines, such as those of the SNOMED CT-AU. Therefore, the committee added 4 steps to the guideline: extract terms, preprocess source terms using clinical context, select a search term, and use search strategies to find SNOMED CT concepts using a browser.

A guideline review panel was invited to validate the guideline from May 2021 to September 2022. The review panel consisted of 3 mapping experts: a professor with more than 5 years of SNOMED CT mapping experience working at a South Korean university and 2 mapping specialists with more than 3 years of mapping experience. The first author emailed, explaining the purpose of the study, and asked to review the understandability and usability of the developed guidelines and to add more mapping examples, if possible, to the panel. In addition, 2 graduate students with no mapping experience were asked to assess whether the guideline was understandable and helpful throughout the email. These 2 processes were repeated until no more issues were identified.

Results

Overview

The SNOMED CT mapping guidelines developed in this study consisted of 9 steps, as presented in Figure 1. Step 1 defines the purpose and scope of the map. Step 2 describes how to extract source terms from EMRs. Step 3 explains how to translate the source terms into Korean, how to define abbreviations or acronyms, and how to correct spelling or punctuation errors [19]. Step 4 explains how to understand the meanings of the extracted source terms to reflect clinical contexts. Step 5 describes how to select appropriate search terms to improve mapping [26]. Step 6 describes efficient strategies to search for SNOMED CT concepts using a browser [27]. Step 7 describes how to classify the correlation between source terms and target SNOMED CT [5,7,28-31]. Step 8 explains how to validate the adequacy and accuracy of the map [19]. The final step explains how to document the final map [32]. The SNOMED CT examples used in this guideline are taken from the international edition, released on January 31, 2023.

Figure 1. Overview of the developed guideline structure. SNOMED CT: Systemized Nomenclature of Medicine–Clinical Terms.

Step 1: Define the Purpose and Scope of the Map

To proceed with mapping, the mapping specialists must first define the purpose and scope. The purpose of the map can be for national health insurance claims, classification and statistics (eg, mortality and prevalence of specific diseases), knowledge management (eg, decision support system), health information exchange among institutions, studies (eg, Common Data Model), and patient care. Figure 2 presents examples of map purposes and scopes.

The scope of the map refers to the range of source and target codes and granularity and may differ according to its purpose. For example, if the map’s purpose is to provide prevalence
statistics for specific diseases, the source codes are restricted to KCD-7 in the clinical findings domain, and the range of target codes should be SNOMED CT concepts in the clinical findings, person, or event top-level hierarchy. If the map’s purpose is for national health insurance claims, the source codes are restricted to the EDI codes in the procedure domain, and the range of target codes should be SNOMED CT concepts in the procedures’ top-level hierarchy.

Map granularity can also vary according to the purpose of the map. For example, if the map is for national health insurance claims, the local term “alcohol-related seizure,” mapped to KCD-7 code G40.5 (special epileptic syndromes), should be mapped to the abstract SNOMED CT concept 230431001 (situation-related seizures [disorder]). However, if the map is for exchanging health information among institutions, the source code should be mapped at the granular level to SNOMED CT concept 308742005 (alcohol withdrawal-induced convulsion [disorder]).

**Figure 2.** Examples of map purpose and scope. EDI: electronic data interchange; EMR: electronic medical record; KCD: Korean Classification of Disease; SNOMED CT: Systemized Nomenclature of Medicine–Clinical Terms.

---

**Step 2: Extract Terms**

When the purpose and scope of the map are determined, mapping specialists should extract terms from EMRs while considering where and how clinical notes are documented. Patient diagnoses can be extracted from progress notes and other similar information sources. Family and past medical histories can be extracted from progress notes and initial nursing assessment records. Surgical procedure names can be extracted from progress notes and initial nursing assessment records. Evaluation procedures can be extracted from surgical notes. Evaluation procedures can be extracted from laboratory documentation.

These data sources are either structured, semistructured, or unstructured medical records written in natural language; for example, as presented in step A of Figure 3, if we extract a “patient’s mother diabetes mellitus type 2,” first the interface terms from the family medical history sections in EMRs can be identified, and then the “DM” diagnosis and the “mother’s family relationship” can be extracted. The terms written in free text from the semistructured family medical history records sections of EMRs can be identified, and then the diagnosis “2 형 당뇨병 (diabetes mellitus, type 2)” can be extracted. We also can identify terms from unstructured records by reading all the documents and extracting the terms written in free text, such as “Past medical history of diabetes mellitus,” which requires natural language processing (NLP).

When conducting automatic term extraction through NLP, word segmentation throughout semantic analysis is required using a corpus [33,34]. For example, the term “diabetes mellitus” should be extracted based on its meaning, not by extracting “diabetes” and “mellitus” separately. For this process to be possible, a Korean medical corpus can be used, such as the medical terminology database released by the Korean Medical Association [35] or the Korean Medical Library Engine [36]. The results of automatic terminology extraction should be reviewed manually.

Alternatively, terms can simply be extracted from code systems such as KCD-7, EDI, and local hospital codes. In these cases, it is sufficient to extract the terms mapped to the code system without duplication.
**Step 3: Preprocess the Source Terms**

Preprocessing takes daily English terms as the source terms, for instance, by translating Korean terms into English. These Korean English (Konglish) terms are then translated into native English terms, with acronyms or abbreviations defined and spelling or punctuation errors corrected. If terms are written in Korean, preprocessing starts with translating the Korean terms into English. For example, “당뇨” written in Korean must be translated into “diabetes mellitus.” Konglish is often used in Korean clinical settings. It is necessary to change Konglish terms into proper English terms. In Korean clinical settings, for example, an evaluation procedure “초음파검사” is represented by various English words such as in SONO and sonogram. SONO, or sonogram, is the image obtained using ultrasound. If it is the name of an evaluation procedure, it is, therefore, appropriate to translate it into “ultrasound.”

If terms are written in acronyms or abbreviations, which are commonly used by health care providers, preprocessing may be required to rewrite them in complete words. The acronym ASD, used in departments of pediatric cardiology, was defined as “atrial septal defect.” The abbreviation MMG used frequently in South Korea, must be rewritten to “mammography.”

Terms should be corrected when misspelled. For example, “ascending aorta dilatation” should be corrected to “ascending aorta dilatation” to obtain the correct search results. Furthermore, if terms include incorrect punctuation, it is preferable to edit the spacing of the source terms. For example, the search for the source term “DeQuervain’s disease, wrist, Rt.” is missing a space, which must be added to produce “DeQuervain.” Other examples are presented in step B of Figure 3.

**Step 4: Preprocess Source Terms Using Clinical Context**

The terms preprocessed in the previous step can have different meanings depending on the clinical context, such as subject, temporal, or finding contexts. For example, the meaning of the preprocessed source term “diabetes mellitus” varies depending on which part of the EMR it is extracted from; it can be “diabetes mellitus,” “past medical history of diabetes mellitus,” “family history of diabetes mellitus,” or “no family history of diabetes mellitus.” “Past medical history of diabetes mellitus” means that the temporal context is in the past. “Family history of diabetes mellitus” means that the temporal context is current or in the past and that the subject is a person in the patient’s family. “No family history of diabetes mellitus” means that the temporal and subject contexts are the same as “family history of diabetes mellitus,” but the finding context is absent.

Clinical context can be inferred from the source of the structured or semistructured medical records or the text in which the source term is written. If the family history section is the source of the medical record, the subject is a family member of the patient, and the temporal context is the past or the present. If the past medical history section is the source, the subject is the patient, and the temporal context is the past or present. If the clinical context cannot be inferred from the structured or semistructured records, the mapping personnel should read the text in which the source term is written. Examples are presented in step C of Figure 3.

**Step 5: Select a Search Term**

To search for a target SNOMED CT concept, a search term comprising keywords of the source terms that have undergone preprocessing must be selected. The search success rate is increased if the mapping specialist understands the general naming convention used by SNOMED International [26]. For example, a diagnosis or chief complaint can be expressed as “finding or morphology + body structure,” a surgery or procedure name can be expressed as “procedure + body structure,” the past medical history can be expressed as “history + disease name,” and the family medical history can be expressed as “family history + disease name.” The body structure requires “structure” to be attached to its name. Most descriptions used singular rather than plural expressions. Stop words in NLP, which are frequently used words with restricted semantic specificity, should be excluded from search terms to improve the success rate. These stop words include articles (“a,” “an,” “the”), prepositions (“in,” “at,” “with,” “without”), conjunctions (“and,” “as”), and ambiguous adjectives or adverbs (“other,” “alone,” “single,” “side,” etc). Examples are presented in step D of Figure 3.
Step 6: Use Search Strategies to Find SNOMED CT Concepts Using a Browser

Use the First 3 Characters of Words

Searches can start with the first 3 characters of search terms. With the first 3 characters, the mapping specialist can prevent search failures due to differences in the forms of various words, such as nouns, verbs, and the use of the passive voices and differences in singular and passive voices. For example, if the source term “perforated small intestine” is entered into the browser, equivalently matched concepts cannot be obtained. However, if 3 characters of each word—“per sma int”—are entered, the concept “235741002 |Perforation of small intestine (disorder)|” is matched.

When a SNOMED CT concept appears, the parent, child, and sibling concepts—higher, lower, or at the same level in the hierarchy—must be reviewed to determine if it is semantically correct.

Search With Synonyms of Source Terms

If the target SNOMED CT concept is not identified by search terms, synonyms should be used as search terms for the source term. The synonyms can be obtained by searching the medical terminology database released by the Korean Medical Association [34] or by searching the Korean Medical Library Engine [36]. Otherwise, the synonym can be found by including “synonym” in search terms on search engines such as Google, searching the thesaurus, reviewing websites of prestigious medical institutions, or from previous studies. In addition, conversational English terms should be changed so that different terms with equivalent meanings are also considered. For example, the conversational English phrase “removal of prostate stones” should be converted into the medically accurate phrase “removal of prostate calculus” for a successful search. Table 1 lists examples of using synonyms in a search term. If a concept is searched using synonyms, the parent, child, and sibling concepts must be reviewed in the final process to confirm whether the concept is semantically correct.

<table>
<thead>
<tr>
<th>Table 1. Examples of using synonyms.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source term</td>
</tr>
<tr>
<td>Removal of prostate stones</td>
</tr>
<tr>
<td>Abdomen sonogram</td>
</tr>
<tr>
<td>Operation of nystagmus</td>
</tr>
<tr>
<td>Extraction of nail</td>
</tr>
<tr>
<td>Correction for rectocele</td>
</tr>
<tr>
<td>Cervical mucus test</td>
</tr>
</tbody>
</table>

Start With Broad or Narrow Terms

If the SNOMED CT concept is not found with synonyms of search terms, the target concept that is semantically equivalent to the source term can be searched by starting with broader or narrower terms. For example, “IUD ectopic” in the diagnosis means that an intrauterine device is in an abnormal place or position. If “intrauterine device ectopic” is entered as a search term, no search results are obtained. If “intrauterine device” (which is a broad term for a concept in the diagnosis) is searched for and the results are filtered through the “disorder” semantic tag, 16 disorder-related terms are retrieved. Among them, the “malposition of intrauterine contraceptive device” concept is semantically equivalent to the “IUD ectopic” concept. In this process, the parent, child, and sibling concepts must be reviewed to confirm whether the concept is semantically correct, as mentioned above in the final process.

Filter SNOMED CT Concept Using Semantic Tag

Critically, to map semantically equivalent target SNOMED CT concepts, the mapping specialist must filter results by semantic tag in the SNOMED CT browser when entering a search term. These tags are presented in Table 2. For example, if the source term “cold” in the diagnosis is entered in the browser without filters, 2 lexically matched concepts 82272006 (common cold [disorder]) and 84162001 (cold sensation quality [qualifier value]) with the same description “cold” are obtained. However, if the results are filtered through the semantic tag “disorder,” the semantically equivalent concept, 82272006 (common cold [disorder]), is immediately obtained. The diagnosis “renal cell carcinoma” must be filtered through the “disorder” semantic tag, and mapped to 702391001 (renal cell carcinoma [disorder]), not 41607009 (renal cell carcinoma [morphologic abnormality]).
Step 7: Classify Mapping Correlations

The mapping results can be classified according to the correlation between the source concepts and SNOMED CT concepts or according to map cardinality, depending on the number of concepts.

Mapping Classification According to the Correlation

According to the correlation, the mapping results can be classified as “1193548004 [Exact match between map source and map target (foundation metadata concept)],” “1193549007 [Narrow map source to broad map target (foundation metadata concept)],” or “1193551006 [Map source not mappable to map target (foundation metadata concept)].” When the source term is matched to the equivalent SNOMED CT concept, the map is classified as a “1193548004 [Exact match between source and map target (foundation metadata concept)].” When the source term is matched to broader SNOMED CT concepts, the map is classified as a “1193549007 [Narrow map source to broad map target (foundation metadata concept)].” When no concept broadly matches a source term, the map is classified as “1193551006 [Map source not mappable to map target (foundation metadata concept)].”

As an example, the source term “breast cancer, upper inner quadrant” in a diagnosis is equivalently matched to 373082000 (malignant neoplasm of breast upper inner quadrant [disorder]), and the map is classified as “1193548004 [Exact match between map source and map target (foundation metadata concept)].” Another example is that the source term “aortic valve stenosis occurred after mitral valve replacement” in the diagnosis has no equivalent SNOMED CT concept that can be equivalently mapped, so it can be mapped to the broadly matching “703223000 [Postprocedural aortic valve stenosis (disorder)].” In this case, the map is classified as “1193549007 [Narrow map source to broad map target (foundation metadata concept)].”

Classify Mapping According to Map Cardinality

If a single source code does not map to a single SNOMED CT concept, it is necessary to classify the map’s cardinality using a complex map. The mapping results can be classified as either “one to one” or “one to many” according to their cardinality. When a single source code is mapped to a single SNOMED CT concept, the map is classified as a “one to one.” When a single source code with broad or multiple meanings is mapped to multiple SNOMED CT concepts, it is classified as “one to many.” For example, the source term “right or left hemicolectomy” has 2 meanings: “right hemicolectomy” and “left hemicolectomy.” The source term is mapped to “359571009 [Right colectomy (procedure)]” and “82619000 [Left colectomy (procedure)],” respectively, and the map is classified as “one to many.”

Step 8: Validate the Map

The map is validated with internal and external validation. Internal validation may vary depending on how many individuals participated in the mapping and validation processes. When 2 mapping experts are involved, one maps the source terms and the other reviews the mapping results. Otherwise, both mapping experts could map the same source terms and then compare the maps constructed by each other. When more than 2 mapping experts are involved, a validation can be conducted by dividing them into 2 groups—mapping and reviewing groups. The map is deemed to be correct if the mappers and reviewers select the same results. If the maps differ, the results should be evaluated in a group discussion with the other mapper to agree on which SNOMED CT concept to use.

External validation can also be used, in which clinical or mapping experts who were not involved in the mapping process verify the validity of the mapping results. Methods for performing external validation include reviewing the sample mapping results and obtaining mapping results that were difficult to map or that were not agreed upon in an internal group discussion.

Step 9: Build the Final Map Format

There are 2 types of maps—simple and complex maps. A simple map is a representation of mapping from a term in other code systems to a SNOMED CT concept; this is comprised of the source code (id, term), target SNOMED CT (id, fully specified name), and map correlation. A complex map is a representation of mapping from a term in other code systems to one or more SNOMED CT concepts; this is comprised of the source code (id, term), target SNOMED CT (id, fully specified name), map correlation, and cardinality. If map cardinality is 2 or more, rows are added to the final map. When documenting the final map, versions of source codes and SNOMED CTs, mapping...
dates, and mapper information must be included. Figure 4 presents examples of the mapping format.

Figure 4. Examples of final map formats. SNOMED CT: Systemized Nomenclature of Medicine–Clinical Terms.

<table>
<thead>
<tr>
<th>Simple map</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source code</td>
</tr>
<tr>
<td>ID</td>
</tr>
<tr>
<td>B02</td>
</tr>
<tr>
<td>B05</td>
</tr>
<tr>
<td>A18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Complex map</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source code</td>
</tr>
<tr>
<td>ID</td>
</tr>
<tr>
<td>D00</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Discussion

Principal Results

A SNOMED CT mapping guideline has been developed for the terms used to document clinical findings and procedures in EMRs. It was developed based on a review of previous mapping guidelines and the literature and the experiences of the authors and the map guideline development committee members. During the development of this mapping guideline, we reflected on the methods of resolving the difficulties experienced in KCD-7 and EDI code mapping, such as preprocessing the source terms and selecting search terms. We also reflected on the methods and examples identified while teaching mapping specialists working in local institutions to map the terms used to document clinical findings and procedures in EMRs to the SNOMED CT at the Korea Human Resource Development Institute of South Korea since 2019 [37]. The mapping training program was evaluated as having high satisfaction among mapping specialists with an average of 4.25 out of 5 points in 2021 (1=very low to 5=very high) and was reported as helpful in conducting mapping during work when they returned to their job with 3.75 out of 5 points in 2021 (1=not helpful to 5=very helpful). The guideline developed in this study will therefore be useful for local mapping specialists to map the terms used to document clinical findings and procedures in EMRs to the SNOMED CT. The guideline will improve the mapping quality performed at each medical institution.

The guidelines developed in this study contain detailed mapping steps from the Korean perspective, not only in the new steps such as term extraction, syntactic and semantic preprocessing, search term selection, and searching strategies in the SNOMED International browser but also in the steps adopted from previous mapping guidelines and studies [19,27,38]. For example, South Korean EMRs are written in both English and Korean. When written in Korean, translation is mandatory during preprocessing, and even when written in English, the Korean way of expressing a medical concept in English differs from that in English-speaking countries, so additional preprocessing is also required. Since Korean mapping specialists are limited by their knowledge of English synonyms, examples of synonyms were included to make the guideline easier to understand. The guideline can, therefore, be applied in local medical institutions to map Korean terms used to document clinical findings and procedures in EMRs to SNOMED CT. They can also be used to develop SNOMED CT mapping guidelines in other countries.

Limitations

This study had some limitations. Since the guideline focused on specific clinical domains, other clinical domains, such as medicine, were not included. Future studies are required on the development of mapping guidelines for terms used to document other clinical domains. In addition, this guideline does not include a guideline for mapping to SNOMED CT post-coordinated expression. The post-coordinated expressions frequently used in South Korea can be added as a new concept in the Korean extension of SNOMED CT.

Conclusions

This study developed a SNOMED CT mapping guideline for the terms used to document clinical findings and procedures in EMRs at local institutions in South Korea. The guideline was based on existing mapping guidelines, the findings of previous studies, and the mapping and teaching experiences of the authors. The mapping guideline developed in this study consisted of the following nine steps: (1) define the purpose and scope of the map, (2) extract terms, (3) preprocess source terms, (4) preprocess source terms using clinical context, (5) select a search term, (6) use search strategies to find SNOMED CT concepts using a browser, (7) classify mapping correlations, (8) validate the map, and (9) build the final map format. The new guideline can be published on the website of the Korea Health Information Service. The guideline can be applied to local medical institutions when mapping Korean terms used to document clinical findings and procedures in EMRs to SNOMED CT. It will also support local medical institutions in standardizing their local code systems using SNOMED CT. Ultimately, the data quality of each local medical institution will be improved, allowing the data to be fully used in clinical
decision support systems, health information exchange, and clinical research.

Acknowledgments
This study was supported by the Korea Health Information Service, and by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare and Ministry of Science and ICT, Republic of Korea (grant HU22C0168).

Conflicts of Interest
None declared.

References
13. So EY, Park HA. Exploring the possibility of information sharing between the medical and nursing domains by mapping medical records to SNOMED CT and ICNP. Healthc Inform Res 2011;17(3):156-161 [FREE Full text] [doi: 10.4258/hir.2011.17.3.156] [Medline: 22084810]


27. SNOMED CT search and data entry guide. SNOMED CT International. 2017. URL: https://confluence.ihtsdotools.org/display/DOCSSEARCH [accessed 2023-04-08]


33. Korean Medical Library Engine. URL: http://www.kmle.co.kr/ [accessed 2023-03-02]


Abbreviations
- EDI: electronic data interchange
- EMR: electronic medical record
- ICD: International Classification of Diseases
- ICNP: International Classification of Nursing Practice
- KCD: Korean Standard Classification of Disease
- LOINC: Logical Observation Identifiers Names and Codes
- NLP: natural language processing
- OECD: Organization for Economic Co-operation and Development
- SNOMED CT: Systemized Nomenclature of Medicine–Clinical Terms
Practical Considerations for Developing Clinical Natural Language Processing Systems for Population Health Management and Measurement

Suzanne Tamang¹,², PhD; Marie Humbert-Droz¹, PhD; Milena Gianfrancesco³, PhD; Zara Izadi³, PhD; Gabriela Schmajuk³, MD; Jinoos Yazdany³, PhD

¹Division of Immunology and Rheumatology, Stanford University School of Medicine, Stanford, CA, United States
²Department of Veterans Affairs, Office of Mental Health and Suicide Prevention, Program Evaluation Resource Center, Palo Alto, CA, United States
³Division of Rheumatology, University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Jinoos Yazdany, PhD
Division of Rheumatology
University of California, San Francisco
10 Koret Way, Room K-219
San Francisco, CA, 94143
United States
Phone: 1 415 576 1000
Email: jinoos.yazdany@ucsf.edu

Abstract
Experts have noted a concerning gap between clinical natural language processing (NLP) research and real-world applications, such as clinical decision support. To help address this gap, in this viewpoint, we enumerate a set of practical considerations for developing an NLP system to support real-world clinical needs and improve health outcomes. They include determining (1) the readiness of the data and compute resources for NLP, (2) the organizational incentives to use and maintain the NLP systems, and (3) the feasibility of implementation and continued monitoring. These considerations are intended to benefit the design of future clinical NLP projects and can be applied across a variety of settings, including large health systems or smaller clinical practices that have adopted electronic medical records in the United States and globally.

(Keywords: clinical natural language processing; electronic health records; population health science; clinical decision support; information extraction)

Introduction
Natural Language Processing (NLP) has the potential to improve the delivery, quality, and safety of health care [1-7]. There have been numerous research applications, including the extraction of disorders, drugs, and procedures. Moreover, NLP methods have automated the extraction of information that is likely to be undercoded or not coded in a patient’s record, such as the severity of a disorder, their functional status, or social determinants of health [4,6-8]. However, examples of health systems operationalizing clinical NLP tools for real-world clinical decision-making, as well as population health management and quality measurement, are limited. This is a missed opportunity to turn rich, unstructured data into structured information that can be used for quality and performance initiatives within a health system or a professional field, or to make national-level comparisons [2,9-12].

To address the challenges translating research tools to clinical practice, we present practical considerations for NLP system stakeholders that can be used to position an early-stage research project for use in real-world decision-making and to eventually demonstrate institutional value. Our practical considerations are informed by prior literature and reports that describe a chiasm rather than a synergy between clinical NLP research and clinical practice. For example, Wen et al [13] share the Mayo Clinic’s Desiderata for the implementation of an NLP development delivery platform derived from 2 decades of implementing clinical NLP in their health system. Lederman et al [14] describe how existing clinical NLP systems “have delivered marginal practical utility and are rarely deployed into

https://medinform.jmir.org/2023/1/e37805

JMI|Med Inform 2023 | vo| 11 | e37805 |

(page number not for citation purposes)
health care settings” and call for a new paradigm of clinical NLP research for real-world decision support. Similarly, Newman-Griffis et al [15] call for a new paradigm and general principles for clinical NLP research that are focused on challenges posed by application needs and describe how these challenges can drive innovation in basic science and technology design. Referring to artificial intelligence systems in medicine more broadly, Topol et al [16] have also observed that “deployment of medical AI systems in routine clinical care provides an important yet largely unfulfilled opportunity”. We also draw from our own collective experience developing clinical NLP systems for research studies and in an operational capacity.

Our practical considerations can be used to support the development of applications that can push forward advances in clinical medicine right now. We also assess the current landscape of Clinical NLP tools and techniques on our adjoining public GitHub site, which can be updated by the research community as clinical NLP technologies evolve [17].

**Practical Consideration #1: Are Data and Compute Infrastructure Ready for NLP?**

“Garbage in, garbage out” refers to low-quality data, or “garbage,” that can result in misinformation. It was first used by US Army scientists to provide the intuition that computers cannot think for themselves, and that “sloppily programmed” inputs inevitably lead to incorrect outputs. Although this saying is over a half century old, it applies even more today, when powerful computers can record large amounts of data that are not fit for the intended use in a short amount of time.

Key questions that will help to determine NLP readiness of a new clinical corpus includes the following: (1) Are notes and note metadata reported in a timely way and with reasonable quality? (2) Is the unstructured free-text data ready for NLP techniques (eg, can the data be used to extract clinical concepts with an accuracy that is fit for the intended use)? And (3) Are the NLP algorithms feasible to execute in the production environment?

Assessing the quality of textual data—or “Data Readiness”—confronts on the problem of data quality by providing empirical findings about syntactic and semantic aspects of a clinical corpus as well as the associated note metadata such as patient identifiers, the date and time of the note, and the type of note. We define “quality” within the context defined by Kahn et al [18] where three dimensions are considered, including plausibility, conformance, and completeness. The results of a Data Readiness assessment help to predict the difficulty of building an NLP system for those data. The quality of free-text data can vary significantly between different note types within the same or across different electronic medical record (EMR) systems. For example, discharge summaries typically contain complete sentences and clearly demarcated sections. By contrast, intensive care unit (ICU) progress notes typically contain large quantities of digits that are not explicitly labelled as to whether they are vital signs, ventilator settings, or any of the many other quantitative measures that are monitored in critically ill patients. ICU progress notes also frequently contain large amounts of information in just one or two grammatically unstructured sentences. Ambulatory progress reports can range from a few sentences to longer documents with standardized formats.

In some cases, data sets that do not initially appear to be ready for NLP on an intended task can be further processed or sampled so that the data are more amenable to their intended use. For example, a data source can be preprocessed to remove notes that do not fit predetermined plausibility criteria, such as the known range of system availability to identify notes that have a plausible date, or notes that do not have an indicated date. However, this may not always result in data that are ready for NLP; in these cases, investigators should work with an organization leadership to improve data collection before undertaking an NLP project.

The institutional nuances of EMR clinical documentation processes require clinical NLP systems developed at other institutions to be customized to a new local data set. This uses specific preprocessing steps related to the provenance and structure of the source data. In prior work funded by the Agency for Healthcare Research and Quality that was based on the Rheumatology Informatics System for Effectiveness (RISE) registry, we found that simple summary statistics on note length in characters and words (“tokens”) were helpful to assess the quality of clinical notes from rheumatology practices across the United States [12]. The RISE registry began operation in 2014, and the free-text extraction covered the period between 2014 and 2018. It combines data from over 260 ambulatory outpatient rheumatology practices that collectively use more than 20 different EMR products. To assess the data readiness of RISE for health services research and to better understand the epidemiology of chronic rheumatic diseases, we first used note metadata. For example, we calculated the number of unique clinical notes recorded by year, as indicated by the time stamp of the patient note. Unique notes were determined by each entry of a textual document within the RISE database. We found that many notes had an invalid timestamp, with dates as early as 1800 and as far out as 8018. This suggested an opportunity to improve the quality of these data. We also found that simple summary statistics describing the textual data helped determine the potential informativeness of RISE for scientific and practical applications. Table 1 suggests that RISE contains many relatively short patient notes (mean of 34.57 tokens in 2018) as well as some longer, more traditional patient notes and letters (SD 203.01 tokens). These types of summary statistics are an important first step in NLP data readiness assessments.
Table 1. Mean, SD, minimum, mode, and maximum note length\(^a\) and word count\(^b\) for free-text patient Rheumatology Notes submitted to the American College of Rheumatology’s data registry, by year.

<table>
<thead>
<tr>
<th>Year</th>
<th>Note count</th>
<th>Length of note</th>
<th>Word count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>Min(^c) Mode</td>
<td>Max(^d)</td>
</tr>
<tr>
<td>2010</td>
<td>891,837</td>
<td>96 (353)</td>
<td>1 17</td>
</tr>
<tr>
<td>2011</td>
<td>1,238,711</td>
<td>128 (554)</td>
<td>4 17</td>
</tr>
<tr>
<td>2012</td>
<td>2,412,737</td>
<td>118 (559)</td>
<td>3 19</td>
</tr>
<tr>
<td>2013</td>
<td>3,409,806</td>
<td>120 (597)</td>
<td>3 19</td>
</tr>
<tr>
<td>2014</td>
<td>5,394,083</td>
<td>209 (1069)</td>
<td>1 19</td>
</tr>
<tr>
<td>2015</td>
<td>7,715,894</td>
<td>211 (1547)</td>
<td>1 19</td>
</tr>
<tr>
<td>2016</td>
<td>9,812,735</td>
<td>233 (1356)</td>
<td>1 19</td>
</tr>
<tr>
<td>2017</td>
<td>11,685,000</td>
<td>242 (1468)</td>
<td>1 19</td>
</tr>
<tr>
<td>2018</td>
<td>5,301,039</td>
<td>239 (1415)</td>
<td>1 19</td>
</tr>
<tr>
<td>Total</td>
<td>50,222,840</td>
<td>205 (1271)</td>
<td>1 19</td>
</tr>
</tbody>
</table>

\(^a\)Please note that 2018 is a partial year. Note length is indicated by non–whitespace characters and symbols.

\(^b\)Word count was estimated after deidentification of the Rheumatology Informatics System for Effectiveness corpus.

\(^c\)Min: minimum.

\(^d\)Max: maximum.

\(^e\)Not available.

To assess the readiness of data for specific linguistic analysis tasks, such as part-of-speech tagging or named-entity recognition, there are a variety of other descriptive statistics based on corpus linguistics that can be used to assess the quality of textual data. Some of these focus on gross characteristics of the data, such as the extent to which documents have clearly identifiable sections and the nature of the data in those sections. For example, lists such as diagnoses and medications usually have relatively clear boundaries, while family and individual medical histories may not. The presence or absence of sentence boundaries, as well as the length of sentences, are also important predictors of the effort required to build high-performing language processing tools. Other descriptive statistics assess textual data on the level of individual words. For example, textual genres with high levels of repeated word use (eg, fever and pain) can be easier to process than textual genres with high levels of words that only appear once (eg, misspellings and typographic errors).

In addition to data that are ready for NLP, automated information extraction algorithms require infrastructure that will allow for the efficient processing of large volumes of new patient notes. There must be discussions at the design phase of the project to ensure that any research products can be operationally tested, and if warranted, translated to operational infrastructure. It is also important for the project to be updated and maintained if being used longitudinally with routine updates of notes.

If a project has no feasible pathway to operationalize the NLP system for real-world decision support, it might be possible that new resources, including institutional computing infrastructure, could be recommended and acquired.

### Practical Consideration #2: What are the Incentives for Adopting the NLP System?

Key questions that will help determine if the proper incentives are in place to support a Clinical NLP system are as follows: (1) will the NLP help to address an existing clinical need? (2) is there support from clinical leadership for the ongoing use of the NLP system? and (3) is there a financial incentive to adopting the NLP system?

Reporting from structured data has been the mainstay in health care practice for decades. The Sentinel active surveillance system for medical products and Observational Medical Outcomes Partnership (OMOP) initiatives helped to pioneer the use of common data models to support regulatory initiatives [19,20]. Building on OMOP’s common data models, the Observation Health Data Science Initiative’s extension has extended the OMOP schema to incorporate unstructured data with the “NOTE” and “NOTE_NLP” tables. It is likely that EMR databases will become even more powerful for regulatory initiatives when they can jointly leverage various data modalities such as patient notes or images for the purpose of improved patient care. However, in the absence of a specific clinical need that a system is designed to address, and without the proper incentives to use the system, it is unlikely that a system will be adopted for clinical uses such as decision support, regardless of performance on a research task. A successful system for population and precision health must be innovative, pragmatic enough to be deployed in a production environment and directly aligned with organizational incentives and clinical leadership’s priorities. It should support interoperability but also allow for customization to the nuances of different health systems. We discuss some of these challenges in the next section.
In cases where there is little or no organizational incentive to adopt a clinical NLP system, it is unlikely to succeed past the research phase. Therefore, working with leadership to identify the potential value to a health system and finding possible incentives to adopt such a system are important first steps.

**Practical Consideration #3: Feasibility of Implementation and Evaluation**

Key questions that will help to determine the feasibility of implementing and evaluating a clinical NLP system include the following: (1) What is the task (ie, clinical need) that this system seeks to address? (2) Are the clinical concepts of interest captured in structured data? If so, are there limitations to what can be extracted? (3) If NLP is justified, are simple NLP techniques enough or are more complex algorithms warranted? (4) Can the Clinical NLP tool be developed and implemented in a reasonable timeline to fulfill stakeholder needs? (5) What are potential sources of bias, considering factors such as the NLP approach, the data used to train the NLP tool, and the population to which it is applied?

An important early consideration is regarding the target population. In cross-validation over random folds, models are trained and tested over the same population. However, in practice, models are often developed in a training data set but applied to novel data that may originate from a different underlying population of patients or clinicians. Differences in clinical practice and workflow patterns, as well as lack of homogeneity in clinical language (as described above), can have large impacts on the transportability of models from where they were developed to a given target population. This is important to factor into the training assessment (eg, being aware of overfitting) and possibly also into model development. If an available external test set exists that represents the target population, it should be tested as part of the model development process to ensure that the NLP tool is portable and externally valid. Ideally, performance metric reporting should be required for all tools meant to be transportable outside of their training corpus.

There are multiple strategies for mitigating bias and improving portability of NLP tools. One source of bias may arise from the specific type of note used to develop a model; for example, an NLP tool developed only on ICU notes, pathology reports, or notes within a certain specialty may not generalize to other note types or clinical settings. Therefore, different note types should be incorporated into the training corpus, if in fact, the target corpus is intended to involve multiple types. Additionally, as previously described, incorporating a secondary data set that represents the target population for testing, apart from the primary data set used for training, can help ensure that the model is transportable and performs well across health care settings, EMRs, and patient populations.

To evaluate model performance, one must decide at which level the assessment should occur, that is, at the mention, document, or patient level. NLP models can be evaluated by their precision (positive predictive value), recall (sensitivity), specificity, \( F_1 \)-score (harmonic mean of precision and recall) and overall accuracy compared to a “gold-standard” test set of reviewed text [5,9]. However, the text-specific evaluation may not be as important as the document or even patient-level performance, especially if multiple mentions per patient occur, or structured data fields are being incorporated into the evaluation in conjunction with NLP annotations. Therefore, although at the mention level, the NLP model may correctly identify a patient as positive, it may be that it is only when combined with the additional information (other mentions, lab results, etc) that the output and model performance are clinically important.

As important as model performance at the time of development is, more crucial may be the model performance over time. Validation of NLP models is key both retroactively and prospectively, as data change longitudinally. It is important for models to be evaluated continually to determine whether they should be fine-tuned and updated, and whether any biases exist. For example, this may involve updating rule-based code to reflect changes in language representation or reevaluating or redeveloping deep learning–based NLP models.

If a clinical NLP system does not address a known and ideally high-priority clinical need, it is less likely to be adopted into practice. However, it may be possible to adapt the system to address a need identified by organizational leadership. If it does not initially show good performance, continued development may help to improve the clinical systems accuracy, especially if linguistic annotation data can be generated and made available for training a better model. Lastly, in some cases where additional expertise can be used for a project, it may be possible to meet a project deadline that would otherwise not be possible. Importantly, having a strategy, including a business plan, for maintaining deployed models is important to ensuring that their clinical application is sustained.

**The Potential Role of NLP in Real-World Decision Support**

NLP has the potential to improve population health outcomes in the United States. For example, in the inpatient care setting, NLP systems could reliably identify individuals with symptoms of diarrhea reported in progress notes and feed these data into algorithms for *Clostridioides difficile* testing. Inpatients with falls documented in clinical notes could trigger alerts to discontinue sedatives or narcotics. In the outpatient setting, NLP can be used to assess the severity of a disease or a postoperative complication. The NLP of free-text patient notes also creates opportunities for national, routine quality and performance measurement, which can support improvement in the value of health care delivered to patients at highest risk for poor outcomes [9-12,21-23]. As health systems across the United States move toward whole-person care paradigms, NLP systems can also be used to identify important clinical decision support factors that are undercoded or altogether absent from structured data sources in patient records, such as the presence of behavioral, psychosocial, and economic risk factors.

Predictive analytics is another area where incorporating clinical text has the potential to improve population health [5-7,24]. Most models for population-level risk stratification that use
health care data have exclusively relied on structured data, but several groups have demonstrated that in certain domains, adding information from clinical text can improve performance. Studies in this area reflect a wide range of tasks from predicting hospital readmissions to identifying patients at risk for suicide [2-13,17,21,22,24,25]. Such models can be used operationally to more accurately target a subset of a population for specific interventions designed to address modifiable risk factors.

Applications of NLP to streamline and facilitate quality and safety reporting are also emerging [9-12]. Federal reporting of quality and safety measures often places considerable burden on clinicians, sometimes requiring duplicate entry of similar concepts in the text of clinical notes as well as in structured fields that can be queried to calculate performance. Reliable extraction of relevant information from clinical notes would not only alleviate burdensome data entry but also greatly expand the types of concepts included in reporting programs. For example, guidelines in hematology support the routine collection of disease activity scores for patients with rheumatoid arthritis, but not all EMRs have structured fields to input these scores. Electronic quality measures that extract this information automatically from structured fields might miss scores that are documented only in clinical notes. NLP could be used to extract these scores and improve the validity and reliability of such quality measures[9-12].

While these and other applications of NLP have the potential to improve health care and population health, the successful deployment and dissemination of these applications has been limited. Given these barriers, how should the field move forward? In addition to our three considerations, we think it is critical that multiple stakeholders provide input from the start of NLP projects. Practicing clinicians can ensure the focus of the work is clinically relevant, fulfills an unmet need, and is aligned with current clinical workflows; clinical informaticians can provide insight into whether data systems are available to scale valid NLP algorithms, and health care administrators can lend insight into IT resources required and the feasibility of scaling and sustaining systems. Until there is stakeholder alignment and investment in a project, impact and scalability are likely to be limited. Similar to many new technologies in medicine, alignment often requires the development of the NLP program as a value proposition that either clearly impacts operational efficiency, revenue, quality and safety, or patient outcomes. Moreover, stakeholders need to be integrated into the software development life cycle to ensure the product’s ongoing implementation is successful.

**Conclusion**

The analysis of unstructured free-text patient data enables new ways in which scientific questions can be studied and health care can be delivered. Although such uses are promising, leveraging the clinical text data collected in the EMR and using these data in health care operations are not without substantial caveats. Opportunities to better align state-of-the-art systems developed by researchers to support the measurement of patient-reported outcomes and to support high-quality health care delivery can likely lead to improved outcomes. With a focus on designing practical applications that are aligned with clinical requirements and organizational incentives, the considerations listed here can be used to design a project-specific checklist for a variety of stakeholders. We also summarized the procedures for considering appropriate use of NLP in health and survey the current landscape of Clinical NLP tools. To support future work in this area, we have provided software and data set summaries, license, and other access requirements on our adjoining GitHub site, which we hope will serve as a continuously updated resource for the research community as technologies evolve.

**Conflicts of Interest**

None declared.

**References**


**Abbreviations**

- **EMR**: electronic medical record
- **ICU**: intensive care unit
- **NLP**: natural language processing
- **OMOP**: Observational Medical Outcomes Partnership
- **RISE**: Rheumatology Informatics System for Effectiveness
Original Paper

An End-to-End Natural Language Processing Application for Prediction of Medical Case Coding Complexity: Algorithm Development and Validation

He Ayu Xu¹*, PhD; Bernard Maccari²*, MSc; Hervé Guillain³, MD, DrPH; Julien Herzen², PhD; Fabio Agri⁴,⁵*, MBA, MD; Jean Louis Raisaro¹*, PhD

¹Biomedical Data Science Center, Lausanne University Hospital, Lausanne, Switzerland
²Unit8 SA, Lausanne, Switzerland
³Public Health Solutions Ltd, Promasens, Switzerland
⁴Department of Administration and Finance, Lausanne University Hospital, Lausanne, Switzerland
⁵Department of Visceral Surgery, Lausanne University Hospital, Lausanne, Switzerland
*these authors contributed equally

Corresponding Author:
He Ayu Xu, PhD
Biomedical Data Science Center
Lausanne University Hospital
CHUV, Centre hospitalier universitaire vaudois Rue du Bugnon 21
Lausanne, 1011 Switzerland
Phone: 41 0795566886
Email: he.xu@chuv.ch

Abstract

Background: Medical coding is the process that converts clinical documentation into standard medical codes. Codes are used for several key purposes in a hospital (eg, insurance reimbursement and performance analysis); therefore, their optimization is crucial. With the rapid growth of natural language processing technologies, several solutions based on artificial intelligence have been proposed to aid in medical coding by automatically suggesting relevant codes for clinical documents. However, their effectiveness is still limited to simple cases, and it is not yet clear how much value they can bring in improving coding efficiency and accuracy.

Objective: This study aimed to bring more efficiency to the coding process to improve the selection of codes by medical coders. To achieve this, we developed an innovative multimodal machine learning–based solution that, instead of predicting codes, detects the degree of coding complexity before coding is performed. The notion of coding complexity was used to better dispatch work among medical coders to eventually minimize errors and improve throughput.

Methods: To train and evaluate our approach, we collected 2060 cases rated by coders in terms of coding complexity from 1 (simplest) to 4 (most complex). We asked 2 expert coders to rate 3.01% (62/2060) of the cases as the gold standard. The agreements between experts were used as benchmarks for model evaluation. A case contains both clinical text and patient metadata from the hospital electronic health record. We extracted both text features and metadata features, then concatenated and fed them into several machine learning models. Finally, we selected 2 models. The first used cross-validated training on 1751 cases and testing on 309 cases aiming to assess the predictive power of the proposed approach and its generalizability. The second model was trained on 1998 cases and tested on the gold standard to validate the best model performance against human benchmarks.

Results: Our first model achieved a macro–$F_1$-score of 0.51 and an accuracy of 0.59 on classifying the 4-scale complexity. The model distinguished well between the simple (combined complexity 1-2) and complex (combined complexity 3-4) cases with a macro–$F_1$-score of 0.65 and an accuracy of 0.71. Our second model achieved 61% agreement with experts' ratings and a macro–$F_1$-score of 0.62 on the gold standard, whereas the 2 experts had a 66% (41/62) agreement ratio with a macro–$F_1$-score of 0.67.

Conclusions: We propose a multimodal machine learning approach that leverages information from both clinical text and patient metadata to predict the complexity of coding a case in the precoding phase. By integrating this model into the hospital coding
system, distribution of cases among coders can be done automatically with performance comparable with that of human expert coders, thus improving coding efficiency and accuracy at scale.

(JMIR Med Inform 2023;11:e38150) doi:10.2196/38150

KEYWORDS
medical coding; natural language processing; NLP; complexity prediction; prediction; decision support; machine learning; model; clinical decision support application; multimodal modeling; coding; algorithm; documentation; health record; electronic health record; EHR; development

Introduction

Background

Medical coding [1] is the translation of health care diagnoses and procedures into standard diagnosis and procedure codes using medical classifications and controlled terminologies. It is a strategic activity for funding hospitals and, therefore, its optimization is a priority in health care systems under financial pressure. In many countries worldwide, including Switzerland, hospital funding is based on the so-called Prospective Payment System [2,3] mechanism. In the Swiss Prospective Payment System, for example, inpatient stays are assigned to diagnosis-related groups [4] according to diagnosis and procedure codes derived from medical documentation, and each hospital stay is paid according to the diagnosis-related group to which it is assigned. Therefore, medical coding is closely linked, on the one hand, to medical documentation, and on the other hand, to hospital revenues. In addition to establishing reimbursement claims, medical codes are used for several other goals, such as setting budgets for planned hospitalizations or evaluating the quality of care by means of indicators such as complication rates after surgery.

The diagnosis and procedure codes of a specific case (ie, inpatient stay) are derived from clinical documentation such as discharge letters, surgical reports, physicians’ and nurses’ notes, and laboratory and radiologic results. The International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) [5], is usually used for coding diagnoses, whereas the classification system used to code procedures can vary from country to country [6].

Codes are manually entered into a hospital information system. In Switzerland, there are >200 coding rules that govern code entry and must be applied by medical coders. The latter are health care professionals who have undergone specific training for this purpose. However, despite training, medical coding remains a complex, quickly evolving, time-consuming, and error-prone task. In our tertiary academic medical center, medical coding staff have been divided into specialty teams since 2018. In a batch of cases, 50% are distributed to a “common pot,” and the other 50% are distributed to the corresponding specialty teams of medical coders. The cases in the “common pot” are distributed randomly to each team. A higher percentage of cases for the specialty teams is not envisaged for 3 reasons. First, it could lead to a loss of knowledge in general coding. Second, it could cause boredom for medical coders. Third, it will not always be possible to guarantee a sufficient number of cases for certain teams. Thus, a way to increase the efficiency of the current distribution of work without going toward a counterproductive overspecialization [7] is to force cases requiring high expertise to be assigned to experienced and specialist coders. This approach is only possible by detecting the complexity of the cases in advance before they are distributed and coded.

In recent years, artificial intelligence (AI) methods have been increasingly proposed to improve the efficiency and accuracy of medical coding. Their main goal has been to support medical coders in finding the most appropriate diagnosis and procedure codes for a given medical documentation. Conventional models, deep learning models such as convolutional neural networks and long short-term memory, and transformers have been trained and tested on automatic coding tasks using publicly available data sets in English [8-13]. Recently, this work has also been expanded to non-English corpora such as the French corpus [14,15]. In addition to the academic approach, commercial software for automatic coding has also been developed and introduced to the market. For example, commercial software such as ID SUISSE [16] applies rule-based algorithms to perform automatic coding. Their principle is to use a prebuilt dictionary of ICD-10 codes and their text labels, try to find clinical text that matches the labels, and then convert the text to ICD-10 codes. More recent tools such as Collective Thinking [17] and 360 Encompass (3M) [18] have improved the rule-based algorithms with machine learning (ML) techniques. Finally, solutions such as Sumex [19] rely on statistical methods to analyze the distributions and combinations of ICD-10 codes to identify possible inconsistencies in the coding patterns.

Despite the increasing number of available solutions, the effectiveness of automatic coding is still limited. Among the best-performing ML models, although precision can reach approximately 75%, the macro–F1-score could only achieve 10% to 12% [12,20,21]. The results indicate that even the best models can only capture a small portion of medical codes from free text. Therefore, the improvement of medical coding using AI-assisted strategies remains an open challenge (Kaur R, unpublished data, July 2021).

Objectives

The purpose of our study was not to find a way to predict ICD-10 codes from medical records. Instead, it was to improve coding quality and efficiency by predicting coding complexity before the coding process. Our primary objective was to bring more efficiency to the coding process to improve the quality of coding by medical coders, and the means to achieve this is an innovative solution using ML. The innovation is to use ML to detect complexity, which is then used to better dispatch the work among medical coders. To the best of our knowledge, this
approach has never been used before. It allows for a more efficient distribution of cases according to coders’ abilities and experience. As such, we will be able to minimize potential human errors because of random assignment and uneven distributions of coding expertise within hospitals’ coding divisions or units. Eventually, by knowing the coding complexity up front, simple cases can be assigned to beginners or nonspecialist coders or AI-assisted systems to maximize their utility while complex cases for which AI-assisted tools are still inefficient are assigned to coding specialists or at least to experienced medical coders.

Depending on the amount of clinical documentation to be examined and other factors such as the length of stay or the diversity of medical specialists involved in the treatment of a patient, coding a case may be a simple or a really complex task. Once a case has been coded, it is typically easy for the person who has done so to classify the case into a complexity level, which represents the complexity of the coding activity. However, predicting the complexity level of a case up front is very time-consuming for a human coder as it requires a deep analysis of the entire documentation, which eventually is equivalent to conducting the coding process directly.

To predict the complexity of a coding task in the precoding phase in an automatic way, we used advanced natural language processing (NLP) techniques to analyze clinical texts and extract features that are predictive of the complexity of cases. We proposed an end-to-end approach that integrates the NLP and ML model into the hospital clinical data warehouse and end-user coding system. Our NLP and ML model predicts case complexity with an accuracy comparable with that achieved by expert human coders. Its beta version is currently under deployment at Lausanne University Hospital. To the best of our knowledge, we are the first to propose and develop this innovative approach.

The remainder of the paper is organized as follows. The application details are presented in the Methods section, and the performance and analysis are presented in the Results section. In the Discussion section, we discuss the values and importance of our application as well as the use of NLP in health care.

Methods

Ethics Approval

The Cantonal Ethics Commission for research on human beings of Canton Vaud granted a full waiver for this study given the its retrospective and quality assurance nature under Req-2022-00677.

Overview

We describe a typical medical coding workflow in Figure 1. After an inpatient (patient who is hospitalized overnight) is treated in the hospital, a discharge letter is produced. Medical coders analyze the diagnosis in the discharge letter and translate the diagnosis into International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) codes. Sometimes the coders need to refer to other clinical documents (eg, intervention protocol and laboratory reports) to translate the information accurately. The diagnosis-related group codes are computed based on the ICD-10 codes and are sent to the insurance companies for billing. The insurance companies reimburse the bills to the hospital based on the received diagnosis-related group codes. If the insurance companies find mistakes in the codes, they ask for revisions from the coding service. We provide an overview of our decision support system in Figure 2 and describe its integration into the hospital information system in Figure 3.

Figure 1. The general coding procedure in hospitals. DRG: diagnosis-related group; ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision.
Figure 2. Workflow of this study. (A) We extracted 2060 cases from the clinical data warehouse at Lausanne University Hospital (CHUV). The cases are rated by coders (B) with complexity ranging from 1 (simplest) to 4 (most complex). (C) We performed feature engineering and trained models on the labeled cases. (D) The final model can produce both predictions of the complexity and its confidence in the predictions.

Figure 3. Integration of our model into the coding service. (A) When an inpatient visits the hospital and their medical case has been produced, the clinical text and patient metadata are stored in our clinical data warehouse. (B) A workflow manager will extract new medical cases regularly and send the data to our model. (C) Our model is containerized and deployed to an execution environment, where it performs the prediction for received cases. (D) Model predictions, together with the confidence of the predictions, are presented to the end users through a user interface to support task distribution in the coding service.

Definition of Complexity
We use the term “coding complexity” to characterize the time and expertise required of medical coders to assign diagnostic codes to medical cases.

Expertise can be defined as the level of experience, medical knowledge, and mastery of coding rules. Therefore, a medical case can be complex by applying many coding rules without being difficult but increasing the possibility of attention errors. Other cases may be complex and difficult because of the medical knowledge they require for proper coding. Therefore, complexity was the measure chosen to categorize the cases.

If coding a medical case does not require much time and deep expertise, the coding complexity is low (level 1; Figure 4). Conversely, if coding a medical case requires a lot of time and deep expertise, the coding complexity is high (level 4; Figure 4).

Coding complexity, similar to pain or satisfaction, is a subjective quantity. A potential objective way of defining coding complexity can be provided by the automatic coding models. By passing the medical cases through automatic coding models and manually examining the confidence score and the completion and accuracy of ICD-10 code predictions, we could divide the cases into simple and complex groups. However, owing to the limited performance (ie, the very low recall score) of current automatic coding models regardless of language [12,20,21], this approach will not bring much value to our situation. Furthermore, if coding complexity could be measured using simple objective data (eg, similar to blood pressure), our multimodal modeling approach would be useless. Thus, in this study, our definition of coding complexity will focus on the subjective ratings provided by medical coders, aiming to minimize subjectivity by using ML approaches and to predict the subjective scores of complexity.
To train our ML model, we extracted 2060 medical cases from hospitalized patients (inpatients) in 2021. We organized 2 annotation phases, each lasting 1 week, for 28 coders to rate the cases’ complexity. During each annotation phase, the coders rated the complexity of the given cases based on an evaluation grid (Figure 4).

**Data Collection and Preprocessing**

**Data Source and Data Annotation**

A medical case contains 2 types of data: a patient’s medical dossier and patient metadata (Textbox 1). We collected 2060 cases in total from the annotation phases. We note that the coding team at our hospital consisted of coders specialized in different medical domains. Hence, during annotation, we also kept track of whether a case was coded by a specialist. For example, if the responsible unit for a case was the internal medicine unit and the coder who coded this case was specialized in cardiology cases, the case was considered as not coded by its specialist coder.

Of the 2060 collected cases, 1998 (96.99%) were annotated by 28 medical coders, with each case coded by only 1 coder to maximize the size of the annotation set. As different medical coders may have different perceptions of the complexity of the same case, we evaluated the interrater reliability by asking 2 expert coders to code another 3.01% (62/2060) of cases. These 62 cases also represented our gold standard to create benchmarks for the models’ performance. For case selection, we first trained several models using the 1998 cases; then used the best model’s prediction to predict the complexity of several cases from our data warehouse; and, finally, randomly selected 62 out of the predicted cases while making sure that the complexity distribution of these 62 cases followed the same complexity distribution as the annotated data set. Each of the 62 cases was rated by each of the expert coders, and they were considered specialists for all cases. These 62 cases are referred to as the gold-standard set.

**Textbox 1.** Data collected for training and testing the model.

- Patient metadata: responsible medical service, number of movements between medical services, age, gender, civil status, whether the patient was deceased, length of stay, and whether the case was coded by a specialist
- Medical dossier: discharge letter of each service, operating procedure, intervention reports, and death letter

**Metadata Preprocessing**

The missing patients’ metadata were imputed based on the nature of the data. For numerical values such as age and length of stay, the missing values were imputed with the median of the existing values because of their skewed distributions (Figure 5). For categorical values such as gender and civil status, the missing values were imputed with the mode of existing values.
Text Data Preprocessing

We tested both classic term frequency-inverse document frequency (TF-IDF)–based text encoding and ML-based text encoding, and different text preprocessing steps were applied accordingly. For TF-IDF text encoding, we first tokenized the text; then removed the stop words; and, finally, replaced the entities with their entity type. The second and third steps were used to reduce the noise and increase the frequency of important words to provide a better signal for the model. An example of processed text is presented in Textbox 2.

For ML-based text encoding such as fastText (Facebook AI Research lab) and transformers, no preprocessing was applied.

Textbox 2. An example of text preprocessing results.

- Original text: Le patient susnommé a séjourné dans notre service du 01.02 au 03.02, date de son retour à domicile.
- Processed text: ["patient," "susnommé," "séjourné," "service," "<date>," "<date>," "<date," "domicile," "]
Model Design

Overview

The overall approach of the model design was as follows. First, we extracted features from the preprocessed metadata and text data. Second, we tested 2 modeling approaches: framing the problem as a classification problem or as a regression problem. On the basis of the modeling approach, we used different metrics to evaluate the model performance.

Feature Engineering

As the values for the patients’ metadata have different scales, we applied standardization (z score) to the numerical data and one-hot encoding to the categorical data.

To extract features from free text, we used 2 methods: TF-IDF and word embeddings.

TF-IDF provides a numerical weight of how important a word is to a collection of documents (Multimedia Appendix 1). We tested 2 configurations of the TF-IDF method: using the top 10,000 frequent terms or using the top 1000 frequent terms. We found that, using the top 10,000 frequent terms, the models performed better than using only the top 1000 frequent terms. Thus, in the following sections, we only report the results from the TF-IDF vector using the top 10,000 frequent terms.

Word embeddings provide the vectorized representation of a word based on the context in which it appears. We tested three types of word embeddings: (1) word2vec [22,23] embeddings trained on 2.5 million clinical texts (12 GB) collected from the hospital’s clinical data warehouse; (2) the pooled output (CLS tokens) of the state-of-the-art French-language transformer model French-Language Understanding via Bidirectional Encoder Representations from Transformers (FlauBERT) [24], which was pretrained on 71 GB of French text collected from the internet; (3) the fastText supervised approach [25] with embeddings initialized with the pretrained word2vec embeddings of (1)—we tested fastText as it provided the subword approach that could reduce the impact of the out-of-vocabulary (OOV) issue. A detailed analysis of OOV for this study is provided in Multimedia Appendix 1.

Textbox 3 shows the sizes of the vectors extracted using the different methods. The detailed conversion methods are presented in Multimedia Appendix 1.

Textbox 3. Vector sizes of text feature engineering.

- Term frequency-inverse document frequency (vectors were extracted using scikit-learn [version 1.0.1]): 10,000
- fastText (initialized with customized embedding; fastText embeddings were extracted using fastText [version 0.9.2; Facebook Artificial Intelligence Research lab]): 100
- word2vec (customized; word2vec embeddings were trained using Gensim [version 4.0.0; RARE Technologies, Ltd]): 100
- French-Language Understanding via Bidirectional Encoder Representations from Transformers (FlauBERT; the FlauBERT embeddings and fine-tuned model were implemented using Hugging Face [version 4.17.0; Hugging Face, Inc]): 768

Model Architecture

The complexity of cases ranges from 1 to 4 with discrete values; thus, we can treat it as either a multi-class classification problem or as a regression problem. The tested models are presented in Figure 6.

For both classification and regression, we used different feature combinations as inputs to train the models. The combinations were as follows: (1) metadata only, (2) word embeddings only, (3) TF-IDF vectors only, and (4) TF-IDF concatenated with metadata.

The overall process of model implementation is summarized in Figure 7. During training, we applied 5-fold cross-validation to reduce overfitting. As the labels were unbalanced, we used stratified sampling for cross-validation in the classification models. We performed hyperparameter tuning of the most promising features and models. For TF-IDF, we optimized the number of words considered in the vocabulary (topmost frequent words) and text preprocessing (lower case, lemmatization, removal of stop words, and removal of nonalphanumeric tokens). For the gradient-boosted trees model, we tuned the number of estimators, learning rate, and maximum depth. Hyperparameters were tuned based on the average performance over all folds in the cross-validation sets using Bayesian optimization.

In addition, we tested the fine-tuning of the FlauBERT sequence classification model using the Hugging Face transformer library [26]. The FlaubertForSequenceClassification application programming interface provides a pretrained FlauBERT model with a classification layer of size 1024 on top. It takes raw text as input and outputs the predicted classes (in our case, which is the complexity level). Among all our experiments, our best results were obtained using the fine-tuned FlauBERT-base uncased model. Notably, we froze the first 11 encoder layers and trained the last encoder layer and the classification layer to limit overfitting. We also weighted each class differently in the cross-entropy loss to account for imbalance. We used the maximum sequence length of 512 tokens and a batch size of 32. In this manuscript, we only report the fine-tuned FlauBERT results obtained using this configuration.
Figure 6. Comparison of performance using different models and input features on the 5-fold–cross-validated training data set (1751 cases) and the best model performance on the test set (309 cases). Dashed vertical lines represent the baseline model results. Models are ranked based on the classification macro–F1-score in the figure. *Average per service: for a given case in a given service, it always predicts the average complexity of cases in this service. A total of 29 services have an average complexity of 2, a total of 5 services have an average complexity of 3, and a total of 1 service has an average complexity of 1. **Majority vote: always predicts the majority class (in our case, complexity 2) and serves as a baseline for model prediction performance. FlauBERT: French-Language Understanding via Bidirectional Encoder Representations from Transformers; TF-IDF: term frequency-inverse document frequency.
Figure 7. Feature engineering and modeling approach using word embeddings and patient metadata as model inputs. The fine-tuned French-Language Understanding via Bidirectional Encoder Representations from Transformers text classification model is not included in this flow.

Data Imbalance
Our data labels were strongly imbalanced, and we tried to overcome this issue by using oversampling and undersampling techniques. Our best model was trained using Synthetic Minority Oversampling Technique [27] for oversampling underrepresented classes followed by random undersampling for overrepresented classes. We also chose metrics to penalize models that did not predict underrepresented classes, such as the macro-$F_1$-score. Ordinal classification can also be an interesting “hybrid” approach. However, we leave trying more sophisticated classification approaches for future work.

Technological Stack
The ML pipeline leverages spaCy (version 3.1; Explosion AI) for preprocessing texts (using the French-language model “fr_core_news_md”), scikit-learn (version 1.0.1) to build complex pipelines that can work with cross-validation, and Optuna (version 2.10.0; Preferred Networks, Inc) to conduct hyperparameter searches. It also eases the deployment of the selected model as preprocessing is part of a single serialized pipeline. The other tools used to try other approaches were fastText for document classification, Gensim (RARE Technologies, Ltd) to manipulate pretrained word embeddings, and Hugging Face Transformers (Hugging Face, Inc) to use pretrained transformer models. Training was performed on a virtual machine with 64 central processing unit cores, allowing us to parallelize training, and an Nvidia RTX 3090 graphics processing unit for larger deep learning models.

The first version of the selected model is being deployed with Machine Learning Model Operationalization Management infrastructure in our medical coding service. The deployment details are presented in Multimedia Appendix 1.

Results
Metadata Analysis
Each team of coders had a set of medical specialties. We considered that a case was annotated by a specialist if the annotator was part of a team from one of the specialties involved in the case. Following this logic, 63.98% (1318/2060) of the cases were annotated by a specialist. We used this as a feature during training. At inference time, we could choose to request a prediction for whether the case would be coded by a specialist.

The distribution of the numerical metadata and categorical metadata is presented in Figure 5. To check if any of the metadata had significant predictive power on coding complexity, we performed Pearson correlations between the numerical metadata features and the complexity ratings; we also performed statistical tests on categorical features such as patient gender and marital status (Table 1). The results show that, in the precoding phase, features such as sentence length and number of medical services visited during a stay did not have strong effects on coding complexity. In the postcoding phase, the number of ICD-10 codes and Swiss Classification of Surgical Procedures codes showed correlations with coding complexity.

With these results, we propose that a future direction of NLP-
or AI-assisted coding could use the metadata and clinical text to predict the number of codes that a case may produce and then compare it with the actual codes obtained after the coding process to perform quality checks in the postcoding phase.

**Table 1.** Pearson correlations between the numerical metadata features and the complexity ratings in both the pre- and postcoding phases and statistical tests of the categorical features and complexity ratings in the precoding phase.

<table>
<thead>
<tr>
<th>Numerical features</th>
<th>Correlation or statistical test</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of tokens from all documents in a stay</td>
<td>0.44</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Number of documents produced in a stay</td>
<td>0.33</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Number of medical services visited during a stay</td>
<td>0.02</td>
<td>.35</td>
</tr>
<tr>
<td>Duration of the stay</td>
<td>0.41</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Age</td>
<td>0.25</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Sentence length</td>
<td>0.003</td>
<td>.83</td>
</tr>
</tbody>
</table>

**Categorical features**

| Marital status                                         | $F_{5, 2054}=14.05$             | <.001   |
| Gender                                                 | $t_{2058}=−3.70$                | <.001   |

**Other metadata available after coding**

| Number of ICD-10<sup>a</sup> codes                     | 0.55                            | <.001   |
| Number of CHOP<sup>b</sup> codes                       | 0.46                            | <.001   |
| DRG<sup>c</sup> cost                                   | 0.34                            | <.001   |

<sup>a</sup>ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision.

<sup>b</sup>CHOP: Swiss Classification of Surgical Procedures.

<sup>c</sup>DRG: diagnosis-related group.

**Coder Rating Analysis**

The complexity ratings of the cases are shown in Figure 8A. The most common rating was complexity 2 (1127/2060, 54.71% of cases), and the least common rating was complexity 4 (58/2060, 2.82% of cases). We used stratified sampling to select the training and test sets; hence, their distributions were nearly identical to the true distribution shown in Figure 8A.

The original medical service of a case may also affect its complexity. Figure 8B shows that the cases from the Department of Palliative Care have the highest average complexity, whereas cases from the Department of Thoracic Surgery have the lowest average complexity.

By analyzing the gold-standard set, where all cases were rated by 2 experts, we found that even the expert coders did not always agree with each other. Of the 62 cases, the 2 experts agreed on 41 (66%). However, they disagreed by more than one complexity level in only 3% (2/62) of cases (Table 2). The interrater reliability (Cohen κ score) was 0.49 between the 2 expert coders. If we consider one expert as the ground truth and the other expert as a predictive model, the macro−$F_1$-score of this “predictive model” can only achieve 0.67 (Figure 9), a moderately good score showing that the task can be learned but models will not achieve a very high performance.

The reason why coders rate the same case with different complexity levels is mainly subjectivity. This is also a reminder that subjective-rated labels are often noisy, and no model can achieve a perfect performance. The ratio of agreement between 2 expert coders gives us an idea of the performance we could expect from a model. If we consider one expert as the model that predicts complexities and the other expert gives true complexity labels, then the highest accuracy that this model (the former expert) can achieve is 66%. In this sense, when later analyzing our model’s performance, the 66% accuracy can be considered as one of the benchmarks. However, given the strong imbalance in the complexity labels, we should rely as well on the confusion matrix to compare the annotator-annotator agreement with the model-annotator agreement.

However, as mentioned in the Model Design section, our samples were highly imbalanced, and the accuracy metric lacked the ability to measure the model’s performance comprehensively according to the sample distribution. As there were 54.71% (1127/2060) of cases rated with a complexity of 2, a naive model that predicts 2 all the time could reach an accuracy of 54.71%, but it provides no value for solving our problem. To consider the imbalanced sample distribution, we used the macro−$F_1$-score together with accuracy to measure the model performance. The macro−$F_1$-score between the 2 coders was 0.67, which was considered as the other benchmark that we used to evaluate the model’s performance.
Figure 8. (A) The distribution of complexity ratings over all 2060 cases. (B) Average complexity rating by service. The green bars show the top 5 services, and the red bars show the bottom 5 services. CHT: thoracic surgery; ION: immuno-oncology; MIN: infectious diseases; OBS: obstetrics; PED: pediatrics; RHU: rheumatology; SIA: adult intensive care; SIP: pediatric intensive care; SPL: palliative care; URG: emergency department.
Figure 9. (A) The complexity rating comparison between 2 expert coders on the gold-standard set. (B) The comparison between the validation model’s predictions and average expert ratings on the gold-standard set. (C) The comparison between 2 expert coders’ ratings on the gold-standard set when grouping into simple (complexity 1 and 2) and complex (complexity 3 and 4) cases. (D) The comparison between average expert ratings and the validation model’s predictions on the gold-standard set when grouping into simple and complex cases. The average expert ratings are rounded up to the next largest integer.

Table 2. Absolute difference between expert 1 and expert 2 complexity ratings. The accuracy reached by expert coders was approximately 66% (41/62; N=62).

<table>
<thead>
<tr>
<th>Absolute difference in complexity ratings between expert coders 1 and 2 (number of complexity levels)</th>
<th>Cases, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>41 (66)</td>
</tr>
<tr>
<td>1</td>
<td>19 (31)</td>
</tr>
<tr>
<td>2</td>
<td>2 (3)</td>
</tr>
<tr>
<td>3</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>
Model Analysis

Overview

First, we wanted to study whether our approach worked on predicting coding complexity for medical cases. We made use of all the 2060 annotated cases (n=1998, 96.99% 1-coder–rated and n=62, 3.01% gold-standard cases). We split the 2060 cases into a training set (n=1751, 85% of cases) and a test set (n=309, 15% of cases) and tested our model architecture. Then, to validate the model’s performance with expert coders’ benchmarks, we left the 3.01% (62/2060) of gold-standard cases out as the test set and trained a model with the same architecture but with more training data (1998/2060, 96.99% of cases).

The Main Model

To train the models, we started by using either patient metadata only or word embeddings or TF-IDF vectors only as input features. The best-performing model using patient metadata was gradient-boosted trees (macro−F₁-score=0.46; accuracy=0.61 for classification; R²=0.15 for regression). The best-performing model using word embeddings was the fastText classification model (macro−F₁-score=0.47; accuracy=0.57; initialized with customized embeddings), and the best-performing model using TF-IDF vectors was gradient-boosted trees (macro−F₁-score=0.45; accuracy=0.62 for classification; R²=0.26 for regression).

The model using word embeddings did not outperform the model using TF-IDF vectors. Thus, we combined the TF-IDF vectors with metadata as input features to integrate information from both patient metadata and medical dossiers. The best-performing model used gradient-boosted trees and achieved a macro−F₁-score of 0.51 and accuracy of 0.59 on the cross-validated training set and a macro−F₁-score of 0.46 and accuracy of 0.58 on the test set. Figure 6 shows the performance comparison between different models on the 5-fold–cross-validated training data set and the test set. The detailed numbers can be found in Multimedia Appendix 1.

As performing well on underrepresented classes is important in our case, we report the macro−F₁-score as the first metric. Macro−F₁-score is the average of the F₁-score per class and is not weighted by the number of instances in the class. Unlike accuracy, this metric penalizes each class equally. On the basis of the macro−F₁-score, we selected our best model as the gradient-boosted trees trained with the combined TF-IDF and metadata features (referred to as the main model).

The confusion matrix (Figures 10A and 10B) shows that our main model confused complexity-2 and complexity-3 cases during training and testing. Figure 9A shows that, even for expert coders, there was no clear distinction when rating complexity 2 and 3 for a case. The difficulty to distinguish between complexity 2 and 3 could be due to the similarity between the 2 classes of cases. We noticed that our main model also had difficulties distinguishing between complexity 3 and 4 during training and testing. This performance could be due to the lack of examples. Although we performed oversampling using Synthetic Minority Oversampling Technique on cases with a complexity of 3 and 4, it still lacked variability in complexity-4 cases.

We then tried to merge complexity-1 and complexity-2 cases as “simple” cases and complexity-3 and complexity-4 cases as “complex” cases and tested the model as a binary classifier. The results (Figures 10C and 10D) show that the model performed well on distinguishing between simple and complex cases. On the training set, the model achieved a macro−F₁-score of 0.62 with an accuracy of 0.71. On the test set, the model achieved a macro−F₁-score of 0.65 with an accuracy of 0.71.
The Validation Model

To validate our model approach and compare it with experts' benchmarks, we trained a validation model using the 96.99% (1998/2060) of 1-coder–rated cases and tested it on the 3.01% (62/2060) of gold-standard cases. The architecture of the validation model was the same as that of the main model.

The comparison between the 2 expert coders' ratings (Figure 9A) shows that most of the expert coders' disagreements were on complexity-2 and complexity-3 cases, and the overall agreement ratio between the 2 coders was 66% (41/62), with a macro–$F_1$-score of 0.67. Table 3 and Figure 9B show the comparison between our validation model and the 2 experts' ratings on the gold-standard set. The model agreed on 53% (33/62) of the cases with expert coder 1 and in 63% (39/62) of the cases with expert coder 2. The validation model achieved a 61% agreement ratio with the average ratings of both experts, with a macro–$F_1$-score of 0.62.
Table 3. Comparison between our validation model’s predictions and 2 expert coders’ ratings on the gold-standard set.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Percentage of agreement</th>
<th>Pearson correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert coder 1 vs expert coder 2</td>
<td>66</td>
<td>0.70&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Model vs expert coder 1</td>
<td>53</td>
<td>N/A&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Model vs expert coder 2</td>
<td>63</td>
<td>N/A</td>
</tr>
<tr>
<td>Model vs ceiled mean of 2 expert coders</td>
<td>61</td>
<td>0.70&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup><sub>P<.001</sub>.  
<sup>b</sup><sub>N/A: not applicable.</sub>

When merging the 4 complexity levels into 2 (simple vs complex; Figures 10C and 10D), the agreement ratio between the 2 coders became 84% (52/62) with a macro-$F_1$-score of 0.76, and the agreement ratio between model predictions and average expert ratings became 0.89 with a macro-$F_1$-score of 0.82. The results indicate that the model is comparable with human experts’ performance and predicts in a very similar manner to that of human experts (Figures 9A and 9B).

Interestingly, for the gold-standard cases, our validation model managed to predict complexity-4 cases 100% correctly, which was different from the main model’s performance during training and testing (Figures 10A and 10B). As there were only 4 selected cases with a complexity of 4 owing to the sampling for expert cases, these cases could be extremely complex and, thus, easy for the model to identify.

Compared with other models that can provide higher accuracy but lower $F_1$-score, both the main model and the validation model were more practical in our concrete use case as it is important to predict diverse complexity levels rather than keep predicting a complexity of 2 for all cases (Multimedia Appendix 1).

**Classification Versus Regression**

We summarize the pros and cons of both approaches given our use case in Textbox 4.

---

**Textbox 4.** Pros and cons of the classification and regression approaches.

- **Prediction confidence**: many classification models output the confidence in the predicted class as a probability, whereas regression models typically do not provide such information out of the box (although CIs are sometimes possible). Confidence is useful for end users, meaning that they can disregard predictions with low confidence. It can also be used in the active learning module (Figure 11) to select new cases (with low prediction confidence and strong disagreement between prediction and coder perception) to retrain the model.

- **Interpretability of results**: using a classification approach enables the computation of $F_1$-scores, accuracy, and confusion matrices. These are more intuitive for end users. Note that, for regression, it is still possible to round prediction to apply these metrics.

- **Order of labels**: complexity scores are naturally ordered. Therefore, given a case annotated with a complexity of 4, a model should be penalized more for predicting a complexity of 1 than for predicting a complexity of 3. Regression methods consider order, whereas classification methods do not.
Discussion

Principal Findings

We presented different ML models that can predict the complexity of coding medical cases with 4 complexity levels. We first trained the models on all 2060 annotated cases. When only using patient metadata, the best model (gradient-boosted trees) could achieve a macro-$F_1$-score of 0.46, an accuracy of 0.61 for classification, and an $R^2$ of 0.15 for regression. By applying NLP methods to extract information from clinical text, the best model (fastText initialized with customized embeddings) could achieve a macro-$F_1$-score of 0.47 and an accuracy of 0.57 for classification. When combining patient metadata and NLP-extracted information, the best model (the main model in the Model Analysis section) achieved a macro-$F_1$-score of 0.51 and an accuracy of 0.59 on the cross-validated training set and a macro-$F_1$-score of 0.46 and an accuracy of 0.58 on the test set.

To evaluate our model approach with experts’ benchmarks, we trained our validation model using the same architecture as the main model on all except the gold-standard cases. Our validation model achieved an accuracy of 0.61 with a macro-$F_1$-score of 0.47 and an accuracy of 0.57 for classification. When combining patient metadata and NLP-extracted information, the best model (the main model in the Model Analysis section) achieved a macro-$F_1$-score of 0.51 and an accuracy of 0.59 on the cross-validated training set and a macro-$F_1$-score of 0.46 and an accuracy of 0.58 on the test set.

To the best of our knowledge, this is the first study to apply NLP and ML models to help differentiate the complexity of coding medical cases.

Clinical Importance

Lausanne University Hospital in Switzerland has 2 missions: guaranteeing medical services in an area and serving as a referral hospital. The dominance of cases with a complexity level of 2 (referred to as case 2) in the labeled sample cases can be explained by this double activity as the hospital not only concentrates on university or referred complex cases but also receives normal cases similar to other hospitals.

In our current medical coding service, the cases to be coded are distributed 50% to the team of the specialty and 50% to a "common pot." This team versus common pot distribution is done randomly without considering the complexity of the cases, leaving complex cases in the common pot and, conversely, depriving the common pot of "simple" cases of specialized resources. Note that, in our case, coders can still choose complex cases from the common pot even if the case is not in their specialty. Many coders care about diversity or learning other types of cases. The integration of this model enables them to choose the complexity consciously.

The dominance of cases 2 will have the effect of pushing a lot of cases into the common pot, reducing the number of cases arriving to teams of different specialties and, hence, reducing the ratio of common pot to specialists. The quality of coding of complexity-3 and complexity-4 cases will be improved as they will be redirected to the specialty teams or senior coders. However, this will also be at the risk of lowering the quality of coding of cases 2, which will end up in the common pot. Therefore, it will be necessary to maintain a 50/50 ratio between...
the common pot and the teams or senior coders and force cases 2 to be coded by teams or seniors as well. This adjustment will enhance the quality of coding of cases 3 and 4 and a maximum of cases 2. After our system is deployed, the new distribution considering the complexity predicted by our NLP and ML model will be monitored in terms of satisfaction of the coding teams and accuracy of coding. Furthermore, we will analyze the accuracy of coding in relation to the predicted case complexity to adjust the model design and more efficiently allocate the case distribution to coders.

In our current model, the complexity of the cases is defined by the coders from our medical service and is rated subjectively. By analyzing the model predictions for a variety of cases, it is possible to summarize the common features shared by the high-complexity cases and those shared by the low-complexity cases. The summarized features can be used to build a set of objective rules that can be shared with other clinical services or the medical coding services of other hospitals. For small hospitals or clinical services, which do not always have sufficient resources to train and build their own ML models, this set of rules can help them distribute the cases more efficiently. In contrast, if the summarized features could not distinguish well between the simple and complex cases, it may reflect that the case complexity is a subjective rather than objective measure. In this situation, the best way to generalize this subjective measure is to build a model, such as in our approach, to learn the highly nonlinear subjective measures.

The complexity of coding a medical case can approximately reflect the complexity of the corresponding clinical case. Our application can not only improve resource allocation in medical coding services but also be generalized to other clinical services. Indeed, coding complexity levels can also be used in decision-making processes to help arbitrate resource allocation among professionals in the same department but affiliated with different clinical services within the department. For example, in the surgery department, a similar approach can be applied to help study the need for resources for different subspecialties based on the volume of treated cases but also on their relative complexity. The generalized application can be integrated into different digital health care systems for automatic task assignment to avoid conflicts in an unfair workload distribution.

**Technical Importance**

OOV is an issue that can impair model performance. Although the word2vec embeddings used in this study were trained on our own clinical data, OOV was still present as the corpus we used to train the embeddings might not have been sufficient to cover all the clinical terms used in the medical discharge documentation. To mitigate the impact of OOV, we tested the fastText subword approach. However, as shown in the Model Analysis section, the model performance was not much improved because of the low OOV ratio of our data set, which was only approximately 8% in the 2060 selected cases for this study. We provide a detailed analysis of OOV in our corpus in Multimedia Appendix 1.

As new clinical documents are produced every day, our deployed model could also face the impaired performance caused by the OOV issue. The solution we propose in this paper to reduce the impact is to monitor the evolution of new OOV with respect to the training data set and retrain the word embeddings when needed. During the retraining phase, we will not only retrain the word embeddings but also retrain the models with coder feedback to further improve the model performance from the perspective of both feature engineering and model engineering.

In our study, we used FlauBERT, which is a pretrained French-language transformer, in 2 different ways. The first way to use it is to generate word embeddings as text features for model inputs. We then also tested a Hugging Face [26] implementation of the sequence classification model using FlauBERT. A detailed description of this approach is presented in Multimedia Appendix 1. The best performance using the transformer model directly achieved a macro–$F_1$-score of 0.47, which is similar to other models that only receive text as features. The model performance did not improve as much as expected. The reason could be that our data set was too small (only 2060 cases) compared with the size of the transformer model. Regarding this, we will continue collecting coder feedback on the predicted cases and use them to train the model continuously. With these approaches, we hope to improve the transformer model performance in the future.

We found that using TF-IDF vectors as text features provided better prediction performance than using word embeddings as text features. The fastText and FlauBERT embeddings were pretrained on a nonclinical corpus; thus, the represented context of the word could deviate from the context used in the clinical text. As shown in the Metadata Analysis section, the median document length per stay was 909 tokens. Common pretrained transformer-based models handle up to 512 tokens, and it is not obvious which subset of the document should be selected to pass to the model. Although it is possible to overcome this limitation by embedding each chunk of 512 tokens and averaging their embeddings, we believe that a substantial improvement over other methods is needed to justify the computation cost. Furthermore, fastText and word embeddings both perform averaging over all vectors of each document, which may dilute the signal too much given the number of tokens. In contrast, TF-IDF can preserve some of this information, which could be the reason why TF-IDF vectors outperformed word embeddings in our task. A future direction to improve the model performance could be to combine TF-IDF vectors with word embeddings as text features. TF-IDF vectors can be used as a weight of importance for the words, whereas word embeddings can represent the contexts of the words. By combining the two, we could obtain vectors that represent both the importance and context of the words comprehensively. Another possible approach to improve the model performance is to build a rule-based model from coders’ experiences and then combine the rule-based model with the ML model, which can increase both the interpretability and flexibility of the prediction. As the complex cases are more likely to have multiple laboratory tests and clinical examinations, we could also include this structured clinical information for future feature engineering.
By comparing our model’s predictions with the expert coders’ ratings, we found that the model could achieve an expert performance level (Figure 9). As rating case complexity is relatively subjective, even expert coders do not always agree with each other. This introduced another level of complexity to our study. However, by learning 1998 cases from the training set, our model’s performance became comparable with that of the experts.

One of the advantages of our model is that we used a multimodal approach. Structured data such as patient metadata can provide quantitative information about patients’ status. Clinical text can provide rich information on diagnostic and other assessments of patients, which are not usually presented in the structured data. By combining the two, we are able to maximize the information needed to evaluate the complexity of a clinical case. Our study used 1 model to process data of different modalities and make predictions. In future work, we propose using dedicated models for each data modality and combining the predictions of multiple models using another ML model to make the final prediction. The benefits of using multiple models are that (1) it is easy to plug in new data and new models into the architecture, which makes the model flexible to extend, and (2) it is easier to perform feature engineering and interpret the model’s prediction.

The advantage of classification models over regression models in our study was that classification models allowed us to produce the confidence of the predictions. By showing both the predicted complexity level and the confidence of the prediction, we are able to provide comprehensive information to end users. However, there are also limitations to our model. Of the 2060 cases we collected for this project, 54.71% (1127/2060) were labeled as complexity-2, and only 2.82% (58/2060) were labeled as complexity-4. The unbalanced data set affects the performance of the classification models, meaning that the models have a higher tendency to predict complexity 2 for a given case. This problem was tackled by oversampling the underrepresented cases and undersampling the overrepresented cases. The results showed that the model performed better with oversampling and undersampling techniques (Multimedia Appendix 1).

Our model will be integrated into our current coding system with an active learning module. Figure 11 shows the integration architecture. The model reads patient metadata and medical dossiers regularly from our clinical data warehouse through a workflow manager. The predictions are presented in the user interface of the coding software. When coders find that the prediction deviates from the perceived complexity, they can put their corrections in a feedback field. Coders’ feedback is stored and sent to the model for retraining. This integration architecture allows us to track and continuously improve the performance of the model.

Future Work
Future work can be carried out on different aspects. To improve the model prediction performance, we can continue working on feature and model engineering. In addition to the data we used in this study, there could be other patient data that can be useful to predict the complexity of cases. Regarding the text features, we could try different combinations of NLP tools to maximize the information extraction from clinical text. We will also continue working on reducing the OOV impact by retraining the word embeddings (both word2vec and fastText) and TF-IDF vectors every 6 months and use coder feedback as new training samples to retrain the models. To make full use of the advanced transformer models, we will not only keep training using the new samples but also explore ways to incorporate patient metadata into the model design. We will also work together with coders to establish a sound and interpretable rule-based model and then combine it with the ML model. The hybrid model can provide both flexibility and good reasoning in distinguishing cases.

Currently, most NLP applications focus on AI-assisted coding using rule-based or ML models. As stated before, the rules framing medical coding complexity are dynamic and change over time, preventing the rapid learning of the tool. Instead of using AI-assisted tools only for coding, it is possible to extend the AI-assisted scope from case preselection to postcoding quality checks. Our approach provides a possibility to preselect cases that are suitable for automatic coding and other cases for manual coding. After a case is coded, AI-assisted tools can provide a post hoc analysis of the code categories and combinations, aiming to find possible mistakes in the codes. This can be done by studying previous coded cases using statistical and NLP analysis.

We also aim to continuously evaluate the application’s impact on our medical coding service. After the integration, we will monitor the average time a coder spends coding a case and the average number of mistakes a coder makes for each case. By comparing the time and accuracy before and after the integration, we can obtain a quantitative measure of how much improvement the model can bring to the coders’ daily work.

In addition to monitoring the quality of coding, we will keep tracking the coders’ user experience. With the help of the active learning module, we are able to collect coders’ feedback on the model’s predictions. The model will be retrained based on coders’ feedback through iterations to improve the prediction performance. As discussed in the Clinical Importance section, our application can not only help with task distribution to current coders but also be used to select cases for training junior coders. Junior coders will receive simple cases at the beginning and gradually receive more complex cases. This approach can give junior coders enough exposure to a variety of cases with respect to their capabilities as well as evoke their interests in medical coding.
Acknowledgments

The authors thank the 2 expert coders, Mireille Nya Buvelot and Lionel Comment, and all coders in the Coding Division for their contribution to complexity annotations. They also thank Dr Mostafa Ajalloeian for providing advice on this project.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Illustrations on the text feature engineering, imbalanced data processing, MLOps infrastructure, model comparison table, OOV analysis, and transformers fine-tune methods.

[DOCX File, 538 KB - medinform_v11i1e38150_app1.docx ]

References


19. Sumex Suite: The Sumex Suite is an established invoice verification solution tailored to the needs of Swiss insurance companies. ELCA. URL: https://www.elca.ch/en/sumex-suite [accessed 2022-03-14]


Abbreviations

AI: artificial intelligence
FlauBERT: French-Language Understanding via Bidirectional Encoder Representations from Transformers
ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision
ML: machine learning
NLP: natural language processing
OOV: out of vocabulary
TF-IDF: term frequency-inverse document frequency

©He Ayu Xu, Bernard Maccari, Hervé Guillain, Julien Herzen, Fabio Agri, Jean Louis Raisaro. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.01.2023. 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.