JMIR Medical Informatics

Clinical informatics, decision support for health professionals, electronic health records, and eHealth infrastructures
Volume 7 (2019), Issue 4 ISSN: 2291-9694

Contents

Review

Electronic Consultation in Primary Care Between Providers and Patients: Systematic Review (e13042)
Freda Mold, Jane Hendy, Yi-Ling Lai, Simon de Lusignan. . . . 4

Original Papers

Identifying Opioid Use Disorder in the Emergency Department: Multi-System Electronic Health Record–Based Computable Phenotype Derivation and Validation Study (e15794)
David Chartash, Hyung Paek, James Dziura, Bill Ross, Daniel Nogee, Eric Boccio, Cory Hines, Aaron Schott, Molly Jeffery, Mehul Patel, Timothy Platts-Mills, Osama Ahmed, Cynthia Brandt, Katherine Couturier, Edward Melnick. . . . 18

Challenges With Continuous Pulse Oximetry Monitoring and Wireless Clinician Notification Systems After Surgery: Reactive Analysis of a Randomized Controlled Trial (e14603)
Prathiba Harsha, James Paul, Matthew Chong, Norm Buckley, Antonella Tidy, Anne Clarke, Diane Buckley, Zenon Sirko, Thuva Vanniyasingam, Jake Walsh, Michael McGillion, Lehana Thabane. . . . 28

Usability Factors Associated With Physicians' Distress and Information System–Related Stress: Cross-Sectional Survey (e13466)
Tarja Heponiemi, Sari Kujala, Suvi Vainiomäki, Tuulikki Vehko, Tinja Lääveri, Jukka Vänskä, Eeva Ketola, Sampsa Puttonen, Hannele Hyppönen. . . . 39

Cohort Selection for Clinical Trials From Longitudinal Patient Records: Text Mining Approach (e15980)
Irena Spasic, Dominik Krzeminski, Padraig Corcoran, Alexander Balinsky. . . . 49

Extracting Clinical Features From Dictated Ambulatory Consult Notes Using a Commercially Available Natural Language Processing Tool: Pilot, Retrospective, Cross-Sectional Validation Study (e12575)
Jeremy Petch, Jane Batt, Joshua Murray, Muhammad Mamdani. . . . 67

Combining Contextualized Embeddings and Prior Knowledge for Clinical Named Entity Recognition: Evaluation Study (e14850)
Min Jiang, Todd Sanger, Xiong Liu. . . . 78

Using a Large Margin Context-Aware Convolutional Neural Network to Automatically Extract Disease-Disease Association from Literature: Comparative Analytic Study (e14502)
Po-Ting Lai, Wei-Liang Lu, Ting-Rung Kuo, Chia-Ru Chung, Jen-Chieh Han, Richard Tsai, Jorng-Tzong Horng. . . . 89

XSL•FO
RenderX

Impact of Automatic Query Generation and Quality Recognition Using Deep Learning to Curate Evidence From Biomedical Literature: Empirical Study (e13430)
Muhammad Afzal, Maqbool Hussain, Khalid Malik, Sungyoung Lee. 104

Efficient Reuse of Natural Language Processing Models for Phenotype-Mention Identification in Free-text Electronic Medical Records: A Phenotype Embedding Approach (e14782)
Honghan Wu, Karen Hodgson, Sue Dyson, Katherine Morley, Zina Ibrahim, Ehtesham Iqbal, Robert Stewart, Richard Dobson, Cathie Sudlow.
1 2 4

Building a Semantic Health Data Warehouse in the Context of Clinical Trials: Development and Usability Study (e13917)
Romain Lelong, Lina Soualmia, Julien Grosjean, Mehdi Taalba, Stéfan Darmoni. 138

Characterizing Artificial Intelligence Applications in Cancer Research: A Latent Dirichlet Allocation Analysis (e14401)
Bach Tran, Carl Latkin, Noha Sharafeldin, Katherina Nguyen, Giang Vu, Wilson Tam, Ngai-Man Cheung, Huong Nguyen, Cyrus Ho, Roger Ho. 155

Fast Prediction of Deterioration and Death Risk in Patients With Acute Exacerbation of Chronic Obstructive Pulmonary Disease Using Vital Signs and Admission History: Retrospective Cohort Study (e13085)
Mi Zhou, Chuan Chen, Junfeng Peng, Ching-Hsing Luo, Ding Feng, Hailing Yang, Xiaohua Xie, Yuqi Zhou. 168

Exploiting Machine Learning Algorithms and Methods for the Prediction of Agitated Delirium After Cardiac Surgery: Models Development and Validation Study (e14993)
Hani Mufti, Gregory Hirsch, Samina Abidi, Syed Abidi. 177

A Bayesian Network Analysis of the Diagnostic Process and its Accuracy to Determine How Clinicians Estimate Cardiac Function in Critically Ill Patients: Prospective Observational Cohort Study (e15358)
Thomas Kaufmann, José Castela Forte, Bart Hiemstra, Marco Wiering, Marco Grzegorczyk, Anne Epema, Iwan van der Horst, SICS Study Group. 198

Differential Diagnosis Assessment in Ambulatory Care With an Automated Medical History–Taking Device: Pilot Randomized Controlled Trial (e14044)
Adrien Schwitzguebel, Clarisse Jeckelmann, Roberto Gavinio, Cécile Levallois, Charles Benaïm, Hervé Spechbach. 209

Automatic Detection of Hypoglycemic Events From the Electronic Health Record Notes of Diabetes Patients: Empirical Study (e14340)
Yonghao Jin, Fei Li, Varsha Vimalananda, Hong Yu. 219

Fast Healthcare Interoperability Resources (FHIR) as a Meta Model to Integrate Common Data Models: Development of a Tool and Quantitative Validation Study (e15199)
Emily Pfaff, James Champion, Robert Bradford, Marshall Clark, Hao Xu, Karamarie Fecho, Ashok Krishnamurthy, Steven Cox, Christopher Chute, Casey Overby Taylor, Stan Ahalt. 229

Navigating Through Electronic Health Records: Survey Study on Medical Students' Perspectives in General and With Regard to a Specific Training (e12648)
Anne Herrmann-Werner, Martin Holderried, Teresa Loda, Nisar Malek, Stephan Zipfel, Friederike Holderried. 250

Primary Care Physicians' Experience Using Advanced Electronic Medical Record Features to Support Chronic Disease Prevention and Management: Qualitative Study (e13318)
Rana Rahal, Jay Mercer, Craig Kuziemsky, Sanni Yaya. 260

XSL•FO
RenderX

Key Factors Affecting Ambulatory Care Providers' Electronic Exchange of Health Information With Affiliated and Unaffiliated Partners: Web-Based Survey Study (e12000)
John Pendergrass, Ranganathan Chandrasekaran. . . . 271

The Impacts of the Perceived Transparency of Privacy Policies and Trust in Providers for Building Trust in Health Information Exchange: Empirical Study (e14050)
Pouyan Esmaeilzadeh. . . . 282

Developing a Reproducible Microbiome Data Analysis Pipeline Using the Amazon Web Services Cloud for a Cancer Research Group: Proof-of-Concept Study (e14667)
Jinbing Bai, Ileen Jhaney, Jessica Wells. . . . 307

A Deep Learning Approach for Managing Medical Consumable Materials in Intensive Care Units via Convolutional Neural Networks: Technical Proof-of-Concept Study (e14806)
Arne Peine, Ahmed Hallawa, Oliver Schöffski, Guido Dartmann, Lejla Fazlic, Anke Schmeink, Gernot Marx, Lukas Martin. . . . 316

Opportunities and Challenges of Telehealth in Remote Communities: Case Study of the Yukon Telehealth System (e11353)
Emily Seto, Dallas Smith, Matt Jacques, Plinio Morita. . . . 329

Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation (e14325)
Patrick Wu, Aliya Gifford, Xiangrui Meng, Xue Li, Harry Campbell, Tim Varley, Juan Zhao, Robert Carroll, Lisa Bastarache, Joshua Denny, Evropi Theodoratou, Wei-Qi Wei. . . . 342

How Online Reviews and Services Affect Physician Outpatient Visits: Content Analysis of Evidence From Two Online Health Care Communities (e16185)
Wei Lu, Hong Wu. . . . 355

Measuring Regional Quality of Health Care Using Unsolicited Online Data: Text Analysis Study (e13053)
Roy Hendrikx, Hanneke Drewes, Marieke Spreeuwenberg, Dirk Ruwaard, Caroline Baan. . . . 371

Interpretability and Class Imbalance in Prediction Models for Pain Volatility in Manage My Pain App Users: Analysis Using Feature Selection and Majority Voting Methods (e15601)
Quazi Rahman, Tahir Janmohamed, Hance Clarke, Paul Ritvo, Jane Heffernan, Joel Katz. . . . 380

Age Assessment of Youth and Young Adults Using Magnetic Resonance Imaging of the Knee: A Deep Learning Approach (e16291)
Ana Dallora, Johan Berglund, Martin Brogren, Ola Kvist, Sandra Diaz Ruiz, André Dübbel, Peter Anderberg. . . . 392

Impact on Readmission Reduction Among Heart Failure Patients Using Digital Health Monitoring: Feasibility and Adoptability Study (e13353)
Christopher Park, Emamuzo Otobo, Jennifer Ullman, Jason Rogers, Farah Fasihuddin, Shashank Garg, Sarthak Kakkar, Marni Goldstein, Sai Chandrasekhar, Sean Pinney, Ashish Atreja. . . . 409

Viewpoint

Digital Health and the State of Interoperable Electronic Health Records (e12712)
Jessica Shull. . . . 242

XSL•FO
RenderX

Review

Electronic Consultation in Primary Care Between Providers and Patients: Systematic Review

Freda Mold[1], BSc, PhD; Jane Hendy[2], BSc, PhD; Yi-Ling Lai[3], MA, PhD; Simon de Lusignan[4], BSc, MBBS, MSc, MD (Res)

[1]Faculty of Health and Medical Sciences, University of Surrey, Guildford, United Kingdom

[2]Brunel Business School, Brunel University London, Uxbridge, United Kingdom

[3]Faculty of Business and Law, University of Portsmouth, Portsmouth, United Kingdom

[4]Nuffield Department of Primary Care Health Science, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Freda Mold, BSc, PhD
Faculty of Health and Medical Sciences
University of Surrey
Kate Granger Building, Surrey Research Park
Priestley Road
Guildford, GU2 7YH
United Kingdom
Phone: 44 1483 684636
Email: Freda.mold@surrey.ac.uk

Abstract

Background: Governments and health care providers are keen to find innovative ways to deliver care more efficiently. Interest in electronic consultation (e-consultation) has grown, but the evidence of benefit is uncertain.

Objective: This study aimed to assess the evidence of delivering e-consultation using secure email and messaging or video links in primary care.

Methods: A systematic review was conducted on the use and application of e-consultations in primary care. We searched 7 international databases (MEDLINE, EMBASE, CINAHL, Cochrane Library, PsycINFO, EconLit, and Web of Science; 1999-2017), identifying 52 relevant studies. Papers were screened against a detailed inclusion and exclusion criteria. Independent dual data extraction was conducted and assessed for quality. The resulting evidence was synthesized using thematic analysis.

Results: This review included 57 studies from a range of countries, mainly the United States (n=30) and the United Kingdom (n=13). There were disparities in uptake and utilization toward more use by younger, employed adults. Patient responses to e-consultation were mixed. Patients reported satisfaction with services and improved self-care, communication, and engagement with clinicians. Evidence for the acceptability and ease of use was strong, especially for those with long-term conditions and patients located in remote regions. However, patients were concerned about the privacy and security of their data. For primary health care staff, e-consultation delivers challenges around time management, having the correct technological infrastructure, whether it offers a comparable standard of clinical quality, and whether it improves health outcomes.

Conclusions: E-consultations may improve aspects of care delivery, but the small scale of many of the studies and low adoption rates leave unanswered questions about usage, quality, cost, and sustainability. We need to improve e-consultation implementation, demonstrate how e-consultations will not increase disparities in access, provide better reassurance to patients about privacy, and incorporate e-consultation as part of a manageable clinical workflow.

(JMIR Med Inform 2019;7(4):e13042) doi:10.2196/13042

KEYWORDS

referral and consultation; health services accessibility; primary health care; general practice; patient access to records; patient portals; Web-based access

XSL•FO
RenderX

Introduction

Background

The growth and ageing of the global population combined with increased expectations place enormous pressures on primary health care. Greater use of technology is seen as a partial solution to the complex challenges of delivering health care to an increasing and ageing population with more chronic disease. This is reflected in health policy in the United Kingdom, the United States, and elsewhere [1]. Technology-supported consultations provide more flexible, though different, style of the clinician-patient relationship. However, adoption has been a challenge [2], and there is limited evidence of benefit [3,4].

The United Kingdom has taken a strong interest in using technology to deliver care [5], mainly driven by the increased cost of emergency administrations. Between 2012 and 2013, there were 5.3 million emergency admissions to UK hospitals, at a cost of approximately £12.5 billion representing a 47% increase over the previous 15 years [6]. These increases have led to growing interest as to whether remote care reduces what is considered unnecessary doctor's appointments or avoidable hospital admissions. However, to be commissioned and mainstreamed into everyday practice, an innovation must show that it can provide significant system-level advantages effectively providing *more for less*. For example, one of the worlds' largest remote care trials, a whole system demonstrator project saw improvement in patients' quality of life [7-9]. Telemedicine has also shown benefits in terms of health outcomes, hospital admission, and in terms of cost-effectiveness [10-12].

In this study, we focus on electronic consultations (e-consultations) situated within primary care. Remote care comes in many forms, including telephone, video, text messaging, email consultations, Web-based portals for prescription orders, appointment booking, and patient access to online health records, or any combinations of all these [13], recognizing that research in this area is heterogeneous [14]. We have excluded telemedicine and telemonitoring and generally specialist-based care that focus on the long-term management of chronic conditions.

E-consultations are feasible, and reliable, and convenient [15], although in common with other digital innovation challenging to implement [16]. Despite the growing use of computerized medical records [17], it has been challenging to incorporate e-consultations into clinical workflow [18,19]. To date, trials show little or no significant difference between usual care and intervention groups in terms of clinical outcomes [20].

Objectives

The aim of this review was to assess the evidence of delivering e-consultations using secure email, messaging or video links in primary care. The objectives were as follows: (1) understand how e-consultations affect patients' access to services, their frequency of use and satisfaction, and any impact on health outcomes; (2) investigate professional and workforce issues, including potential changes in workload or flow (actual and perceived) and barriers to use; and (3) identify possible organizational or technology barriers and solutions to implementation.

Methods

Design

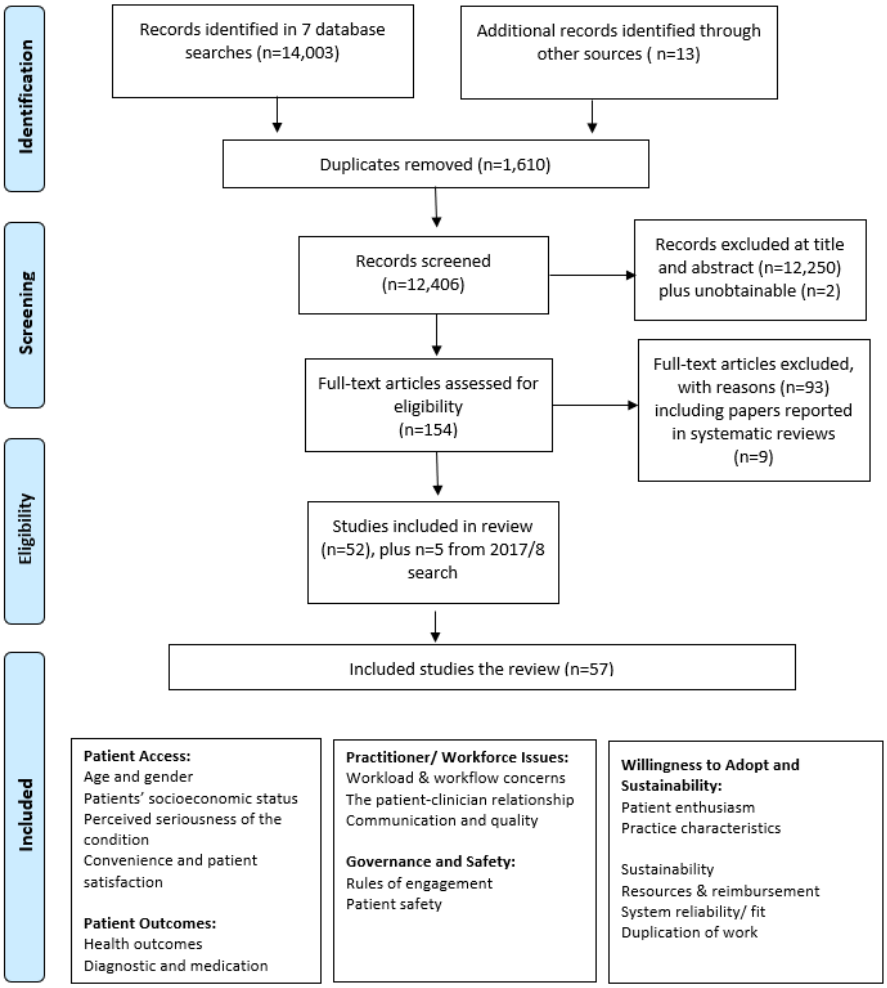
This systematic review follows Preferred Reporting Items for Systematic Reviews and Meta-Analyses [21] guidelines (Figure 1). The study aims were structured using the population, intervention, comparator, and outcome format [22]. The study *population* was defined as users or nonusers of e-consultation services, including both patients and carers and clinicians as well as support staff in primary care. The *intervention* related to synchronous or asynchronous e-consultation service used in primary care. Any *comparison* was used, including usual care. Several *outcomes* were identified including the following:

1. Patient(s): changes to service use including access to services (by specific patient groups, disorder or attributes of the user, frequency of attendance, and satisfaction), and impact on health outcomes.
2. Professional or workforce: workload and barrier to e-consultation implementation, impact on professional identity, consultation or revisit rates, and finally (if the information is available) quality and safety (ie, complaint numbers).

The protocol was registered on PROSPERO, the international database of systematic reviews, registration number CRD42015019152.

XSL•FO
RenderX

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses.



Information Sources and Searches

Advanced searches were performed across a range of bibliographic databases, including, the Cochrane Library, general medical databases (MEDLINE, EMBASE, CINAHL, via EBSCO platform), PsycINFO, EconLit, and Web of Science. A search was performed in the database OpenGrey for unpublished material.

Search strings were developed according to the index terms (Medical Subject Headings [MeSH] for MEDLINE) of each database together with keywords within the title or abstract using Boolean searches (AND, OR) with truncation and wildcard functions used (Multimedia Appendix 1).

This is an emergent and developing area, so recently published research was of key interest. We searched the literature from January 1, 1999, to March 1, 2017. No limits were placed on the evidence type (type of document, ie, systematic review), country of origin, or language of literature. Search results were exported into EndNote (v7.2.1). The search yielded 14,016 references, of which 1610 were duplicates and 12,406 were screened.

Setting and Participants

The systematic review focused on primary care and ambulatory care settings. Our principal participants in this study were patients and their family, caregivers (users and nonusers of e-consultations), and health care professionals (clinicians, allied health professionals, practice support staff, and managers). The technology is also relevant and was included in this review, focusing on current implementation, design, and the Information and Technology infrastructure underpinning e-consultations.

Eligibility Criteria

Search results were checked against the predefined inclusion and exclusion criterion (see Multimedia Appendix 2 for excluded studies). The inclusion criteria were based on the following: (1) a range of health care conditions, including any long-term chronic conditions managed in primary care (diabetes and hypertension) or routine conditions (skin conditions and sleep issues); (2) any asynchronous and synchronous use of emails and visual or video technologies (eg, Skype) used by both patients, carers and health care professionals in the e-consultations; and (3) no limitations were placed on the type of study (randomized controlled trial [RCT], qualitative, quantitative, and economic impact); however, study protocols were excluded as they do not contain original outcome data or review evidence.

XSL•FO
RenderX

Exclusions were studies focusing on telephone use alone (without the use of email, video or messaging) and any experimental studies which fail to provide specific outcomes measures or reported quality measures for service evaluation purposes only (eg, National Health Service Information Centre Quality and Outcomes Framework summary data). Finally, studies were excluded if they reported the use of medical records, email or telephone to recruit participants to research projects. This review only includes studies that performed e-consultations with primary care staff, with services performed in other settings (the community, secondary, or tertiary care) being excluded. Other studies were excluded if they focused on health promotion or education tools, which was not the primary focus of this review. Specifically, we were interested in e-consultations impact on access and health outcomes related to an illness event, rather than on long-term preventative strategies. Budgetary constraints excluded the authors from including studies that needed to be translated. Finally, to avoid possible bias and overreporting, studies were excluded if their results were already reported on in included review article [23]. All included studies were required to involve the patient in the e-consultation with their primary care provider. As such, provider-to-provider interactions were excluded from this review.

Data Selection

Evidence was sourced and retrieved by members of the research team (FM and YL). Results from searches were stored electronically. An initial screening of titles and abstracts was independently conducted by 2 team members (YL and FM). Inclusion queries were resolved through discussion at team meetings. Inclusion decisions were recorded using EndNote (v7.2.1). Further exclusions occurred once full texts were retrieved and when papers failed to meet the inclusion criteria or were a poor fit.

Data Extraction

Independent duel data extraction was undertaken by 2 researchers using a predesigned data extraction form (DEF) reflecting the core objectives of the study, including aims and objectives, study design, setting, type of e-consultation, outcome measures, comparator groups, and key findings. Data extracted also focused on a range of clinical outcomes (such as hemoglobin HbA_{1c} and blood pressure), behavioral outcomes (patient-clinician interaction, perceptions, acceptance, and system use), and organizational issues (such as functionality, usability, cost, and workflow). The DEF aimed to assist the authors to consistently retrieve the core contents of each study and aid in the organization of material before analysis.

Data Analysis and Quality Assessment

The analysis was executed in several stages. The first stage was the identification of the themes arising from the literature. The themes were developed over a series of meetings when the researchers clustered the results into higher order categories that seem to have coherence when summarized together. The aim of the clustering was to devolve a large and varied number of results into a smaller number of more easily understood, salient issues. The analysis was supported using a 3-stage thematic analysis process previously used [24,25] and guided by the Mayring framework [26]. The second stage included the assessment of evidence quality. Finally, themes were grouped against each of the research objectives to build up a comprehensive overview of the evidence. The analysis was undertaken by FM and JH with periodic input from the wider team.

Critical Appraisal

Studies based on qualitative, quantitative, and mixed methods designs were subject to critical appraisal, using the Mixed Methods Appraisal Tool (MMAT 2011 version) [27,28]. The MMAT tool uses criteria scored from 0% to 25%, with the overall score being 100. The interrater reliability of the MMAT was 0.94 [27]. No quality threshold was imposed, but caution was used to not overemphasize the contribution of evidence which had a low score (50% and less; n=7 papers, 25%). In reporting findings, greater emphasis has been placed on the literature with a higher MMAT score (>50% and above; n=41). For this work to be transparent, we have reported the MMAT score table (see Multimedia Appendix 3).

Results

Study Characteristics

A total of 57 studies were included in the review (n=57), including evidence from a range of countries, the United States (n=30) and the United Kingdom (n=13), with the remaining from Australia (n=3), Sweden (n=3), Finland (n=3), Canada (n=3), Denmark (n=1), and Italy (n=1), enabling greater ability for the findings to be generalizable (See Multimedia Appendix 4).

A variety of study designs were used, although the majority employed quantitative methods including descriptive designs such as surveys, and analysis of service frequency data (n=22) [29-50], quasi-experimental, cohort, or cross-sectional designs (n=10) [51-60], or RCTs (n=2) [61,62]. There was also a range of qualitative study designs using case studies, interviews, and focus groups (n=13) [63-75]. Only 6 studies had a mixed method design [76-81]. A total of 4 review findings were included [20,82-84].

A total of 5 overarching themes were identified across the literature: patient access, patient outcomes, workforce issues, governance and safety, and factors that impact on willingness to adopt and sustainability.

Patient Access

Age and Gender

The sociodemographics of patients using e-consultations was mixed. Users of e-consultations [29,38,81,82] and secure messaging [40,55] were primarily women [29,38,40,41,43,55,81,82] who used these services during working hours [29], presumably because of issues of convenience [41] in terms of organizing care or treatment for dependents (young children or older relatives) [30]. However, the evidence is far from conclusive, as 1 study found no statistical difference between genders [58], and another study found that more men (59/87) than women used the service

XSL•FO
RenderX

(28/87) [54]. The mean age of e-consultation users also varies. Some studies report prevalent users as being younger (45.9 vs 50.3 years, *P*<.01) [58], some as being 31-49 years (63/87, 77%) [54,82], middle-aged (50-65 years) [55], or over 60 years of age [43].

A study comparing patient characteristics receiving face-to-face or e-consultation in primary care (sinusitis and urinary tract infection [UTI]) found older people (≥65 years) to be less likely to use e-consultations (sinusitis, 28/475, 5.8%; UTI, 9/99, 9%, *P*<.001) [38]. In a similar study, age (over >65) was also associated with being less likely to use secure messaging (odds ratio [OR] 0.65, 95% CI 0.59-0.71) [55]. Early evaluation of e-consultations in one clinic suggested older patients found the concept of e-consultations confusing [81]. In contrast, a systematic review in 2014 suggests concerns about older patients being confused by them may be unjustified, and benefit could be gained if offered the right support [82].

Patients' Socioeconomic Status

Direct measures of socioeconomic status or failure to have health insurance, which we took as an indirect measure of socioeconomic status, were associated with limited affordability and access to emerging technologies [71]. Socioeconomically disadvantaged patients or those with poorer self-reported health were less likely to express an interest in communicating about their care using email or the internet [35]. In addition, patients who used email to communicate with their clinician were significantly associated with a higher annual family income (*P*=.007; >US $70,000) [34,43]. This group was reported to communicate with their clinician twice as much as those on lower incomes (<US $10,000-29,999) [34]. Moreover, a study investigating the characteristics of e-consultation patients found a high number of employed patients (for conditions such as sinusitis, 355/475, 74.7%; or UTI, 59/99, 60%; *P*<.001), suggesting out-of-office access is important for those in work [38].

In contrast, 1 study suggests the lack of medical insurance increased the odds of using 2-way visual and audible contact with health providers (OR 0.83, 95% CI 0.72-0.97) [41]. The cost of e-consultations for patients (email via a portal) varies between US $35 [29] and US $39 [39]. Earlier work found there may be a cost threshold, with 60.1% (149/248) of patients willing to pay up to US $10 or more per year. Only 31.0% (77/248) of patients were willing to pay up more—up to US $50 or more per year for secure email contact [31]. Willingness to pay did not differ by age (*P*=.06) [31].

Perceived Seriousness of the Condition, Convenience, and Patient Satisfaction

Patients reported using e-consultations when they did not perceive that a face-to-face consultation as warranted, even if conditions were chronic and long term such as diabetes and hypertension [29,57,79], or in cases where symptoms were routine or nonurgent, such as skin conditions, low-level pain, sleep issues, hemorrhoids, coughs, or sinusitis [29,48,79,81,83]. Unlike other studies, email contents analysis in 1 study suggests emails are useful when patients want to request information (symptom updates) or simple provider action (referrals, medications, treatments, or test result information) [63]. This suggests e-consultation [67,83] and online primary care visits [29] offer a convenient means through which to manage low-risk, nonurgent health concerns.

Differences also emerged when using technology to receive test results. Although many patients were willing to use email to obtain test results for cholesterol (1045/1229, 85.02%), less were willing to use this mode of contact for more serious conditions such as receiving a brain computed tomography scan test result (725/1229, 58.99%) [34]. Perceived seriousness also impacted on the mode of communication, with patients reporting favorable attitudes toward email but not text message or a Web page for the delivery of blood test results [44].

Convenience was the primary reported reason for choosing an e-consultation by patients across multiple studies [35,38,41,45,48,67,79,83]. Patient satisfaction [32,51,59,66,70] with immediate care received was increased [81] in the short term at 6 months [52]. Studies exploring the possible long-term impact of e-consultations over face-to-face encounters reported similar findings [40,52]. One study found no significant difference in the 30-day adjusted visit frequency at follow-up (2.35 visits per year before and 2.35 after portal messaging, *P*=.93) [40]. The subgroup analysis at 1 year of follow-up found an adjusted nonsignificant decrease of 0.1 visits per year (2.44 visits per year before the first message) and 2.34 after (*P*=.14) [40].

Timeliness of responses was important to patients using email [33,74,81] and was associated with satisfaction [84]. Patients had high expectations regarding the timeliness of responses for various Web-based services. Almost all patients in 1 study (2011/2260, 88.98%) expected a reply from email messages from clinicians within 24 hours, and 67.96% (1536/2260) expected responses or access to laboratory results within a 24-hour period [34]. More than 50% of patients expected a reply within 8 hours [34] and preferably the same day [74].

A range of studies found specific advantages to using e-consultations including improved access to care [66,70,83], both in the delivery of care outside of standard working hours [73] and care delivery to remote areas, time saved [32,36,45,73], and cost-saving including lost wages [73]. One evaluation study, of joint teleconsultations among general practitioners (GPs), specialists, and patients, found cost-saving for patients between €1,000.06 and €2700.50 by patients avoiding travel to emergency departments and for in-clinic visits or diagnostic examinations [50]. Finally, video and email consultations provide both patients and clinicians with opportunities to learn about health conditions and their management, through information and image sharing [65,74], offering the potential for more active patient engagement in the care process [52,63,82].

Joint e-consultations among GPs, specialists, and patients resulted in significantly higher levels of patient satisfaction (mean difference 0.33 scale points, 95% CI 0.23-0.43, *P*<.001) [62]. Satisfaction was also associated with a reduction of distance travelled [38] (average decrease of 170 kms) [32] or 1-way distance saved per patient (average 65 miles) [36]. Not surprisingly, greater e-consultation use was associated with the

XSL•FO
RenderX

winter months [38], especially for patients (and families) using video consultations in rural and remote communities [73].

Patient Outcomes

There is a lack of good quality evidence demonstrating positive patient outcomes from e-consultations because of the heterogeneity of existing evidence making an accurate assessment of benefits difficult [20]. In addition, there are limitations as to the longevity of follow-up data in trial material, again limiting the generalizability of any findings [20]. There were, however, several areas of potential benefit highlighted. Survey evidence suggests how telemedicine was as good as or even better than face-to-face consultation concerning the explanation of care to patients [32]. Email consultations were also shown to be clinically feasible in terms of diagnostic accuracy [84].

E-consultations may also play a role in the management of symptoms [51,57]. A study focusing on the management of hypertension in rural areas, using videoconferencing, found that the intervention group had a higher proportion of patients with blood pressure within treatment goals (systolic blood pressure, 140 mmHg; diastolic blood pressure, 90 mmHg), both at baseline and at follow-up, compared with a comparison group [57]. The intervention group was shown to have a higher probability of meeting their target blood pressure goal (OR 2.7, 95% CI 1.4-5.2) over the comparison group [57]. The quality of physical examinations in e-consultations was significantly worse regarding effectiveness (2.3 vs 4.9 for the face-to-face visit, $P<.001$), but history taking and therapeutic effectiveness were not significantly different [59].

Workforce

Several studies report clinicians' reluctance to use email with their patients because of increased workload concerns [37,40,46,84]. Clinicians reported improved efficiencies as email or secure messaging was described as taking little additional time [70] and encouraged care access [79]. However, as time is cumulative, even small additions, for example, between 2 and 6 min per email consultation [84], may lengthen the working day [70,76]. A quasi-experimental study reported how offering access to visit notes or email contact to patients was actually easier than expected and resulted in no change in the volume of messaging from patients [51]. Indeed, few clinicians reported longer visits (0%-5%) or more time answering patients' questions outside of face-to-face visits (0%-8%) [51]. Practice size has little effect on the overall workload [51]. Similarly, an evaluation of an email service found email services did not have any adverse time implications [66]. As such, practice partners were satisfied that the service worked effectively and did not negatively impact their day-to-day workload [66].

A retrospective cohort study of patients (n=2357) using electronic messaging (both secure messages and e-consultations) via a portal found, after the first message surge, no significant visit frequency differences (mean 2.35 annual visits per patient both before and after the first message, $P=.93$) [40]. Subgroup analysis indicated no significant change in the frequency of visits between high messaging users, or for those who had used messaging for longer. In other studies, e-consultations were found not to reduce telephone consultations [79] or number of office visits [70]. Evidence focusing on return visits to primary care found no significant differences in rates of early return visits for the same reason (e-consultations 20.2%, 46/228; face-to-face 19.6%, 98/500; $P=.86$) [58]. Similarly, a pilot study found less than <10% of patients who had an e-consultation (*similar to email*) required a follow-up face-to-face appointment [78]. Only the presence of moderate or more comorbidities was a significant predictor (OR 1.95, 95% CI 1.20-3.17; $P<.01$) relating to return visits for the same reason [58]. A small questionnaire to determine the feasibility of conducting follow-up visits using videoconferencing compared with face-to-face visits reported no significant difference in either group at 6 months [52]. Overall, findings from multiple studies suggest the use of e-consultations may complement in-person delivery (or could be a useful adjunct) to routine care [68,79,84], but this is reliant on the seriousness or risks associated with specific health conditions [58,68,79].

The Patient-Clinician Relationship

E-consultation was reported to impact on the patient-clinician relationship. The quality and safety of communication between groups may be affected as well as the interpersonal relationship (both positively and negatively). Access to physician notes and electronic messaging impacted on who initiated the direction of contact [70] and quality of the clinician and patient communication (content and tone) [51,63,73,79,83,84]. The ability to immediately exchange information (in a timely manner either asynchronous or synchronously) was reported to potentially improve the therapeutic relationship [84]. Clinicians felt patients' access to visit notes and electronic messaging strengthened their relationship with some patients because of a sense of enhanced trust, transparency, communication, and shared decision making [51,79]. Email exchange was also viewed as a useful tool to enable patients to express individual concerns and building a partnership, which was supportive and patient centered [63,83]. Video consultations in remote areas were also seen as an effective way to maximize home support, bring comfort to users in their own homes, and bring providers and families together from various regions [73].

In contrast, there were concerns about how e-consultations might negatively impact on the clinician-patient relationship [68]. These concerns include the need for professionals to communicate using nontechnical language [69] and their need to manage multiple tasks simultaneously (such as recording information), which might impact on the perceived engagement and attentiveness of the clinician in the Web-based interaction [75]. Indeed, in circumstances where nurses were present with clinicians in the e-consultation, clinicians themselves sometimes felt like outsiders, as the nurse and patient were better able to form a mutual bond via nonverbal communication and empathetic skills (such as maintaining eye contact) [75].

Governance and Safety

Within this review, governance, quality, and safety issues emerged in various forms, but not widely researched [39]. Only 1 study, a retrospective analysis of secure messaging and e-consultations was undertaken to assess the potential risk of

XSL•FO
RenderX

time-sensitive symptoms, such as chest pain or dyspnea [39]. Only 6 hospitalizations were related to a previous secure message (0.09% of secure messages), and 2 hospitalizations were related to previous e-consultations (0.2% of e-consultations, 2/892) [39]. Quality emerged in terms of the mode of care delivery either in terms of offering patients' information which impacts on their future service use, such as offering information which decreases the need for face-to-face encounters [60], enabling further opportunities to identify new problems during e-consultations [36] or raising perceptions of medicolegal liability [79].

Clinicians also raised concerns related to the lack of guidance about the *rules of engagement* [67], such as if an email is left answered [79] or level of confidence about taking medical history via e-consultations rather than face-to-face [52]. In response to the lack of guidance, GPs and patients have introduced their own rules of contact. These rules were not comprehensive and did not cover all eventualities [67]. Lack of formal practices and guidance was a recurring issue across the evidence [74,76,83]. A final concern is whether instructions through email can be adequately understood and correctly acted upon as intended by the sender [20,79] and whether some questions were appropriate for discussion via email [74].

Factors That Impact on Willingness to Adopt and Sustainability

Willingness to use technologies can be broadly divided into 2 related themes: the patient perspective and professional or organizational perspective. Low response rates among users were prevalent across studies [37,56,76], indicating differences in use depending on the level of experience between first users and those who are more experienced [36,46,76,81].

Patient enthusiasm was often dependent on their previous experience of using technology to manage their health [56]. In a longitudinal study comparing pre and post attitudinal changes to e-consultation found that first-time users were more likely to have a positive view, whereas experienced users were more negative (P=.025), suggesting patient use may tail off over time [54]. Other factors impact on patients' willingness to try e-consultations, including perceived severity of the condition (minor complaints) [79] and the actual mode of communication (secure email, direct access to records or laboratory results) [44].

General practices' willingness to adopt may also manifest in terms of the actual characteristics of the general practice (size and location) [71], with smaller practices in more deprived areas being less likely to use email [77]. Clinicians working in group practices were reported to be more in favor of using video technology for consultations [49].

In terms of sustainability, e-consultation may have repercussions in respect of further work across settings. A pilot mixed methods study found that specialist consultation requests made into primary care clinicians [78] resulted in GPs being asked to offer more patient advice, order diagnostic tests, or commence a new course of treatment [78]. Other work has echoed this potential service *push* to other health care providers with teleconsultations, resulting in a small number of additional diagnostic examinations (n=8) and hospitalizations (n=6) [50]. Similarly, an RCT examining whether e-consultations (called virtual outreach in the study) among GPs, specialists, and patients would reduce follow-up appointments found more e-consultation patients than the standard group being offered a follow-up appointment (502/971, 51.6%, vs 400/971, 41.1%; OR 1.52, 95% CI 1.27-1.82; P<.001) [62]. There was, however, variability associated with rates of follow-up according to specialty and site [62].

With regard to implementation and sustainability, there is limited evidence available about the cost-effectiveness of e-consultations, but the high cost of buying telemedicine equipment [46] and expense of implementing this technology is a concern for health care professionals [61].

Costs of clinicians' time to support joint consultations were unlikely to be offset against subsequent savings to health care services in the short term [61]. The total use of UK health care (NHS) resources over 6 months suggests that the overall mean cost per patient is significantly higher in the joint consultation group than the standard outpatient group by approximately £100 [61]. The significant reduction in tests and investigations in the joint consultation group resulted only in small cost reduction *downstream* [61]. Similarly, other studies recommend future long-term follow-up (over 6 months) to determine downstream outcomes and full evaluation of cost-effectiveness [62].

Delays in service delivery was also an additional concern with the provision of out-of-hours services. A small study assessing delayed response to patients' secure email messages (messages not opened after 12 hours or nonresponse after 36 hours) found both kinds of delays were higher on weekends (P<.001) (Friday-Sunday) [40]. Delay was more likely to be experienced by patients aged over 50 years (605/2357, 25.66% delayed; P=.013) [40]. The study suggests that these delays could be addressed by automatically rerouting messages to a 24-hour staffed support service or another mechanism to manage this after-hour workflow [40]. Provision of logistical support for a range of e-consultation methods may, therefore, be significant to enable long-term and efficient implementation of systems in primary care [62]. In addition, in 1 study, facilities which offered user support for those wanting secure messaging were found to have higher rates of adoption (2.13%) over other providers (1.52%; P=.006) [56].

Other notable barriers to implementation include commissioners' incentives (or direction of cost) for the introduction of remote services [65], the impact of size and location of practices [71], and organizational resistance [59,77]. From the provider's perspective, a mixed method study suggests email communication could be embedded into everyday practice and be remunerated similarly to usual clinic time, thereby potentially offering a new structure of care [79]. The direction of cost is illustrated in 1 study exploring the experience of Greek health care providers and their patients with the introduction of an e-consultation service [65]. The study found that there was no incentive for the health care system to introduce e-consultations as often patients incurred the cost of their own travel to the mainland for health care [65]. Implementation may also be

XSL•FO
RenderX

influenced by whether e-consultations in practice were resource- or reimbursement-driven [37,71,81].

The final sustainability consideration is system-level fit, the extent that e-consultations can integrate into existing services and the scalability of implementing this technology.

Scottish research on the uptake of an electronic clinical communication system reported that although the current system was beneficial, issues around system reliability, incompatibility of systems, and duplication of data hindered widespread uptake [45]. The main perceived barrier to adoption were views about the instability of computer networks across the region [45]. Technology design was also seen as critical in relation to ease of use and functionality for both patients and health care professionals [36,46,76,81] and can be directly linked to uptake or adoption [76]. Functionality is also important to clinicians [46,81]. This emerged in reference to possible technical failure, level of previous and current training needs, experiences of technology use (both positive and negative), and the condition, state and age of the available technology [61].

Mixed Methods Appraisal Tool Results

The overall MMAT study quality was moderate, with only 11 studies identified as excellent (100%). However, use of the MMAT, aided both description and appraisal of studies, helping to highlight the need for robust and larger trials as well as to fully explore the level of risk, both real and perceived [58,79].

As previously mentioned, generalizability of some studies was limited [46,55,78] in many cases by low participant numbers [37,52,68] or single or low number of study sites [34,35,37,40,43,51,66]. Owing to the heterogeneity of (OR and hazards ratio) measured outcomes across studies, the study team decided not to conduct a meta-analysis, as this may have resulted in a misrepresentation of the data.

Discussion

A total of 5 themes emerged which addressed our review objectives. These themes were patient access, patient health outcomes, workforce issues, governance and safety, and finally willingness to adopt and sustainability of e-consultations.

Patient Access

In understanding how e-consultations affect patients' access to services, there is evidence to suggest that e-consultations work well for some patient groups but not for others impacting on access, with the elderly and the poor less likely to use these services [36,39,56,72]. As such, there was a disparity between different users and under what circumstances patients are more willing to use e-consultations systems and why.

Patient Health Outcomes

There was also a lack of evidence of whether patient health outcomes improve with e-consultations [20]. Indeed, a potential limitation to this study is the dearth of studies reporting health outcomes from e-consultations. As such, there is a need for further high-quality studies to fully evaluate the usefulness of e-consultations in primary care, especially on how patient outcomes are affected and the long-term impact of e-consultations on the patient-clinician interactions.

Workforce Issues

In investigating professional and workforce issues, evidence suggests that e-consultations may increase patient expectations of care delivery [34] and complement existing in-person care [68,79,84]. There were, however, differences in the perceived rise of work demand for clinicians and the actual manifestation of raised workloads reported in studies, with clinicians reporting little additional time [70] or volume of messaging from patients [51].

E-consultations may also impact on the patient-clinician relationship in terms of changing the quality of the communication [51,63,73,79,83,84], either by fostering an enhanced sense of trust or transparency in communication [51,79] or highlighting communication deficiencies regarding the interpersonal skills needed to manage Web-based interactions [69,75].

Governance and Safety

The review highlights the lack of evidence or guidance about any rules of engagement for technology consultations and the challenges this presents to patient safety [66,74,76-78,83,85]. An appropriate consultative discussion to clarify *terms and conditions* and guidance may enhance professionals' confidence in using these systems and positively impact on implementation and sustainability of e-consultation.

Further research is also needed to explore the value and perceived benefit of care provision beyond core working hours (8 am to 6.30 pm, Monday to Friday). Expectations of timeliness arising from this review may lead to pressures in other areas of the health care system, such as secondary care services (accident and emergency providers). Despite the challenges of providing comprehensive care coverage to meet changing demographics and health care demands, early research does suggest the need to manage and deliver care outside of traditional infrastructures [86].

Consideration also needs to be given to quality and safety concerns, especially in relation to the accuracy of e-consultations diagnoses, or whether differences emerge in the quality and safety of prescribing (face-to-face vs e-consultation), including by whom—physician or advanced practitioner [87,88].

Willingness to Adopt and Sustainability

Finally, identifying possible organizational or technological issues related to the implementation of e-consultations found little evidence of studies being sustainable in the longer term (up to 1 year) [40,52]. Therefore, consideration needs to be given to whether these systems are only useful at specific time points in the patient journey, for example, newly diagnosed patients with specific conditions, or whether e-consultations could be more broadly applied across conditions. Indeed, studies into a willingness to pay were also underrepresented [61], and caution is reported in other studies suggesting the need to adequately fund organizations before establishing video consultation as routine in general practice [49]. This perhaps suggests a need for further research, to capture longer term

XSL•FO
RenderX

economic data related to e-consultation, an important consideration for any provider considering implementation [40,48,89]. Adopting e-consultations may also enable greater communication between clinicians [71], across specialist and primary care [73,78], and a broader range of geographical urban and rural areas [33,71,82].

Strengths and Limitations

In a fast-moving field, it is impossible for reviews to always include the latest developments, and some of these may be commercialized without publication. In addition, we faced the challenge of appraising if recent studies carried out in outpatient clinics are relevant to primary care [90-94]. Finally, in conducting this review, we also appreciate there are some technology and infrastructure differences between the countries, including limitations in using emails to communicate with patients. This may also have limited the reporting of results, especially if some studies were not translatable into English.

Conclusions

E-consultations are intended to address the growing demand for care from general practice. Policies and new funding opportunities that support innovative ways of care delivery may encourage a cultural shift in how patients interact with professionals and manage their own care, while also shaping the way primary care professionals use and manage technology in their practice to provide safe and efficient care.

There are 3 key messages identified from this review which may be considered important in the future developments of e-consultations. First, the review provides some insight into who, why, and when specific patient groups may be disproportionally disadvantaged or advantaged by using Web-based systems. Second, consideration needs to be given to providing a better understanding of patients' views about privacy and security of their data, so patient privacy and confidentiality are ensured. This may include exploring patients' views across different health conditions or time points, as perceived seriousness of their conditions is one key factor influencing willingness to consult electronically. Finally, issues impacting on professional's use of and perceptions of e-consultation may also be a limiting factor in terms of adoption. Fears of extra workload, expectations of quick response time, insufficient guidelines or training about the *rules of online engagement*, and effective communication strategies were all factors impacting on use.

Our review suggests that e-consultations may improve aspects of care delivery, but there remains uncertainty about which potential users to target. Improved e-implementation is a high priority, as well as the further work needed to develop innovations which support equitable primary care access and delivery.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Resubmission - Example Search String.
[DOCX File , 14 KB - medinform_v7i4e13042_app1.docx]

Multimedia Appendix 2
Resubmission - Inclusion and Exclusions.
[DOCX File , 27 KB - medinform_v7i4e13042_app2.docx]

Multimedia Appendix 3
Resubmission - MMAT Quality Appraisal.
[DOCX File , 25 KB - medinform_v7i4e13042_app3.docx]

Multimedia Appendix 4
Resubmission - Evidence Tables.
[DOCX File , 106 KB - medinform_v7i4e13042_app4.docx]

References

1. Steventon A, Bardsley M, Billings J, Dixon J, Doll H, Hirani S, Whole System Demonstrator Evaluation Team. Effect of telehealth on use of secondary care and mortality: findings from the Whole System Demonstrator cluster randomised trial. Br Med J 2012 Jun 21;344:e3874 [FREE Full text] [doi: 10.1136/bmj.e3874] [Medline: 22723612]
2. Greenhalgh T, Vijayaraghavan S, Wherton J, Shaw S, Byrne E, Campbell-Richards D, et al. Virtual online consultations: advantages and limitations (VOCAL) study. BMJ Open 2016 Jan 29;6(1):e009388 [FREE Full text] [doi: 10.1136/bmjopen-2015-009388] [Medline: 26826147]
3. The Royal College of General Practitioners. Future Vision - Case Studies URL: https://www.rcgp.org.uk/policy/future-vision/case-studies.aspx [accessed 2019-10-14]

XSL•FO
RenderX

4. Armfield NR, Gray LC, Smith AC. Clinical use of Skype: a review of the evidence base. J Telemed Telecare 2012 Apr;18(3):125-127. [doi: 10.1258/jtt.2012.SFT101] [Medline: 22362829]
5. Barlow J, Hendy J, Chrysanthaki T. Scaling-up remote care in the United Kingdom: lessons from a decade of policy intervention. In: Glascock A, Kutzik DM, editors. Essential Lessons for the Success of Telehomecare - Why It's not Plug and Play. Amsterdam: OS Press; 2012.
6. National Audit Office. 2013. Emergency admissions to hospital: managing the demand URL: https://www.nao.org.uk/wp-content/uploads/2013/10/10288-001-Emergency-admissions.pdf [accessed 2017-08-01]
7. Steventon A, Bardsley M, Billings J, Dixon J, Doll H, Beynon M, et al. Effect of telecare on use of health and social care services: findings from the Whole Systems Demonstrator cluster randomised trial. Age Ageing 2013 Jul;42(4):501-508 [FREE Full text] [doi: 10.1093/ageing/aft008] [Medline: 23443509]
8. Henderson C, Knapp M, Fernández JL, Beecham J, Hirani SP, Cartwright M, Whole System Demonstrator evaluation team. Cost effectiveness of telehealth for patients with long term conditions (Whole Systems Demonstrator telehealth questionnaire study): nested economic evaluation in a pragmatic, cluster randomised controlled trial. Br Med J 2013 Mar 20;346:f1035 [FREE Full text] [doi: 10.1136/bmj.f1035] [Medline: 23520339]
9. Bower P, Cartwright M, Hirani SP, Barlow J, Hendy J, Knapp M, et al. A comprehensive evaluation of the impact of telemonitoring in patients with long-term conditions and social care needs: protocol for the whole systems demonstrator cluster randomised trial. BMC Health Serv Res 2011 Aug 5;11:184 [FREE Full text] [doi: 10.1186/1472-6963-11-184] [Medline: 21819569]
10. Michaud TL, Zhou J, McCarthy MA, Siahpush M, Su D. Costs of home-based telemedicine programs: a systematic review. Int J Technol Assess Health Care 2018 Jan;34(4):410-418. [doi: 10.1017/S0266462318000454] [Medline: 30058505]
11. Batsis JA, DiMilia PR, Seo LM, Fortuna KL, Kennedy MA, Blunt HB, et al. Effectiveness of ambulatory telemedicine care in older adults: a systematic review. J Am Geriatr Soc 2019 Aug;67(8):1737-1749. [doi: 10.1111/jgs.15959] [Medline: 31066916]
12. Kruse CS, Soma M, Pulluri D, Nemali NT, Brooks M. The effectiveness of telemedicine in the management of chronic heart disease - a systematic review. JRSM Open 2017 Mar;8(3):2054270416681747 [FREE Full text] [doi: 10.1177/2054270416681747] [Medline: 28321319]
13. Giordano R, Clark M, Goodwin N. The King's Fund. 2011. Perspectives on Telehealth and Telecare. Learning from the 12 Whole System Demonstrator Action Network (WSDAN) URL: https://www.kingsfund.org.uk/sites/default/files/field/field_publication_file/perspectives-telehealth-telecare-wsdan-paper-nov11.pdf [accessed 2017-11-14]
14. Hanlon P, Daines L, Campbell C, McKinstry B, Weller D, Pinnock H. Telehealth interventions to support self-management of long-term conditions: a systematic metareview of diabetes, heart failure, asthma, chronic obstructive pulmonary disease, and cancer. J Med Internet Res 2017 May 17;19(5):e172 [FREE Full text] [doi: 10.2196/jmir.6688] [Medline: 28526671]
15. Vimalananda VG, Gupte G, Seraj SM, Orlander J, Berlowitz D, Fincke BG, et al. Electronic consultations (e-consults) to improve access to specialty care: a systematic review and narrative synthesis. J Telemed Telecare 2015 Sep;21(6):323-330 [FREE Full text] [doi: 10.1177/1357633X15582108] [Medline: 25995331]
16. Hendy J, Chrysanthaki T, Barlow J, Knapp M, Rogers A, Sanders C, et al. An organisational analysis of the implementation of telecare and telehealth: the whole systems demonstrator. BMC Health Serv Res 2012 Nov 15;12:403 [FREE Full text] [doi: 10.1186/1472-6963-12-403] [Medline: 23153014]
17. Mold F, Raleigh M, Alharbi NS, de Lusignan S. The impact of patient online access to computerized medical records and services on type 2 diabetes: systematic review. J Med Internet Res 2018 Jul 6;20(7):e235 [FREE Full text] [doi: 10.2196/jmir.7858] [Medline: 29980499]
18. Sifferlin A. The doctor will Skype you now. Telemedicine apps aim to replace nonemergency visits. Time 2014 Jan 13;183(1):12. [Medline: 24640400]
19. de Lusignan S, Mold F, Sheikh A, Majeed A, Wyatt JC, Quinn T, et al. Patients' online access to their electronic health records and linked online services: a systematic interpretative review. BMJ Open 2014 Sep 8;4(9):e006021 [FREE Full text] [doi: 10.1136/bmjopen-2014-006021] [Medline: 25200561]
20. Atherton H, Sawmynaden P, Sheikh A, Majeed A, Car J. Email for clinical communication between patients/caregivers and healthcare professionals. Cochrane Database Syst Rev 2012 Nov 14;11:CD007978. [doi: 10.1002/14651858.CD007978.pub2] [Medline: 23152249]
21. PRISMA Statement. URL: http://www.prisma-statement.org/ [accessed 2014-11-14]
22. Stillwell SB, Fineout-Overholt E, Melnyk BM, Williamson KM. Evidence-based practice, step by step: asking the clinical question: a key step in evidence-based practice. Am J Nurs 2010 Mar;110(3):58-61. [doi: 10.1097/01.NAJ.0000368959.11129.79] [Medline: 20179464]
23. Higgins JP, Deeks JJ. Chapter 7: Selecting studies and collecting data. In: Higgins JP, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1. London: The Cochrane Collaboration; 2008.
24. Mold F, Forbes A. Patients' and professionals' experiences and perspectives of obesity in health-care settings: a synthesis of current research. Health Expect 2013 Jun;16(2):119-142 [FREE Full text] [doi: 10.1111/j.1369-7625.2011.00699.x] [Medline: 21645186]

XSL•FO
RenderX

25. Mold F, Ellis B, de Lusignan S, Sheikh A, Wyatt JC, Cavill M, et al. The provision and impact of online patient access to their electronic health records (EHR) and transactional services on the quality and safety of health care: systematic review protocol. Inform Prim Care 2012;20(4):271-282 [FREE Full text] [Medline: 23890339]
26. Mayring P. Forum: Qualitative Social Research. 2000. Qualitative Content Analysis URL: http://www.qualitative-research.net/index.php/fqs/article/view/1089/2386 [accessed 2014-11-12]
27. Pace R, Pluye P, Bartlett G, Macaulay AC, Salsberg J, Jagosh J, et al. Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. Int J Nurs Stud 2012 Jan;49(1):47-53. [doi: 10.1016/j.ijnurstu.2011.07.002] [Medline: 21835406]
28. Pluye P, Gagnon M, Griffiths F, Johnson-Lafleur J. A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in Mixed Studies Reviews. Int J Nurs Stud 2009 Apr;46(4):529-546. [doi: 10.1016/j.ijnurstu.2009.01.009] [Medline: 19233357]
29. Brunett PH, DiPiero A, Flores C, Choi D, Kum H, Girard DE. Use of a voice and video internet technology as an alternative to in-person urgent care clinic visits. J Telemed Telecare 2015 Jun;21(4):219-226. [doi: 10.1177/1357633X15571649] [Medline: 25697491]
30. Adamson SC, Bachman JW. Pilot study of providing online care in a primary care setting. Mayo Clin Proc 2010 Aug;85(8):704-710 [FREE Full text] [doi: 10.4065/mcp.2010.0145] [Medline: 20516427]
31. Adler KG. Web portals in primary care: an evaluation of patient readiness and willingness to pay for online services. J Med Internet Res 2006 Oct 26;8(4):e26 [FREE Full text] [doi: 10.2196/jmir.8.4.e26] [Medline: 17213045]
32. Brown-Connolly NE. Patient satisfaction with telemedical access to specialty services in rural California. J Telemed Telecare 2002;8(Suppl 2):7-10. [doi: 10.1177/1357633X020080S204] [Medline: 12217115]
33. Cooper CP, Gelb CA, Rim SH, Hawkins NA, Rodriguez JL, Polonec L. Physicians who use social media and other internet-based communication technologies. J Am Med Inform Assoc 2012;19(6):960-964. [doi: 10.1136/amiajnl-2011-000628] [Medline: 22634078]
34. Couchman GR, Forjuoh SN, Rascoe TG, Reis MD, Koehler B, Walsum KLV. E-mail communications in primary care: what are patients' expectations for specific test results? Int J Med Inform 2005 Jan;74(1):21-30. [doi: 10.1016/j.ijmedinf.2004.08.005] [Medline: 15626633]
35. Denberg TD, Ross SE, Steiner JF. Patient acceptance of a novel preventive care delivery system. Prev Med 2007 Jun;44(6):543-546. [doi: 10.1016/j.ypmed.2007.01.010] [Medline: 17321583]
36. Elliott J, Chapman J, Clark DJ. Videoconferencing for a veteran's pain management follow-up clinic. Pain Manag Nurs 2007 Mar;8(1):35-46. [doi: 10.1016/j.pmn.2006.12.005] [Medline: 17336868]
37. Kittler AF, Carlson GL, Harris C, Lippincott M, Pizziferri L, Volk LA, et al. Primary care physician attitudes towards using a secure web-based portal designed to facilitate electronic communication with patients. Inform Prim Care 2004;12(3):129-138 [FREE Full text] [doi: 10.14236/jhi.v12i3.118] [Medline: 15606985]
38. Mehrotra A, Paone S, Martich GD, Albert SM, Shevchik GJ. Characteristics of patients who seek care via eVisits instead of office visits. Telemed J E Health 2013 Jul;19(7):515-519 [FREE Full text] [doi: 10.1089/tmj.2012.0221] [Medline: 23682589]
39. North F, Crane SJ, Stroebel RJ, Cha SS, Edell ES, Tulledge-Scheitel SM. Patient-generated secure messages and eVisits on a patient portal: are patients at risk? J Am Med Inform Assoc 2013;20(6):1143-1149 [FREE Full text] [doi: 10.1136/amiajnl-2012-001208] [Medline: 23703826]
40. North F, Crane SJ, Chaudhry R, Ebbert JO, Ytterberg K, Tulledge-Scheitel SM, et al. Impact of patient portal secure messages and electronic visits on adult primary care office visits. Telemed J E Health 2014 Mar;20(3):192-198 [FREE Full text] [doi: 10.1089/tmj.2013.0097] [Medline: 24350803]
41. Polinski JM, Barker T, Gagliano N, Sussman A, Brennan TA, Shrank WH. Patients' satisfaction with and preference for telehealth visits. J Gen Intern Med 2016 Mar;31(3):269-275 [FREE Full text] [doi: 10.1007/s11606-015-3489-x] [Medline: 26269131]
42. Rohrer JE, North F, Angstman KB, Oberhelman SS, Meunier MR. Timely response to secure messages from primary care patients. Qual Manag Health Care 2013;22(2):161-166. [doi: 10.1097/QMH.0b013e31828be314] [Medline: 23542371]
43. Wakefield DS, Kruse RL, Wakefield BJ, Koopman RJ, Keplinger LE, Canfield SM, et al. Consistency of patient preferences about a secure internet-based patient communications portal: contemplating, enrolling, and using. Am J Med Qual 2012;27(6):494-502. [doi: 10.1177/1062860611436246] [Medline: 22517909]
44. Grayston J, Fairhurst K, McKinstry B. Using new technologies to deliver test results in primary care: structured interview study of patients' views. Primary Health Care 2009;11(02):142-154 [FREE Full text] [doi: 10.1017/s146342360999034x]
45. Pagliari C, Donnan P, Morrison J, Ricketts I, Gregor P, Sullivan F. Adoption and perception of electronic clinical communications in Scotland. Inform Prim Care 2005;13(2):97-104 [FREE Full text] [Medline: 15992494]
46. Richards H, King G, Reid M, Selvaraj S, McNicol I, Brebner E, et al. Remote working: survey of attitudes to eHealth of doctors and nurses in rural general practices in the United Kingdom. Fam Pract 2005 Feb;22(1):2-7. [doi: 10.1093/fampra/cmh716] [Medline: 15642724]

XSL•FO
RenderX

47. Umefjord G, Malker H, Olofsson N, Hensjö LO, Petersson G. Primary care physicians' experiences of carrying out consultations on the internet. Inform Prim Care 2004;12(2):85-90 [FREE Full text] [doi: 10.14236/jhi.v12i2.112] [Medline: 15319060]
48. Umefjord G, Hamberg K, Malker H, Petersson G. The use of an Internet-based Ask the Doctor Service involving family physicians: evaluation by a web survey. Fam Pract 2006 Apr;23(2):159-166. [doi: 10.1093/fampra/cmi117] [Medline: 16464871]
49. Jiwa M, Meng X. Video consultation use by Australian general practitioners: video vignette study. J Med Internet Res 2013 Jun 19;15(6):e117 [FREE Full text] [doi: 10.2196/jmir.2638] [Medline: 23782753]
50. Zanaboni P, Scalvini S, Bernocchi P, Borghi G, Tridico C, Masella C. Teleconsultation service to improve healthcare in rural areas: acceptance, organizational impact and appropriateness. BMC Health Serv Res 2009 Dec 18;9:238 [FREE Full text] [doi: 10.1186/1472-6963-9-238] [Medline: 20021651]
51. Delbanco T, Walker J, Bell SK, Darer JD, Elmore JG, Farag N, et al. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. Ann Intern Med 2012 Oct 2;157(7):461-470 [FREE Full text] [doi: 10.7326/0003-4819-157-7-201210020-00002] [Medline: 23027317]
52. Riippa I, Linna M, Rönkkö I. A patient portal with electronic messaging: controlled before-and-after study. J Med Internet Res 2015 Nov 9;17(11):e250 [FREE Full text] [doi: 10.2196/jmir.4487] [Medline: 26553595]
53. Granlund H, Thoden C, Carlson C, Harno K. Realtime teleconsultations versus face-to-face consultations in dermatology: immediate and six-month outcome. J Telemed Telecare 2003;9(4):204-209. [doi: 10.1258/135763303322225526] [Medline: 12952690]
54. Hanson D, Calhoun J, Smith D. Changes in provider attitudes toward telemedicine. Telemed J E Health 2009 Jan;15(1):39-43. [doi: 10.1089/tmj.2008.0052] [Medline: 19199846]
55. Ralston JD, Rutter CM, Carrell D, Hecht J, Rubanowice D, Simon GE. Patient use of secure electronic messaging within a shared medical record: a cross-sectional study. J Gen Intern Med 2009 Mar;24(3):349-355 [FREE Full text] [doi: 10.1007/s11606-008-0899-z] [Medline: 19137379]
56. Shimada SL, Hogan TP, Rao SR, Allison JJ, Quill AL, Feng H, et al. Patient-provider secure messaging in VA: variations in adoption and association with urgent care utilization. Med Care 2013 Mar;51(3 Suppl 1):S21-S28. [doi: 10.1097/MLR.0b013e3182780917] [Medline: 23407007]
57. Nilsson M, Rasmark U, Nordgren H, Hallberg P, Skönevik J, Westman G, et al. The physician at a distance: the use of videoconferencing in the treatment of patients with hypertension. J Telemed Telecare 2009;15(8):397-403. [doi: 10.1258/jtt.2009.090509] [Medline: 19948706]
58. Angstman KB, Rohrer JE, Adamson SC, Chaudhry R. Impact of e-consults on return visits of primary care patients. Health Care Manag (Frederick) 2009;28(3):253-257. [doi: 10.1097/HCM.0b013e3181b3efa3] [Medline: 19668067]
59. Dixon RF, Stahl JE. Virtual visits in a general medicine practice: a pilot study. Telemed J E Health 2008 Aug;14(6):525-530. [doi: 10.1089/tmj.2007.0101] [Medline: 18729750]
60. Palen TE, Price D, Shetterly S, Wallace KB. Comparing virtual consults to traditional consults using an electronic health record: an observational case-control study. BMC Med Inform Decis Mak 2012 Jul 8;12:65 [FREE Full text] [doi: 10.1186/1472-6947-12-65] [Medline: 22769592]
61. Jacklin PB, Roberts JA, Wallace P, Haines A, Harrison R, Barber JA, Virtual Outreach Project Group. Virtual outreach: economic evaluation of joint teleconsultations for patients referred by their general practitioner for a specialist opinion. Br Med J 2003 Jul 12;327(7406):84 [FREE Full text] [doi: 10.1136/bmj.327.7406.84] [Medline: 12855528]
62. Wallace P, Barber J, Clayton W, Currell R, Fleming K, Garner P, et al. Virtual outreach: a randomised controlled trial and economic evaluation of joint teleconferenced medical consultations. Health Technol Assess 2004 Dec;8(50):1-106, iii [FREE Full text] [doi: 10.3310/hta8500] [Medline: 15546515]
63. Mirsky JB, Tieu L, Lyles C, Sarkar U. A mixed-methods study of patient-provider e-mail content in a safety-net setting. J Health Commun 2016;21(1):85-91 [FREE Full text] [doi: 10.1080/10810730.2015.1033118] [Medline: 26332306]
64. Roter DL, Larson S, Sands DZ, Ford DE, Houston T. Can e-mail messages between patients and physicians be patient-centered? Health Commun 2008;23(1):80-86. [doi: 10.1080/10410230701807295] [Medline: 18443995]
65. Baldwin LP, Clarke M, Jones R. Clinical ICT systems: augmenting case management. J Manag Med 2002;16(2-3):188-198. [doi: 10.1108/02689230210434925] [Medline: 12211344]
66. Neville RG, Marsden W, McCowan C, Pagliari C, Mullen H, Fannin A. Email consultations in general practice. Inform Prim Care 2004;12(4):207-214 [FREE Full text] [Medline: 15808022]
67. Atherton H, Pappas Y, Heneghan C, Murray E. Experiences of using email for general practice consultations: a qualitative study. Br J Gen Pract 2013 Nov;63(616):e760-e767 [FREE Full text] [doi: 10.3399/bjgp13X674440] [Medline: 24267859]
68. Hanna L, May C, Fairhurst K. The place of information and communication technology-mediated consultations in primary care: GPs' perspectives. Fam Pract 2012 Jun;29(3):361-366. [doi: 10.1093/fampra/cmr087] [Medline: 22006040]
69. Harrison R, Macfarlane A, Murray E, Wallace P. Patients' perceptions of joint teleconsultations: a qualitative evaluation. Health Expect 2006 Mar;9(1):81-90 [FREE Full text] [doi: 10.1111/j.1369-7625.2006.00368.x] [Medline: 16436164]

XSL•FO
RenderX

70. Bishop TF, Press MJ, Mendelsohn JL, Casalino LP. Electronic communication improves access, but barriers to its widespread adoption remain. Health Aff (Millwood) 2013 Aug;32(8):1361-1367 [FREE Full text] [doi: 10.1377/hlthaff.2012.1151] [Medline: 23918479]
71. Davis MM, Currey JM, Howk S, DeSordi MR, Boise L, Fagnan LJ, et al. A qualitative study of rural primary care clinician views on remote monitoring technologies. J Rural Health 2014;30(1):69-78 [FREE Full text] [doi: 10.1111/jrh.12027] [Medline: 24383486]
72. Schattner P, Mathews M, Pinskier N. Promoting e-communication-lessons from a feasibility study. Aust Fam Physician 2008 Mar;37(3):185-188. [Medline: 18345372]
73. Sevean P, Dampier S, Spadoni M, Strickland S, Pilatzke S. Patients and families experiences with video telehealth in rural/remote communities in Northern Canada. J Clin Nurs 2009 Sep;18(18):2573-2579. [doi: 10.1111/j.1365-2702.2008.02427.x] [Medline: 19694885]
74. Hansen CS, Christensen KL, Ertmann R. Patients and general practitioners have different approaches to e-mail consultations. Dan Med J 2014 Jun;61(6):A4863. [Medline: 24947631]
75. Torppa MA, Timonen O, Keinänen-Kiukaanniemi S, Larivaara P, Leiman M. Patient-nurse-doctor interaction in general practice teleconsultations--a qualitative analysis. J Telemed Telecare 2006;12(6):306-310. [doi: 10.1258/135763306778558196] [Medline: 17022839]
76. Greenhalgh T, Hinder S, Stramer K, Bratan T, Russell J. Adoption, non-adoption, and abandonment of a personal electronic health record: case study of HealthSpace. Br Med J 2010 Nov 16;341:c5814 [FREE Full text] [doi: 10.1136/bmj.c5814] [Medline: 21081595]
77. Hanna L, May C, Fairhurst K. Non-face-to-face consultations and communications in primary care: the role and perspective of general practice managers in Scotland. Inform Prim Care 2011;19(1):17-24 [FREE Full text] [doi: 10.14236/jhi.v19i1.789] [Medline: 22118332]
78. Liddy C, Rowan MS, Afkham A, Maranger J, Keely E. Building access to specialist care through e-consultation. Open Med 2013;7(1):e1-e8 [FREE Full text] [Medline: 23687533]
79. Popeski N, McKeen C, Khokhar B, Edwards A, Ghali WA, Sargious P, et al. Perceived barriers to and facilitators of patient-to-provider e-mail in the management of diabetes care. Can J Diabetes 2015 Dec;39(6):478-483. [doi: 10.1016/j.jcjd.2015.07.001] [Medline: 26409770]
80. Albert SM, Shevchik GJ, Paone S, Martich GD. Internet-based medical visit and diagnosis for common medical problems: experience of first user cohort. Telemed J E Health 2011 May;17(4):304-308 [FREE Full text] [doi: 10.1089/tmj.2010.0156] [Medline: 21457013]
81. Padman R, Shevchik G, Paone S, Dolezal C, Cervenak J. eVisit: a pilot study of a new kind of healthcare delivery. Stud Health Technol Inform 2010;160(Pt 1):262-266. [doi: 10.3233/978-1-60750-588-4-262] [Medline: 20841690]
82. Hickson R, Talbert J, Thornbury WC, Perin NR, Goodin AJ. Online medical care: the current state of 'eVisits' in acute primary care delivery. Telemed J E Health 2015 Feb;21(2):90-96. [doi: 10.1089/tmj.2014.0022] [Medline: 25474083]
83. Ye J, Rust G, Fry-Johnson Y, Strothers H. E-mail in patient-provider communication: a systematic review. Patient Educ Couns 2010 Aug;80(2):266-273 [FREE Full text] [doi: 10.1016/j.pec.2009.09.038] [Medline: 19914022]
84. Caffery LJ, Smith AC. A literature review of email-based telemedicine. Stud Health Technol Inform 2010;161:20-34. [doi: 10.3233/978-1-60750-659-1-20] [Medline: 21191155]
85. Pinnock H, Sheikh A. Standards for reporting implementation studies (StaRI): enhancing reporting to improve care. NPJ Prim Care Respir Med 2017 Jun 26;27(1):42 [FREE Full text] [doi: 10.1038/s41533-017-0045-7] [Medline: 28652602]
86. Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O. Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Q 2004;82(4):581-629 [FREE Full text] [doi: 10.1111/j.0887-378X.2004.00325.x] [Medline: 15595944]
87. Mehrotra A, Paone S, Martich GD, Albert SM, Shevchik GJ. A comparison of care at e-visits and physician office visits for sinusitis and urinary tract infection. JAMA Intern Med 2013 Jan 14;173(1):72-74 [FREE Full text] [doi: 10.1001/2013.jamainternmed.305] [Medline: 23403816]
88. Bellon JE, Stevans JM, Cohen SM, James AE, Reynolds B, Zhang Y. Comparing advanced practice providers and physicians as providers of e-visits. Telemed J E Health 2015 Dec;21(12):1019-1026. [doi: 10.1089/tmj.2014.0248] [Medline: 26161623]
89. Edwards HB, Marques E, Hollingworth W, Horwood J, Farr M, Bernard E, et al. Use of a primary care online consultation system, by whom, when and why: evaluation of a pilot observational study in 36 general practices in South West England. BMJ Open 2017 Nov 22;7(11):e016901 [FREE Full text] [doi: 10.1136/bmjopen-2017-016901] [Medline: 29167106]
90. NHS England. 2016. General Practice Forward View (GPFV) URL: https://www.england.nhs.uk/publication/general-practice-forward-view-gpfv/ [accessed 2018-01-03]
91. NHS England. 2016. GP Online Services URL: https://www.england.nhs.uk/gp-online-services/ [accessed 2018-01-03]
92. Greenhalgh T, Shaw S, Wherton J, Vijayaraghavan S, Morris J, Bhattacharya S, et al. Real-world implementation of video outpatient consultations at macro, meso, and micro levels: mixed-method study. J Med Internet Res 2018 Apr 17;20(4):e150 [FREE Full text] [doi: 10.2196/jmir.9897] [Medline: 29625956]

XSL•FO
RenderX

93. Sturesson L, Groth K. Effects of the digital transformation: qualitative study on the disturbances and limitations of using video visits in outpatient care. J Med Internet Res 2018 Jun 27;20(6):e221 [FREE Full text] [doi: 10.2196/jmir.9866] [Medline: 29950290]
94. Hansen AH, Broz J, Claudi T, Årsand E. Relations between the use of electronic health and the use of general practitioner and somatic specialist visits in patients with type 1 diabetes: cross-sectional study. J Med Internet Res 2018 Nov 7;20(11):e11322 [FREE Full text] [doi: 10.2196/11322] [Medline: 30404766]

Abbreviations

DEF: data extraction form
e-consultation: electronic consultation
GPs: general practitioners
MMAT: Mixed Methods Appraisal Tool
OR: odds ratio
RCT: randomized controlled trial
UTI: urinary tract infection

Edited by C Lovis; submitted 06.12.18; peer-reviewed by A Alturkistani, K Groth, L Daines, N Delvaux, MS Aslam; comments to author 30.01.19; revised version received 11.06.19; accepted 07.08.19; published 03.12.19.

Please cite as:
Mold F, Hendy J, Lai YL, de Lusignan S
Electronic Consultation in Primary Care Between Providers and Patients: Systematic Review
JMIR Med Inform 2019;7(4):e13042
URL: http://medinform.jmir.org/2019/4/e13042/
doi:10.2196/13042
PMID:31793888

©Freda Mold, Jane Hendy, Yi-Ling Lai, Simon de Lusignan. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 03.12.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Identifying Opioid Use Disorder in the Emergency Department: Multi-System Electronic Health Record–Based Computable Phenotype Derivation and Validation Study

David Chartash[1], PhD; Hyung Paek[2], MSEE, MD; James D Dziura[3], MPH, PhD; Bill K Ross[4]; Daniel P Nogee[3], MD; Eric Boccio[3], MD; Cory Hines[5], MD; Aaron M Schott[5], MD; Molly M Jeffery[6,7], PhD; Mehul D Patel[5], PhD; Timothy F Platts-Mills[5], MSc, MD; Osama Ahmed[3], BSc; Cynthia Brandt[1,3], MPH, MD; Katherine Couturier[3], MD; Edward Melnick[3], MHS, MD

[1]Yale Center for Medical Informatics, Yale University School of Medicine, New Haven, CT, United States

[2]Information Technology Services, Yale New Haven Health, New Haven, CT, United States

[3]Department of Emergency Medicine, Yale University School of Medicine, New Haven, CT, United States

[4]North Carolina Translational and Clinical Sciences Institute, University of North Carolina School of Medicine, Chapel Hill, NC, United States

[5]Department of Emergency Medicine, University of North Carolina School of Medicine, Chapel Hill, NC, United States

[6]Department of Emergency Medicine, Mayo Clinic, Rochester, MN, United States

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

Corresponding Author:
Edward Melnick, MHS, MD
Department of Emergency Medicine
Yale University School of Medicine
464 Congress Ave
Suite 260
New Haven, CT, 06519
United States
Phone: 1 2037855174
Email: edward.melnick@yale.edu

Abstract

Background: Deploying accurate computable phenotypes in pragmatic trials requires a trade-off between precise and clinically sensical variable selection. In particular, evaluating the medical encounter to assess a pattern leading to clinically significant impairment or distress indicative of disease is a difficult modeling challenge for the emergency department.

Objective: This study aimed to derive and validate an electronic health record–based computable phenotype to identify emergency department patients with opioid use disorder using physician chart review as a reference standard.

Methods: A two-algorithm computable phenotype was developed and evaluated using structured clinical data across 13 emergency departments in two large health care systems. Algorithm 1 combined clinician and billing codes. Algorithm 2 used chief complaint structured data suggestive of opioid use disorder. To evaluate the algorithms in both internal and external validation phases, two emergency medicine physicians, with a third acting as adjudicator, reviewed a pragmatic sample of 231 charts: 125 internal validation (75 positive and 50 negative), 106 external validation (56 positive and 50 negative).

Results: Cohen kappa, measuring agreement between reviewers, for the internal and external validation cohorts was 0.95 and 0.93, respectively. In the internal validation phase, Algorithm 1 had a positive predictive value (PPV) of 0.96 (95% CI 0.863-0.995) and a negative predictive value (NPV) of 0.98 (95% CI 0.893-0.999), and Algorithm 2 had a PPV of 0.8 (95% CI 0.593-0.932) and an NPV of 1.0 (one-sided 97.5% CI 0.863-1). In the external validation phase, the phenotype had a PPV of 0.95 (95% CI 0.851-0.989) and an NPV of 0.92 (95% CI 0.807-0.978).

Conclusions: This phenotype detected emergency department patients with opioid use disorder with high predictive values and reliability. Its algorithms were transportable across health care systems and have potential value for both clinical and research purposes.

(JMIR Med Inform 2019;7(4):e15794) doi:10.2196/15794

XSL•FO
RenderX

KEYWORDS

electronic health records; emergency medicine; algorithms; phenotype; opioid-related disorders

Introduction

Background

In the decade since the Health Information Technology for Economic and Clinical Health Act of 2009 was enacted, US hospitals have achieved greater than 96% adoption of electronic health records (EHRs) [1]. EHRs are projected to store 2314 exabytes (1 exabyte=approximately 1 billion GB) of health data by 2020 [2]. This wealth of data has been touted as *a practically inexhaustible source of knowledge to fuel a learning health care system* [3]. Yet at this time, significant challenges remain for using clinical data for research and optimization of health care delivery [4]. Integral to addressing these challenges and studying an intervention in actual clinical care is the ability to accurately and reliably identify patients with particular diagnoses or medical conditions across heterogeneous systems [4-6]. An EHR-based computable phenotype aims to do precisely that. Henceforth, it is referred to as an EHR-based *phenotype*, defined as a set of data elements and logical expressions used to identify individuals or populations (ie, cohorts) with particular diagnoses or medical conditions via clinical characteristics, events, and service patterns that are ascertained using a computerized query of an EHR system or data repository [5,7]. Phenotypes are typically used in clinical trial recruitment to identify cohorts with specific conditions using diverse data sources [5]. They are also increasingly used to define an authoritative standard for electronic clinical quality measure reporting [8].

An estimated 2.1 million people in the United States have opioid use disorder (OUD) [9], and over 33,000 opioid-related deaths occur annually, a number projected to increase to more than 81,000 by 2025 [10,11]. From 2016 to 2017, emergency departments (EDs) experienced a 30% increase in visits for opioid overdose [12]. Buprenorphine, a partial opioid agonist generally combined with an antagonist (naloxone), is an effective treatment for OUD that decreases mortality (from approximately 5% to 3% annually following an ED visit for opioid overdose), withdrawal symptoms, craving, and opioid use [13-15]. Initiating buprenorphine in the ED doubles the rate of addiction treatment engagement in ED patients with OUD [16]. However, ED-initiated buprenorphine has not yet been adopted into routine emergency care [17,18].

Objectives

Phenotyping could be used as a clinical tool to identify patients likely to benefit from ED-initiated buprenorphine or other interventions and as a research tool to identify patients who should be included in large-scale intervention studies of OUD interventions. We will conduct a multi-system pragmatic trial of user-centered clinical decision support to implement EMergency department-initiated BuprenorphinE for opioid use Disorder (EMBED) across 20 EDs in 5 health care systems [19]. EHR phenotyping will allow pragmatic comparison of the effectiveness of the EMBED intervention to usual care on outcomes in ED patients with OUD in the upcoming EMBED trial (primary outcome—adoption of ED-initiated buprenorphine in routine emergency care). Our objective in this study was to derive and validate an EHR-based computable phenotype to identify ED patients with OUD using structured data; physician validation based on chart review was used as the reference standard. This phenotype will be used to inform patient identification and data collection for the subsequent EMBED pragmatic trial.

Methods

Study Setting and Sample

This phenotype was created for the purposes of identifying patients with OUD who could benefit from ED-initiation of buprenorphine in a subsequent trial or quality improvement initiatives. Therefore, the phenotype only included ED patients who were discharged from the hospital (ie, not admitted as inpatients), were not currently prescribed buprenorphine, methadone, or naltrexone as medication treatment for OUD, and were not pregnant (as buprenorphine with naloxone may not be safe for pregnant women and its use requires more expertise than clinical decision support). This study was performed within the XXXX Health System in YYYY and XXXX Health System in YYYY by identifying a cohort of adults (>18 years of age) with ED encounters between November 1, 2017, and October 31, 2018, in the EHR. The 2 health care systems use different billing companies, but the same EHR vendor (Epic; Epic Systems Corporation). Data were extracted from the EHR of each hospital using local Epic Clarity databases (Epic; Epic Systems Corporation). These data comprised information available within the EHR on the date of service of the ED visit in question. Approval for this study was provided by the Institutional Review Boards of the respective institutions (Protocol IDs 2000022749 [internal validation] and 18-2653 [external validation]).

Clinical Definition of Opioid Use Disorder

Although psychiatric evaluation is the gold standard for diagnosing OUD, within the emergency medicine (EM) context, diagnosis if performed is based on the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5) criteria [20]. The DSM-5 specifies 11 criteria for the diagnosis of OUD, with qualifiers for remission [21]. It specifies that OUD consists of "A problematic pattern of opioid use leading to clinically significant impairment or distress, as manifested by at least two of [...eleven criteria], occurring within a 12-month period." These criteria include opioids taken repeatedly, continuously, and in larger amounts over a longer period than was intended, resulting in sequelae such as tolerance, withdrawal, craving, desire to cut down, failure to fulfill or engage in social and role obligations (such as at work, school, or home), and continued use despite problems related to use.

Electronic Health Record Definition

The computable phenotype algorithm was developed based on data elements from available primary care OUD phenotypes [22,23] with additions and revisions based on the clinical

XSL•FO
RenderX

judgment of an EM attending physician and clinical informaticist as well as available and high-yield structured ED data elements as judged by the health system's medical director of Information Technology (HP). To maximize the yield and performance of the phenotype, 2 separate algorithms were created (Figure 1).

Algorithm 1 is a diagnostic coding–based approach to identifying patients with OUD, utilizing opioid-related International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10) diagnostic codes associated with the ED visit (as coded by a clinician or medical coder, Table 1).

Figure 1. Flow diagram of phenotypes. ED: emergency department; MOUD: medication for opioid use disorder; ETOH: ethyl alcohol; OUD: opioid use disorder.

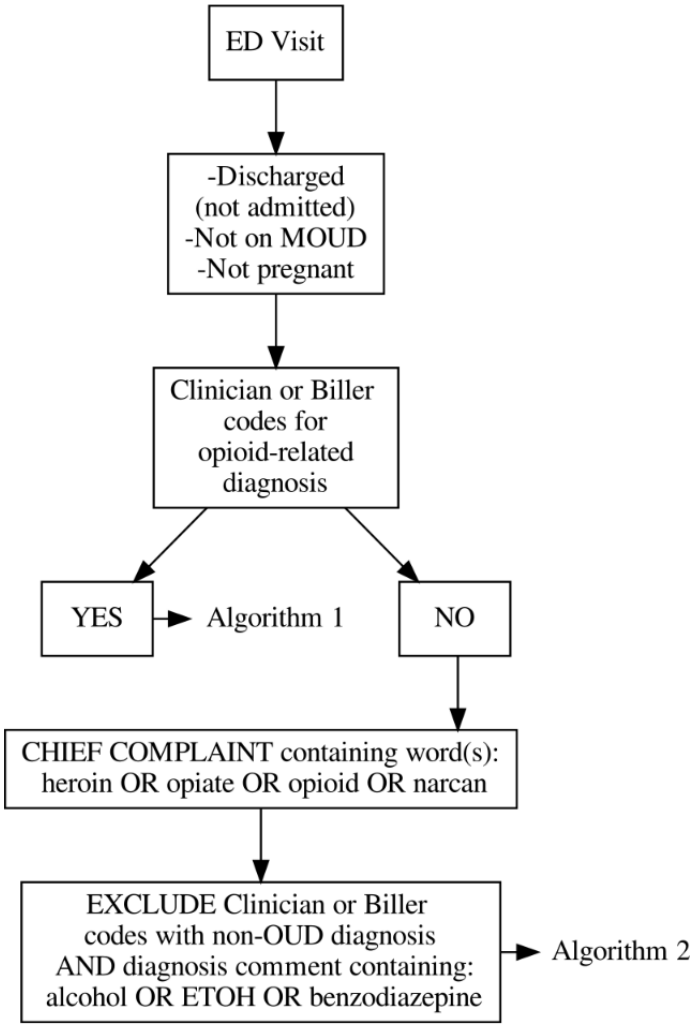


Table 1. List of International Classification of Diseases, Tenth Revision (ICD-10) codes for opioid-related diagnoses used for Algorithm 1 case detection.

ICD-10[a] code	Description
F11	Opioid-related disorders
T40.0	Poisoning by, adverse effect of, and underdosing of opium
T40.1	Poisoning by, adverse effect of, and underdosing of heroin
T40.2	Poisoning by, adverse effect of, and underdosing of other opioids
T40.3	Poisoning by, adverse effect of, and underdosing of methadone
T40.4	Poisoning by, adverse effect of, and underdosing of other synthetic narcotics
T40.6	Poisoning by, adverse effect of, and underdosing of other and unspecified narcotics

[a]ICD-10: International Classification of Diseases, Tenth Revision.

Algorithm 2 identifies patients who have not been captured by Algorithm 1 but have information in their ED chief complaint suggestive of OUD. This algorithm flagged patients if the words *heroin*, *opiate*, *opioid*, or *narcan* were included in their chief complaint for the ED visit. However, as naloxone is often used in ED patients with undifferentiated altered mental status or

XSL•FO
RenderX

overdose, patients with *narcan* in their chief complain who did not have an OUD-related final diagnosis were excluded. Upon preliminary testing of the algorithm, the 2 most frequent false-positive diagnoses were alcohol- and benzodiazepine-related. Visits with these chief complaints but alcohol- and benzodiazepine-related final diagnoses were removed by excluding patients with the words *alcohol* or *benzodiazepine* in their final ED diagnosis.

Structured Query Language Implementation of the Phenotype

After the 2 algorithms were reviewed, finalized, and approved by the investigative team, individual elements of each algorithm were converted into structured query language (SQL). The computable phenotype algorithm was written to be deployed in the Epic EHR systems across both the health systems. Data structures within Epic were mapped to each of the concepts (which were standardized across hospitals in the system) by a clinical informatics expert (HP). This eliminated the necessity of translating concepts to local codes within hospitals. Sample queries were run, and HP verified charts for accuracy. The SQL query and data dictionary (Multimedia Appendix 1, SQL file and data dictionary) were assembled by HP and reviewed for accuracy and comprehensiveness by HP, EM, and DC. The possible values of each variable are described in Figure 1 and expanded upon in Tables 1 and 2. Once all of the elements of the phenotype were codified in SQL, the algorithms were applied to the study population's ED medical records.

Table 2. Algorithm 2 case definition variables.

EHR[a] data variable	Criteria for suspected opioid use disorder
Chief complaint	• Reason for visit contains the words heroin; opiate; opioid • Reason for visit comment contains the word narcan
Diagnosis description	• Not Algorithm 1 positive, that is, does not contain ICD-10[b] codes F11 or T40. 0-40.6 listed in Table 1 • Does not contain the words alcohol, EtOH, benzodiazepine

[a]EHR: electronic health record.

[b]ICD-10: International Classification of Diseases, Tenth Revision.

Evaluation Phase 1: Internal Validation

Following the implementation of the computable phenotype algorithm, internal validation was performed using a sample of 125 charts retrieved from the XXXX health system EHR by a clinical informaticist (HP). A total of 75 charts were intended to be representative of the resulting OUD phenotypes with 50 of these charts meeting Algorithm 1 criteria and the other 25 meeting Algorithm 2 criteria. The other 50 charts were *phenotype negative* (ie, not satisfying criteria for either algorithm)**.** Charts were selected at random from the cohort with ED visits from April 10, 2018, to August 1, 2018, across the health systems and reviewed during August 2018 to October 2018 for the internal validation phase and December 2018 to January 2019 for the external validation phase. As the chart reviewers were given access to the patient's full chart, the time window for the charts was deliberately narrow to avoid postvisit information (eg, of someone who subsequently develops OUD that was not present on the date of the ED visit) confounding the accuracy of the chart review of the ED visit.

Evaluation Phase 2: External Validation

The external validation cohort was constructed by a clinical informaticist (WKR) with 20,000 randomly sampled ED visits occurring between November 1, 2017, and October 31, 2018 across the XXXXX health system. We picked this number of charts given the rate of phenotype-positive charts in the internal validation cohort with a goal of estimating sensitivity of the phenotype based on prevalence in this random sample. A total of 55 charts met Algorithm 1 criteria. Of those not positive for Algorithm 1, 1 chart met Algorithm 2 criteria. Of the remaining negative cases, a 0.25% (50/200) random sample produced 50 charts for review. Cases positive for Algorithm 1 or 2 were combined owing to the low yield of a single chart identified as Algorithm 2 positive.

Chart Review

Each chart was reviewed independently and separately by 2 EM physicians (internal validation: DPN, EB; external validation: CH, AMS) blinded to the results and the algorithms and the decision of the other reviewer. All cases of disagreement were adjudicated by a third EM physician reviewer (internal validation: KC; external validation: TFP) also blinded to the results of the algorithms and the decision of the other reviewers. Reviewers were asked to diagnose patients as *OUD-positive* or *OUD-negative* based upon a review of EHR data available up to and on the date of the ED visit (but not after the ED visit), their clinical judgment, and the DSM-5 OUD diagnostic criteria which were presented to them with each case at the time of review [21]. For cases that were categorized as *OUD-positive*, reviewers were then prompted to select at least 2 of the 11 DSM-5 criteria that informed their diagnosis.

Analysis

Phenotype performance was assessed using descriptive statistics. A standard 2×2 confusion matrix [24] was configured for analysis of the performance of each algorithm in each phase. The reference standard was the adjudicated diagnosis, whereas the test was the phenotype result. For Algorithm 1 in the internal validation phase (Table 3), the top row included the 50 phenotype-positive charts, and the bottom row included the 50 phenotype-negative charts. For Algorithm 2 in the internal validation phase (Table 3), the top row included the 25 phenotype-positive charts, and the bottom row included the 25 phenotype-negative charts. In the external validation phase, the algorithms were combined because of low incidence of

XSL•FO
RenderX

Algorithm 2-positive (Table 3), with 56 positive and 50 negative. Interrater reliability was reported using Cohen kappa. Analyses were conducted with the scikit-learn package (version 0.19.2) in Python (version 2.7.12) for internal validation and Stata (StataCorp, version 14) for external validation.

Table 3. Confusion matrices for validation phases (disease present: reference standard).

Test	Result			
	Reviewers +	Reviewers −	Predictive value	95% CI
Algorithm 1 (internal validation)				
Phenotype +	48	2	0.96[a]	0.863-0.995
Phenotype −	1	49	0.98[b]	0.893-0.999
Algorithm 2 (internal validation)				
Phenotype +	20	5	0.8[a]	0.593-0.932
Phenotype −	0	25	1.0[b]	0.863-1.000[c]
Combined phenotype (external validation)				
Phenotype +	53	3	0.95[a]	0.851-0.989
Phenotype −	4	46	0.92[b]	0.807-0.978

[a]Positive predictive value.

[b]Negative predictive value.

[c]97.5%, one-sided.

Results

Among ED visits resulting in discharge from November 1, 2017, to October 31, 2018, across the 13 EDs in the 2 health care systems, a total of 474,176 unique ED visits (discharged patients only) with an average of 36,475 ED visits per year per site were identified. A total of 2294 of these visits were phenotype-positive with an average of 176 (median 104) phenotype-positive visits per site. Site visit by volume is presented in Table 4.

XSL•FO
RenderX

Table 4. Annual volume of emergency department (ED) visits meeting phenotype criteria (November 1, 2017, to October 31, 2018, ED discharges only).

Validation	Total patients (n)	Total visits (n)	Algorithm 1 (n)	Algorithm 2 (n)
Internal				
Department				
Hospital X I	44,291	67,995	343	49
Health System X II	22,344	29,309	56	11
Health System X III	24,738	38,128	251	46
Health System X IV	27,220	44,505	324	73
Health System X V	44,780	65,837	509	70
Health System X VI	17,797	22,540	25	0
Total	181,170	268,314	1508	249
Average	30,195	44,719	251.3	41.5
External				
Department				
Health System Y I	9818	15,749	37	4
Health System Y II	15,220	25,556	91	2
Health System Y III	22,332	30,912	57	4
Health System Y IV	22,080	38,086	100	4
Health System Y V	5467	6190	24	1
Health System Y VI	34,576	46,335	98	0
Health System Y VII	32,879	43,034	110	5
Total	142,372	205,862	517	20
Average	20,339	29,409	74	3

Internal Validation Cohort

In the internal validation cohort of 125 charts, reviewers disagreed on the classification of 3 charts (agreement=97%; kappa=0.95), with the adjudicator identifying the 2 discordant Algorithm 1 cases as not having OUD and the 1 discordant Algorithm 2 case as having OUD. Algorithm 1 had a positive predictive value (PPV) of 0.96 (95% CI 0.863-0.995) and a negative predictive value (NPV) of 0.98 (95% CI 0.893-0.999; Table 3). Algorithm 2 had a PPV of 0.8 (95% CI 0.593-0.932) and an NPV of 1.0 (one-sided 97.5% CI 0.863-1; Table 3). The most frequently met current DSM-5 criteria were “opioids taken in larger amounts or over a longer period than was intended” or “recurrent use in situations in which it is physically hazardous,” whereas the least frequent criteria were those describing social dysfunction related to the use of opioids (such as “recurrent opioid use resulting in a failure to fulfill major role obligations at work, school, or home” or “important social, occupational, or recreational activities are given up or reduced because of opioid use”).

External Validation Cohort

In the external validation cohort of 106 charts, reviewers disagreed on the classification of 8 charts (agreement=92.5%; kappa=0.85). A total of 3 of the 8 discordant cases were phenotype-positive, of which the adjudicator determined 2 as having OUD. Of the 5 discordant cases that were phenotype-negative, the adjudicator identified 3 as having OUD. The combined phenotype had a PPV of 0.95 (95% CI 0.851-0.989) and an NPV of 0.92 (95% CI 0.807-0.978; Table 3). The most frequently met current DSM-5 criteria were “opioids are often taken in larger amounts or over a longer period than was intended” and “craving, or a strong desire or urge to use opioids,” whereas the least frequent criterion was “important social, occupational, or recreational activities are given up or reduced because of opioid use.”

Discussion

Principal Findings

With an externally validated PPV of 0.95 and NPV of 0.92, the combined phenotype derived and validated for this study performed remarkably well in predicting OUD in ED patients across 2 large health care systems. The strength of the phenotype’s classification performance may be because of the possibility that the algorithm and the reviewers were using similar (if not the same) information from patients’ charts.

In both the internal and external validation chart reviews, the most common DSM criterion selected by the reviewers was “opioids are often taken in larger amounts or over a longer period than was intended.” In the internal validation phase, the second most common criterion was “recurrent opioid use in situations in which it is physically hazardous,” whereas the

XSL•FO
RenderX

second most common criterion in the external validation phase was "craving, or a strong desire or urge to use opioids." Although desire and effort to cut down or control opioid use are specific diagnostic criteria, they were inconsistently applied by the reviewers. As these specific criteria are not explicitly documented in the routine emergency care, the reviewers instead had to infer which criteria to apply to cases using available documentation. In both chart review phases, the least frequently identified criteria were those describing failures in social behavior as they pertained to the use of opioids. This could be because of the fact that ED billing requirements do not require detailed documentation of social history, and the impact of opioids on social behaviors usually has limited value for assisting clinicians in making a diagnosis during emergency care [25].

Given the limitations of ED documentation, our phenotype benefited from incorporation of available structured data elements from the data dictionaries created in previous work to develop EHR phenotypes for primary care patients on chronic opioid therapy at risk for problematic opioid use [22,23]. Given the difference in populations and objectives between this study and the primary care OUD phenotype, it is difficult to compare the differences in their phenotypes' performance. In particular, the previous work focused on the performance of natural language processing for identifying risk for problematic opioid use in patients for whom differences in the signs and symptoms of OUD might be more nuanced: every patient included in that study was on chronic opioid therapy. The goal of that study was to capture the presence of OUD symptoms using free-text notes—a complex machine learning problem. In contrast, our study included a broader population (all patients presenting to the ED were eligible for inclusion in the phenotype-negative sample), and our phenotype drew on structured data elements including diagnoses and chief complaints; these structured data generally reflect the clinical judgment of people who have directly observed the patient and determined that OUD was likely.

A strength of our study compared with previous EHR phenotype work is the external validation of the phenotype's performance via chart review in the second health care system. For example, the HIV EHR phenotype developed by Paul et al [26] performs well, but its transportability and performance in outside health care systems are not known [27]. External validation is particularly important for EHR phenotypes that rely on documentation and diagnostic codes as documentation and diagnostic codes are dependent on local practice patterns by clinicians and coders, both of which could vary within and across health care systems.

The phenotype described here will be used as part of a subsequent pragmatic trial to be deployed across multiple health care systems to identify patients who may have been candidates for ED-initiated buprenorphine—these cases will form the denominator of a measure to assess what proportion of those potentially eligible actually received buprenorphine. As most patients evaluated using the phenotype will screen negative, our approach would likely result in a high number of false negatives in a true epidemiologic evaluation. As we were trying to maximize specificity for a pragmatic trial, the phenotype's classification performance will meet the trial's needs to screen patients for eligibility for ED-initiated buprenorphine with high specificity. Furthermore, the goal was not to definitively determine a diagnosis of OUD for each patient. For clinical practice, any patient identified as having OUD by this phenotype would require confirmation using an in-person assessment. As the capacity and expectation of EDs to treat OUD expands, so also does the value of an accurate EHR phenotype that could be used to identify patients who might benefit from treatment including ED-initiated buprenorphine and referral for ongoing medication treatment for OUD.

Limitations

The primary limitation of this study is the use of retrospective ED chart review as a reference standard for the diagnosis of OUD. Our chart review process was robust and included all clinical documentation up to date of the ED visit. However, a full diagnostic assessment by a psychiatrist or addiction medicine specialist would be the gold standard to establish a diagnosis of OUD. If available, it is possible that such an assessment would differ from chart review alone.

External validation in an outside health care system strengthens the evidence for the generalizability of our phenotype. The external system uses the same EHR vendor but a different billing and coding company. It is unknown how the EHR phenotype would perform in systems using other EHR platforms. In addition, transportability issues have been discovered in preliminary estimates from a third health system because of differences in structured data capture of the chief complaint. In the external validation phase, the second algorithm did not identify a substantial number of cases. This suggests that there are likely local practice patterns in documentation or coding that may have affected the transportability of this EHR phenotype [27]. Although local phenotype development and adaptation could overcome this limitation, the overall classification performance remained strong in the external validation phase. Furthermore, in the internal validation phase, the individual algorithms maintained high PPV values (0.96 and 0.8, respectively) and NPV values (0.98 and 1.0).

In future work, the efficacy of the phenotype algorithms will be tested in the EMBED trial, and the question as to whether these algorithms can function in a pragmatic ED setting will be answered. Statistically, the cohort selection and chart review performed here did not obtain a patient population reflective of the true prevalence of the disease, and as such, sensitivity and specificity calculations would not provide an accurate reflection of the phenotype's performance in a true ED population. Assessment of sensitivity for events with low base rates is inherently unreliable. A very large sample size would be necessary for a precise estimate of sensitivity. Therefore, the assessment of sensitivity in this analysis is limited. To address this limitation, we report only PPV and NPV here. Estimating sensitivity and specificity for the external validation study can be done by inflating the phenotype-negative row numbers in the external validation confusion matrix by the sampling factor (Table 3) to represent 19,944 patients who screened negative. The sampling factor of approximately 399 (19,944/50) would change this row to extrapolated values of 1596 (false negatives)

XSL•FO
RenderX

and 18,354 (true negatives). These extrapolated values would yield a sensitivity of 3.2% with an extremely wide confidence interval and specificity of 99.9%. This wide range could be explained by the false negatives that the phenotype is not identifying. Given the current opioid crisis, we know that the rates of OUD are high, and there are likely many individuals with occult OUD that is not being identified in the ED as they may be presenting with medical complaints unrelated to their OUD comorbidity. For example, abdominal pain and chest pain are the 2 most common presenting complaints to EDs nationally. There is likely a large population of ED patients with these complaints that have OUD that goes unrecognized in the ED. Future work should screen for more precise estimates of undiagnosed OUD in the ED population.

In the upcoming trial, further evaluation of the phenotype algorithms' performance will begin to address the intra- and intersite population sensitivity and specificity. To further refine the algorithm's ability to discern between true and false positives, logistic regression is planned to predict future OUD-coded diagnoses given information from previous visits, such that variables can be removed given their performance in the regression model. Future work will also attempt to quantify the rate of false negatives through extended manual review and to determine whether changes to the algorithm improve sensitivity. The long-term goal of future work is to standardize the representation of the algorithms such that they can be portable beyond Epic to other EHR vendors as well as explore additional information retrieval techniques [28]. A more comprehensive validation could establish more reliable sensitivity estimates by use of a gold standard estimate of true prevalence of OUD in the ED population by screening a representative ED population for OUD with DSM-5 diagnostic criteria [21].

Conclusions

An EHR phenotype derived and internally and externally validated for the purposes of a pragmatic trial to test the effectiveness of user-centered clinical decision support to increase the adoption of ED-initiated buprenorphine performed reliably and accurately to identify ED patients with OUD. The 2 algorithms comprising the phenotype were transportable across health care systems and have potential value for both clinical quality improvement interventions as well as research endeavors. Standardization of the phenotype will support efforts to use clinical phenotyping as an evidence-based tool at the front line of clinical practice.

Acknowledgments

Research reported in this publication was supported within the National Institutes of Health (NIH) Health Care Systems Research Collaboratory, by a cooperative agreement (UG3DA047003) from the National Institute on Drug Abuse of the NIH. This work also received logistical and technical support from the NIH Collaboratory Coordinating Center (U24AT009676). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

DC, ERM, HP, TFP, and JDD significantly contributed to the conception and design of the study. HP, WKR, TFP, MP, and DC acquired and analyzed the data. DPN, EB, KCC, CH, AMS, and TFP performed the clinical chart reviews. HP, DC, DPN, and ERM drafted the initial manuscript. All authors were involved in data interpretation, revised the manuscript, and approved the final version submitted for publication. HP, DC, WKR, and TFP had access to the data in the study and took responsibility for data integrity and accuracy. ERM took responsibility for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
SQL code to retrieve records from Epic Clarity database.
[TXT File , 71 KB - medinform_v7i4e15794_app1.txt]

References

1. Henry J, Pylypchuk Y, Searcy T, Patel V. Health IT Dashboard. 2016. Adoption of Electronic Health Record Systems among US Non-Federal Acute Care Hospitals: 2008-2015 URL: https://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php [accessed 2019-10-15]
2. Cyclone Interactive: Digital Marketing Agency. 2014. Vertical Industry Brief: Digital Universe Driving Data Growth in Healthcare Internet URL: https://www.cycloneinteractive.com/cyclone/assets/File/digital-universe-healthcare-vertical-report-ar.pdf [accessed 2015-12-29]
3. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff (Millwood) 2014 Jul;33(7):1163-1170 [FREE Full text] [doi: 10.1377/hlthaff.2014.0053] [Medline: 25006142]

XSL•FO
RenderX

4. Richesson RL, Green BB, Laws R, Puro J, Kahn MG, Bauck A, et al. Pragmatic (trial) informatics: a perspective from the NIH Health Care Systems Research Collaboratory. J Am Med Inform Assoc 2017 Sep 1;24(5):996-1001. [doi: 10.1093/jamia/ocx016] [Medline: 28340241]
5. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014;21(2):221-230 [FREE Full text] [doi: 10.1136/amiajnl-2013-001935] [Medline: 24201027]
6. Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. Br Med J 2015 May 8;350:h2147. [doi: 10.1136/bmj.h2147] [Medline: 25956159]
7. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. J Am Med Inform Assoc 2013 Dec;20(e2):e226-e231 [FREE Full text] [doi: 10.1136/amiajnl-2013-001926] [Medline: 23956018]
8. Bodenreider O, Nguyen D, Chiang P, Chuang P, Madden M, Winnenburg R, et al. The NLM value set authority center. Stud Health Technol Inform 2013;192:1224 [FREE Full text] [Medline: 23920998]
9. Ahrnsbrak R, Bose J, Hedden S, Lipari R, Park-Lee E. SAMHSA - Substance Abuse and Mental Health Services. 2017. Key Substance Use and Mental Health Indicators in the United States: Results from the 2016 National Survey on Drug Use and Health URL: https://www.samhsa.gov/data/sites/default/files/NSDUH-FFR1-2016/NSDUH-FFR1-2016.htm [accessed 2017-09-01]
10. Seth P, Scholl L, Rudd RA, Bacon S. Overdose deaths involving opioids, cocaine, and psychostimulants - United States, 2015-2016. MMWR Morb Mortal Wkly Rep 2018 Mar 30;67(12):349-358 [FREE Full text] [doi: 10.15585/mmwr.mm6712a1] [Medline: 29596405]
11. Chen Q, Larochelle MR, Weaver DT, Lietz AP, Mueller PP, Mercaldo S, et al. Prevention of prescription opioid misuse and projected overdose deaths in the United States. JAMA Netw Open 2019 Feb 1;2(2):e187621 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.7621] [Medline: 30707224]
12. Vivolo-Kantor AM, Seth P, Gladden RM, Mattson CL, Baldwin GT, Kite-Powell A, et al. Vital signs: trends in emergency department visits for suspected opioid overdoses - United States, July 2016-September 2017. MMWR Morb Mortal Wkly Rep 2018 Mar 9;67(9):279-285 [FREE Full text] [doi: 10.15585/mmwr.mm6709e1] [Medline: 29518069]
13. Kakko J, Svanborg KD, Kreek MJ, Heilig M. 1-year retention and social function after buprenorphine-assisted relapse prevention treatment for heroin dependence in Sweden: a randomised, placebo-controlled trial. Lancet 2003 Feb 22;361(9358):662-668. [doi: 10.1016/S0140-6736(03)12600-1] [Medline: 12606177]
14. Mattick RP, Breen C, Kimber J, Davoli M. Buprenorphine maintenance versus placebo or methadone maintenance for opioid dependence. Cochrane Database Syst Rev 2014 Feb 6(2):CD002207. [doi: 10.1002/14651858.CD002207.pub4] [Medline: 24500948]
15. Larochelle MR, Bernson D, Land T, Stopka TJ, Wang N, Xuan Z, et al. Medication for opioid use disorder after nonfatal opioid overdose and association with mortality: a cohort study. Ann Intern Med 2018 Aug 7;169(3):137-145 [FREE Full text] [doi: 10.7326/M17-3107] [Medline: 29913516]
16. D'Onofrio G, O'Connor P, Pantalon M, Chawarski M, Busch S, Owens P, et al. Emergency department-initiated buprenorphine/naloxone treatment for opioid dependence: a randomized clinical trial. J Am Med Assoc 2015 Apr 28;313(16):1636-1644 [FREE Full text] [doi: 10.1001/jama.2015.3474] [Medline: 25919527]
17. Duber HC, Barata IA, Cioè-Peña E, Liang SY, Ketcham E, Macias-Konstantopoulos W, et al. Identification, management, and transition of care for patients with opioid use disorder in the emergency department. Ann Emerg Med 2018 Oct;72(4):420-431 [FREE Full text] [doi: 10.1016/j.annemergmed.2018.04.007] [Medline: 29880438]
18. Martin A, Mitchell A, Wakeman S, White B, Raja A. Emergency department treatment of opioid addiction: an opportunity to lead. Acad Emerg Med 2018 May;25(5):601-604 [FREE Full text] [doi: 10.1111/acem.13367] [Medline: 29266577]
19. NIH Collaboratory Rethinking Clinical Trials. UG3 Project: Pragmatic Trial of User-Centered Clinical Decision Support to Implement EMergency Department-Initiated BuprenorphinE for Opioid Use Disorder (EMBED) - Rethinking Clinical Trials Internet URL: https://tinyurl.com/y35dz7b8 [accessed 2018-10-25]
20. Bernstein E, Bernstein J, Weiner S, D'Onofrio G. Substance use disorders. In: Tintinalli JE, Stapczynski JS, Ma OJ, Yealy DM, Meckler GD, Cline DM, editors. Tintinalli's Emergency Medicine: A Comprehensive Study Guide. Eight Edition. New York: McGraw-Hill; 2016.
21. American Psychiatric Association. Substance-related addictive disorders. In: Diagnostic and Statistical Manual of Mental Disorders. Arlington, VA: American Psychiatric Association; 2013.
22. Carrell DS, Cronkite D, Palmer RE, Saunders K, Gross DE, Masters ET, et al. Using natural language processing to identify problem usage of prescription opioids. Int J Med Inform 2015 Dec;84(12):1057-1064. [doi: 10.1016/j.ijmedinf.2015.09.002] [Medline: 26456569]
23. Palmer RE, Carrell DS, Cronkite D, Saunders K, Gross DE, Masters E, et al. The prevalence of problem opioid use in patients receiving chronic opioid therapy: computer-assisted review of electronic health record clinical notes. Pain 2015 Jul;156(7):1208-1214. [doi: 10.1097/j.pain.0000000000000145] [Medline: 25760471]
24. Hulley S, Cummings S, Browner W, Grady D, Newman T. Designing Clinical Research. Philadelphia, PA: Lippincott Williams & Wilkins; 2013.

XSL•FO
RenderX

25. Centers for Medicare and Medicaid Services. 1997. 1997 Documentation Guidelines for Evaluation and Management Services URL: https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNEdWebGuide/Downloads/97Docguidelines.pdf [accessed 2019-10-09]
26. Paul DW, Neely NB, Clement M, Riley I, Al-Hegelan M, Phelan M, et al. Development and validation of an electronic medical record (EMR)-based computed phenotype of HIV-1 infection. J Am Med Inform Assoc 2018 Feb 1;25(2):150-157 [FREE Full text] [doi: 10.1093/jamia/ocx061] [Medline: 28645207]
27. Pacheco JA, Rasmussen LV, Kiefer RC, Campion TR, Speltz P, Carroll RJ, et al. A case study evaluating the portability of an executable computable phenotype algorithm across multiple institutions and electronic health record environments. J Am Med Inform Assoc 2018 Nov 1;25(11):1540-1546 [FREE Full text] [doi: 10.1093/jamia/ocy101] [Medline: 30124903]
28. Rasmussen LV, Thompson WK, Pacheco JA, Kho AN, Carrell DS, Pathak J, et al. Design patterns for the development of electronic health record-driven phenotype extraction algorithms. J Biomed Inform 2014 Oct;51:280-286 [FREE Full text] [doi: 10.1016/j.jbi.2014.06.007] [Medline: 24960203]

Abbreviations

DSM-5: Diagnostic and Statistical Manual of Mental Disorders, 5th Edition
ED: emergency department
EHR: electronic health record
EM: emergency medicine
EMBED: EMergency department-initiated BuprenorphinE for opioid use Disorder
ICD-10: International Classification of Diseases, Tenth Revision
NIH: National Institutes of Health
NPV: negative predictive value
OUD: opioid use disorder
PPV: positive predictive value
SQL: structured query language

Edited by G Eysenbach; submitted 08.08.19; peer-reviewed by X Fan, M Graber, R Radecki, C Freiermuth, E Schoenfeld, JT Pollettini, A Follmann, KA Nguyen; comments to author 18.09.19; revised version received 27.09.19; accepted 01.10.19; published 31.10.19.

Please cite as:
Chartash D, Paek H, Dziura JD, Ross BK, Nogee DP, Boccio E, Hines C, Schott AM, Jeffery MM, Patel MD, Platts-Mills TF, Ahmed O, Brandt C, Couturier K, Melnick E
Identifying Opioid Use Disorder in the Emergency Department: Multi-System Electronic Health Record–Based Computable Phenotype Derivation and Validation Study
JMIR Med Inform 2019;7(4):e15794
URL: http://medinform.jmir.org/2019/4/e15794/
doi:10.2196/15794
PMID:31674913

©David Chartash, Hyung Paek, James D Dziura, Bill K Ross, Daniel P Nogee, Eric Boccio, Cory Hines, Aaron M Schott, Molly M Jeffery, Mehul D Patel, Timothy F Platts-Mills, Osama Ahmed, Cynthia Brandt, Katherine Couturier, Edward Melnick. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 31.10.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Challenges With Continuous Pulse Oximetry Monitoring and Wireless Clinician Notification Systems After Surgery: Reactive Analysis of a Randomized Controlled Trial

Prathiba Harsha[1,2], MBBS, MSc; James E Paul[2], FRCPC, MSc, MD; Matthew A Chong[3], MD; Norm Buckley[2], BA (Psych), FRCPC, MD; Antonella Tidy[2], HBSc, CCRA; Anne Clarke[2], RN; Diane Buckley[2], RN; Zenon Sirko[2], BSc; Thuva Vanniyasingam[4], MSc, PhD; Jake Walsh[5], MCSE; Michael McGillion[6], RN, PhD; Lehana Thabane[1,2], PhD

[1]Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada

[2]Department of Anesthesia, McMaster University, Hamilton, ON, Canada

[3]Western University, London, ON, Canada

[4]St. Joseph's Healthcare, Hamilton, ON, Canada

[5]Hamilton Health Sciences, Hamilton, ON, Canada

[6]School of Nursing, McMaster University, Hamilton, ON, Canada

Corresponding Author:
Lehana Thabane, PhD
Health Research Methods, Evidence and Impact
McMaster University
3rd Floor Martha Wing, Room H325
50 Charlton Avenue East, St Joseph's Healthcare
Hamilton, ON, L8N 4A6
Canada
Phone: 1 905 522 1155 ext 33720
Email: thabanl@mcmaster.ca

Abstract

Background: Research has shown that introducing electronic Health (eHealth) patient monitoring interventions can improve healthcare efficiency and clinical outcomes. The VIGILANCE (VItal siGns monItoring with continuous puLse oximetry And wireless cliNiCian notification aftEr surgery) study was a randomized controlled trial (n=2049) designed to assess the impact of continuous vital sign monitoring with alerts sent to nursing staff when respiratory resuscitations with naloxone, code blues, and intensive care unit transfers occurred in a cohort of postsurgical patients in a ward setting. This report identifies and evaluates key issues and challenges associated with introducing wireless monitoring systems into complex hospital infrastructure during the VIGILANCE eHealth intervention implementation. Potential solutions and suggestions for future implementation research are presented.

Objective: The goals of this study were to: (1) identify issues related to the deployment of the eHealth intervention system of the VIGILANCE study; and (2) evaluate the influence of these issues on intervention adoption.

Methods: During the VIGILANCE study, issues affecting the implementation of the eHealth intervention were documented on case report forms, alarm event forms, and a nursing user feedback questionnaire. These data were collated by the research and nursing personnel and submitted to the research coordinator. In this evaluation report, the clinical adoption framework was used as a guide to organize the identified issues and evaluate their impact.

Results: Using the clinical adoption framework, we identified issues within the framework dimensions of people, organization, and implementation at the meso level, as well as standards and funding issues at the macro level. Key issues included: nursing workflow changes with blank alarm forms (24/1030, 2.33%) and missing alarm forms (236/1030, 22.91%), patient withdrawal (110/1030, 10.68%), wireless network connectivity, false alarms (318/1030, 30.87%), monitor malfunction (36/1030, 3.49%), probe issues (16/1030, 1.55%), and wireless network standards. At the micro level, these issues affected the quality of the service in terms of support provided, the quality of the information yielded by the monitors, and the functionality, reliability, and performance of the monitoring system. As a result, these issues impacted access through the decreased ability of nurses to make

XSL•FO
RenderX

complete use of the monitors, impacted care quality of the trial intervention through decreased effectiveness, and impacted productivity through interference in the coordination of care, thus decreasing clinical adoption of the monitoring system.

Conclusions: Patient monitoring with eHealth technology in surgical wards has the potential to improve patient outcomes. However, proper planning that includes engagement of front-line nurses, installation of appropriate wireless network infrastructure, and use of comfortable cableless devices is required to maximize the potential of eHealth monitoring.

Trial Registration: ClinicalTrials.gov NCT02907255; https://clinicaltrials.gov/ct2/show/NCT02907255

(JMIR Med Inform 2019;7(4):e14603) doi:10.2196/14603

KEYWORDS

continuous pulse oximetry; wireless notification; issues; evaluation of issues; clinical adoption framework; remote monitoring; postoperative monitoring; false alarm

Introduction

Background

Although the adoption of technology in the hospital environment is slow compared to other fields, there has been a recent increase in digital health solutions proposed for health care issues as technologies improve [1]. With increased workload demands on health care providers, hospitals have turned to technological solutions to improve efficiency and safety of patient care [2]. Patient assessment in a typical postsurgical ward happens only once every four to six hours or, at times, just once during day shifts and irregularly at night [3-5]. This infrequent monitoring, combined with the need for opioids and sedatives and the risk of respiratory depression, may predispose patients postoperatively to more frequent cardiorespiratory arrests (ie, code blues), intensive care unit (ICU) transfers, and the need for resuscitation [6-8]. Early detection is the key to preventing complications [9]. Pulse oximetry, capnography, and wireless remote automated monitoring with clinician notification systems are some of the methods that are being used to support safe patient care in the face of declining clinical staff complements [9-11].

The VItal siGns monItoring with continuous puLse oximetry And wireless cliNiCian notification aftEr surgery (VIGILANCE) study examined the impact of continuous pulse oximetry (CPOX) on the incidence of postoperative respiratory complications [8]. VIGILANCE was an unblinded randomized controlled trial (RCT), targeting noncardiac postsurgical patients (n=2049) at the Juravinski Hospital in Hamilton, Canada. All trial patients with an anticipated length of stay of at least 24 hours and scheduled to stay in one of two surgical wards (E4 and F4) were randomized to either the standard (n=1019) or the intervention arms (n=1030). The standard arm participants received routine monitoring, including assessments every four hours by nurses. The intervention arm patients received continuous monitoring of blood oxygen saturation (SpO_2) and pulse rate (PR) using a wireless respiratory monitoring system, the Nellcor Oxinet III system (Covidien, Dublin, Ireland), in addition to standard monitoring [8,12]. Both E4 and F4 were mixed surgical wards with postsurgical patients admitted after plastic surgery, mastectomies, general surgery, urology, gynecology, orthopedic, and oncology surgeries. Both the wards have 24 beds and have approximately 1100 elective and emergent admissions for surgery per year. On both wards the nurse-to-patient ratio is 1:4.

Need for Evaluation of Issues and Objectives

The VIGILANCE study represented a timely opportunity for the anesthesia service at our institution to work toward reducing the rate of postoperative respiratory complications [8]. At the time, clinical trials research on electronic Health (eHealth) patient monitoring was a burgeoning field, with little prior experience to draw upon, and the challenges associated with introducing digital health systems into the complex hospital infrastructure were not well studied or appreciated [2,13]. Based on our experience, we found that multiple factors interfered with the implementation and conduct of the VIGILANCE study. The purpose of this report was to engage in a reactive analysis by reflecting on the challenges faced by the VIGILANCE research team during the project implementation, followed by identification and evaluation of issues to facilitate future improvements [14]. Through examination of the issues and challenges we faced, our overall aim was to help foster understanding of the difficulties related to eHealth implementation and prevent future implementation challenges [2]. In so doing, our specific objectives were to: (1) identify issues related to deployment of the eHealth intervention system of the VIGILANCE study; and (2) evaluate the influence of these issues on intervention adoption.

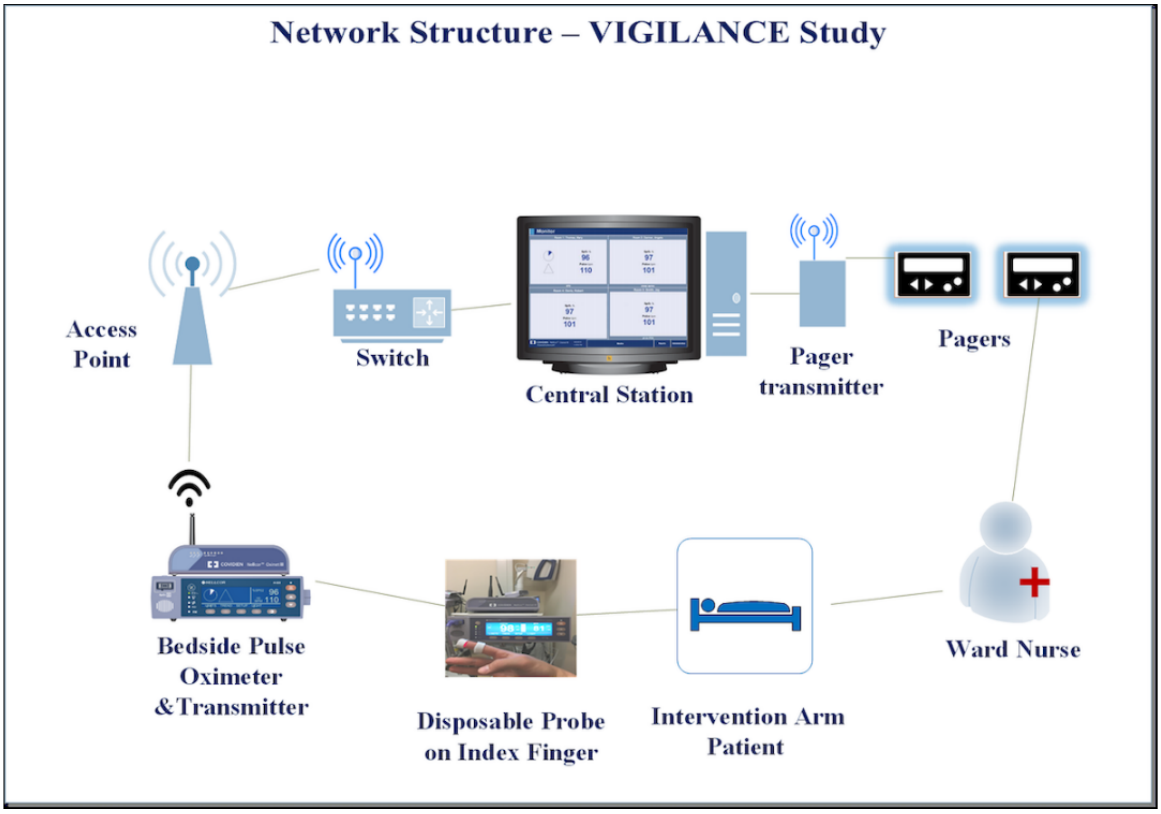
Methods

Vital Signs Monitoring with Continuous Pulse Oximetry and Wireless Clinician Notification After Surgery: Setting and Structure of the Intervention and Network

The monitoring system within the intervention arm allowed for bedside monitoring and wireless pager notification of clinical staff when the alarm threshold was exceeded. Alarms were set at a threshold of SpO_2<90% and PR of ≤50 beats per minute or ≥130 beats per minute, to set a balance between safety and false alarms [8]. The network structure of the monitoring system was comprised of probe, pulse oximeter unit, transmitter, an optional monitor stand, access points, wireless network, switch, central station, pager transmitter, and pagers (Figure 1) [12,15].

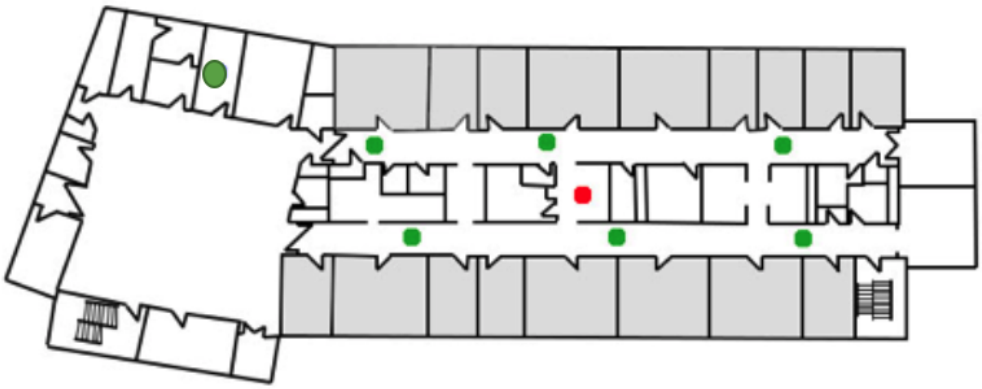
XSL•FO
RenderX

Figure 1. Network Structure of the monitoring system of the VIGILANCE Study. VIGILANCE: VItal siGns monItoring with continuous puLse oximetry And wireless cliNiCian notification aftEr surgery.



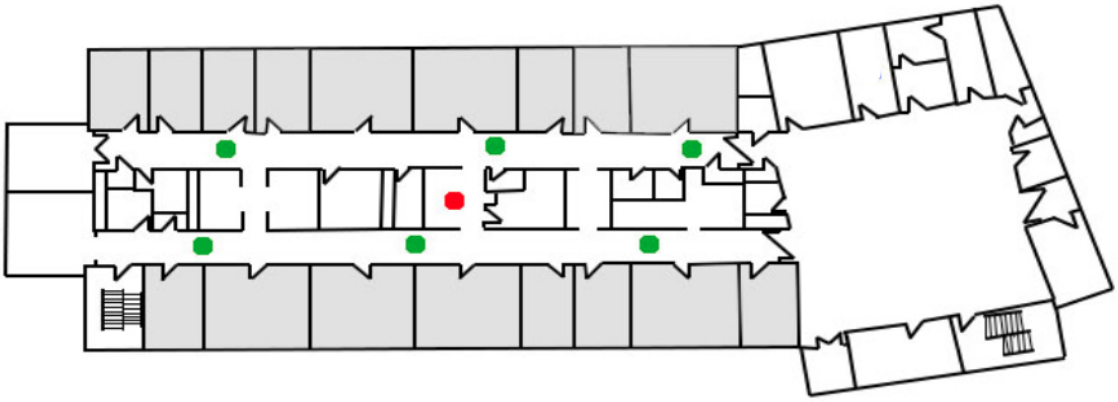
The oximetry probe on the patient's finger was connected to the bedside CPOX monitor through a cable. The CPOX monitor sent patient data through a wired port to the transmitter. The transmitter then converted the data into Ethernet data and wirelessly sent it to the central station via access points. The hospital wireless network structure was made up of the Institute of Electrical and Electronics Engineers (IEEE) standards 802.11a and 802.11g. The IEEE 802.11a standard provided up to 54 megabits per second (Mbps) in a 5 gigahertz (GHz) band, whereas IEEE 802.11g used a 2.4 GHz band [16]. During the installation of the Wireless Local Area Network (WLAN), the Health Information Technology Services (HITS) staff installed the access points after assessing the wireless connectivity, size of the rooms, and structure of the wards [17]. There were seven access points forming a WLAN on the E4 surgical ward (Figure 2) and six access points on the F4 ward (Figure 3).

Figure 2. Access points in E4. Green dots: access points in E4 ward; Red dot: central station; Grey rooms: indicate patient rooms; Unfilled/white rooms: other rooms or spaces.



XSL•FO
RenderX

Figure 3. Access points in F4. Green dots: access points in F4 ward; Red dot: central station; Grey rooms: indicate patient rooms; Unfilled/white rooms: other rooms or spaces.



Information from the access points was wirelessly sent to the hospital's internet routers or switch, which were connected to the central station on their respective wards. The patient data, including the patient name, SpO_2 (%), and PR (beats per minute) along with alarm details, were displayed to the healthcare personnel at their respective central stations. The central station served as the application server and had the Oxinet III software that was required to read the information from the CPOX [12].

Implementation Measures

Testing of the connectivity of the pulse oximeter and the central monitor to the WLAN was done, and the channels were adjusted according to their connectivity after the study was initiated. Prior to study initiation, training and in-service sessions with the continuous monitoring system were held to train the ward nurses. Research nurses visited the ward daily to provide support and to collect the study forms. To reduce the incidence of false notifications due to transient events, notifications were only sent to the nurses after 30 seconds of the event, with a delay of 15 seconds set for both pager notifications and the bedside monitor [8]. Prior to the study starting, the research ethics board application required key personnel of all the involved hospital areas, including the nursing managers of the study wards, to assess the study requirements and comment back to the originator.

Design and Conceptual Framework

The presentation of findings in this evaluation report has been guided by Lau et al's Clinical Adoption (CA) framework [18,19]. The CA framework is an extension of the benefits evaluation framework by Canada Health Infoway, and it is designed to lend guidance to understanding factors, influencing eHealth intervention adoption, in healthcare organizations at macro, meso, and micro levels [18-20]. The overall rationale behind this framework is that, for the successful clinical adoption of technology, the various factors in the framework need to be managed efficiently [18,19]. An underlying premise is that the lower the quality of the technology, as defined by decreased functionality, performance, security, content, availability, and responsiveness, there is an associated decrease of usage, user satisfaction, and acceptance by the stakeholders, and thus overall decreased net benefits [19]. Therefore, this framework was used to understand and organize the various challenges faced during the VIGILANCE study.

For this evaluation, selective constructs were used depending on the issues identified and the context of the project [19,21]. The people, organization, and implementation issues at the meso level were identified [19], and the healthcare standards and funding constructs were included at the macro level [19]. At the micro level, system, information, service quality, use, and net benefits in terms of care quality, access, and productivity, were evaluated [19].

Data Source

During the conduct of the VIGILANCE study, some of the issues that affected the eHealth intervention arm were documented in the case report forms, alarm event form, and nursing user feedback questionnaire. The VIGILANCE study case report forms included items pertaining to patients' deviation from the assigned intervention, the evidence of the type of monitoring received and the reasons for patient withdrawal of the study intervention. These data were captured through the Research Electronic Data Capture (REDCap) system [22]. An alarm event form was used to capture details of alarms and the nursing response to these. Nurses who were assigned to intervention patients completed these forms when patients had any true or false alarms and documented the associated symptoms, along with measures taken to address the alarms. Once the patient was discharged, these alarm event forms were deposited in the study storage box and collected by the research nurse, before being deidentified, scanned, and saved in the study folder in Dropbox [23]. Forms that were not deposited were scanned, along with the patient chart, into the SOVERA (CGI Inc, Montreal, Quebec) hospital health record storage system. Nursing user feedback surveys were also administered to ward nurses after completion of the VIGILANCE study and will be reported in a separate study. Any other issues related to the VIGILANCE pulse oximeter, wireless network connectivity, or nursing workflow, as experienced by the ward staff and the research personnel, were reported to the study coordinator on an ongoing basis.

Data Analysis

Data analysis involved identification of issues from the data sources, categorization of these issues under the meso and macro level of the CA framework, and evaluation of the impact of these on the micro level constructs of the CA framework by reflecting on the VIGILANCE study happenings during

XSL•FO
RenderX

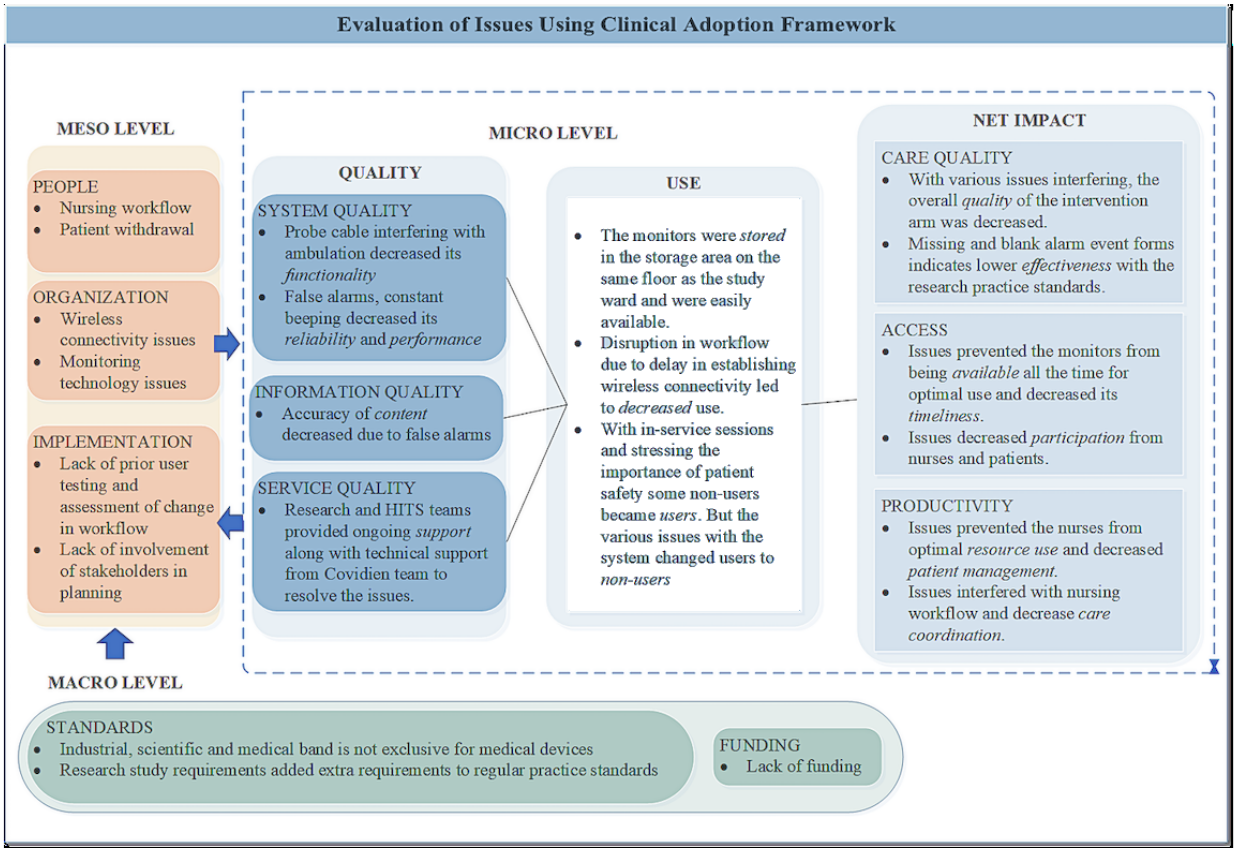
discussions within the study team [14,19]. The problems identified at the meso level were described under people, organization, and implementation categories, and those identified at the macro level were described under standards and funding categories [19]. The impact of these issues on the micro level factors of the CA framework was described under quality, usage, and net impact categories [19]. The identified issues were quantified and are presented using descriptive statistics generated through REDCap, along with counts and percentages for these issues.

Results

Summary

This evaluation report used the CA framework to organize multiple issues that impacted the VIGILANCE intervention (summarized in Figure 4) [18,19]. A detailed analysis of the constructs of the CA framework is included in Multimedia Appendix 1.

Figure 4. Evaluation of issues using the clinical adoption framework. HITS: health information technology services.



People

This includes issues encountered by the key stakeholders: nurses and patients.

Nursing Workflow

Process changes related to the VIGILANCE study protocols resulted in changes to nursing workflows on the study wards. Upon receiving a newly transferred study subject, the nurse assigned to the study patient had to: (1) determine whether the patient was randomized to the standard care or the intervention arm; (2) connect the patient to the monitor; (3) carry the pager; (4) respond to any alarm notifications; and (5) enter the alarm information on the study form. Nursing staff compliance with the study was assessed using the alarm event forms. Among the scanned alarm event forms from the intervention arm, 2.33% (24/1030) were blank without any entry by the nurses and 22.91% (236/1030) of the forms could not be found, indicating decreased compliance with the research practice standards despite multiple in-service sessions. Troubleshooting issues took their time away from actual clinical work.

Patient Withdrawal

After starting on the monitoring, 10.68% (110/1030) of patients withdrew from the CPOX monitoring. Of those 110 individuals, 74 did not provide any reason for withdrawal and the remaining 36 patients provided a total of 44 different reasons. The reasons in the comment section captured various other causes for patient withdrawal. These reasons were categorized and presented in (Table 1).

XSL•FO
RenderX

Table 1. Reasons for patient withdrawal from continuous monitoring (N=110).

Reason	Number of patients, n (%)[a]
No reason provided for withdrawal from monitoring	74 (67.27)
Probe cable	10 (9.07)
Too many false alarms	6 (5.4)
Uncomfortable probe	6 (5.4)
Restriction in ambulation	4 (3.6)
Noise or beeps	4 (3.6)
Confusion or anxiety	3 (2.7)
Monitor malfunction	3 (2.7)
Sleep disturbance	2 (2.7)
Allergic to Velcro	1 (0.9)
Carpal tunnel	1 (0.9)

[a]Percentage calculated will add up to more than 100% as patients reported more than one reason for withdrawal.

Organization Issues

This includes challenges associated with the wireless network connectivity and the monitoring technology.

Wireless Network Connectivity

Research personnel and HITS staff reported that the fundamental structure of the wireless network had the greatest negative impact on VIGILANCE study implementation. The central station failed to display the data being recorded by the oximeters because of a failure in connection at some point in the long cascade of communication (Figure 1). The hospital had upgraded to a newer wireless structure just before the use of these monitors, and a firmware was installed by Covidien to connect to this newer wireless network. This was thought to have caused the connectivity issue initially. Once the monitor lost wireless connectivity, the network connection required reauthentication for security purposes, but the firmware did not optimally support this function. In the hospitals that did not require reauthentication, the firmware was not required for the monitors to connect to the network. This prevented the bedside monitor from connecting promptly to the central monitor and led to nurses taking more time in some cases and in other cases failing to connect the monitor. Although the access points were installed to create a WLAN, the monitors in the rooms farther from the central station had more difficulty in connecting to the WLAN. The CPOX monitors in both E4 and F4 would alternate between the wireless channels 1, 6, or 11 by default and were later set permanently to only channel 6, which resulted in somewhat better stability. Along with unsupported wireless adapter firmware inside the monitor, other medical devices that were connected to the WLAN increased the traffic and caused interference. Interference from nonmedical devices, such as microwaves and other wireless devices, was thought to be another cause for the wireless CPOX failing to connect to the WLAN [24].

Monitoring Technology Issues

Out of 1030 patients, 369 reported at least one monitoring-related issue, and a total of 380 issues were identified. The list of monitoring technology issues for which quantitative data were available is presented below (Table 2).

If a patient's SpO_2 was >90% and they were not bradycardic or tachycardic, the alarm was considered to be false. Among the intervention patients, 30.87% (318/1030) had at least one episode of false alarm. The most common reason for false alarms was movement of the probe. These false alarms resulted in notifications being sent to the nurses, with the nurses going back to the patient room to examine the patient and leading to both a disruption in their workflow and alarm fatigue. Failure of the bedside CPOX monitor to connect or respond was considered a monitor malfunction. Malfunction constituted 9.75% (36/369) of the monitoring technology issues and led to 3.49% (36/1030) of patients receiving standard monitoring.

Table 2. Reported monitoring technology issues (N=369).

Issue	Number of patients, n (%)[a]
At least one false alarm	318 (86.17)
Monitor malfunction	36 (9.75)
Stopped using monitor due to probe or probe cable	16 (4.33)
Stopped using monitor due to false alarms	6 (1.62)
Stopped using monitor due to constant beeping or noise	4 (1.08)

[a]Percentage calculated will add up to more than 100% as patients reported more than one monitoring technology issue.

XSL•FO
RenderX

Constant beeping or noise from a monitor occurred when it was unable to connect to an access point. This issue led to 1.08% (4/369) of the patients with monitoring technology issues to discontinue use of the monitor. An uncomfortable probe or probe cable resulted in 1.55% (16/1030) of the patients withdrawing from the wireless monitoring system. The research personnel and the nurses reported that the bedside monitor was too large for patient rooms. The dimensions were 8.4 cm × 26.4 cm × 17.3 cm [15]. Although the monitor itself was not that big, the broad base of the monitor stand, the intravenous stand, and a chair in the cramped patient cubicle made the nurses feel as if the monitors were bulky.

Implementation Issues

The study design planning and signing off on the ethics approval application of the study did not require the nursing managers to assess the change in workflow prior to study commencement. The ward nurses and HITS team were not part of the study design or planning, and feedback from the nurses was only sought after the study was completed. This led to difficulty managing changes in the nursing workflow and delays in detecting connectivity issues.

Standards and Funding Issues

The frequency that is internationally followed for the Industrial, Scientific, and Medical (ISM) band is between 2.4 and 2.5 GHz, which is not exclusive for medical devices and thus leads to the issue of interference [10]. Congestion caused by multiple medical and nonmedical devices (eg, microwaves and nonhospital devices such as mobile phones) trying to connect to the WLAN resulted in connectivity issues [24]. Research study requirements changed the workflow for the nurses and added extra requirements to their regular practice standards, and there was a lack of funding for involving front-line nurses as part of the study team to lead the project on the wards.

Quality, Use and Net Impact

An evaluation of impact of the meso and macro level issues on the constructs of the micro level is included in Multimedia Appendix 1 and summarized in Figure 4. The key meso and macro level issues identified during the VIGILANCE study impacted system, information, and service quality at the micro level that led to: (1) decreased use; (2) suboptimal system access; (3) decreased care coordination; and (4) decreased effectiveness and efficiency of the system. The wireless connectivity issues and monitor malfunction affected access through a decreased ability of the nurses to make complete use of the monitors, patient withdrawal, change in nursing workflow, false alarms, wireless connectivity, and probe issues affected the care quality of the trial intervention through decreased effectiveness, and productivity was affected by interference with care coordination. Thus, the decreased quality of the eHealth solution led to decreased clinical adoption by stakeholders.

Discussion

Overview

This evaluation report examining the key challenges impacting the implementation of the VIGILANCE trial identified multiple issues in people, organization, implementation, standards, and funding dimensions of the CA framework [19]. Key issues included nursing workflow changes, patient withdrawal, wireless network connectivity, false alarms, monitor malfunction, probe issues, and wireless network standards. These issues led to decreased net benefits and thus decreased clinical adoption of the monitoring system.

Comparison with Prior Work

Ross et al's [2] systematic review discussed factors that influenced eHealth systems in clinical environments. Factors such as the ability of eHealth interventions to adapt to the local environment, system functionality, implementation climate, stakeholder engagement, and stakeholder knowledge and beliefs are consistent with the issues that were identified in this study [2].

In the article by Soomro and Cavalcanti [16], they studied the challenges and opportunities associated with the use of WLAN in hospital environments. The 802.11a and 802.11g wireless network standards, which operate on the distributed coordinated function, work on the random-access mechanism where multiple analog and digital signals are combined and transmitted randomly [16]. When there is an overlap of these signals, the channels will randomly retry to transmit after some time, which might lead to the loss of real-time data [16]. To address this lack in Quality of Service (QoS) support, some of the proposed solutions include: (1) extensions such as 802.11e that can provide priority QoS-based access depending on the type of signal (voice, video, best-effort, background traffic) and parameter-based QoS (allots channel time to each station); (2) guaranteed QoS for distinctive traffics; (3) differentiated services architecture based on traffic and QoS guarantees; and (4) integrated networks with both WLAN and wireless personal area networks [16,25]. Some additional factors that affected WLAN connectivity include: coexistent interference from other devices operating in the same ISM band, different configuration requirements for various devices, and the increasing use of mobile devices [16,26]. With the increasing use of mobile and wireless technology in health care, hospitals must update their infrastructure accordingly [26]. Wireless medical device manufacturers must ensure that devices can coexist with other devices prior to their approval for premarket submission, according to the current guidelines by Food and Drug Administration in the United States [27]. Standards for coexistence and the testing of coexistence of wireless medical devices are currently being developed [27,28]. International groups and the Continua Health Alliance have been formed and are collaborating to standardize medical devices and transmission of data [10,29].

Literature has shown that having a comprehensive approach that involves the stakeholders during the planning of any eHealth implementation yields better results, with increased buy-in, improved workflow, and acceptance of the system [2,30]. In

XSL•FO
RenderX

eHealth projects, issues with change management and omission to test the system prior to implementing have led to project failures [31,32]. User testing before implementation ensures that the system works according to plan and facilitates user buy-in with the digital intervention [31,33]. Champions of the systems have also been identified in the literature as crucial components of eHealth intervention implementation [2,30]. Therefore, involving users in planning the workflow and testing and engaging front-line nurses who could act as champions of the wireless system monitoring would have facilitated the VIGILANCE team in identifying any system issues, streamlining the workflow, and engaging the nurses more efficiently.

False alarms and constant beeping led to patients withdrawing from the continuous monitoring system and interfered with the nursing workflow. Alarm fatigue is a major concern in the hospital environment with the increasing use of monitoring technology in the hospitals, as desensitization of the health care providers due to constant exposure to alarms, beeps, and other noises can put the patient's safety at risk [4,34]. False alarms from the CPOX due to motion have been a significant concern over the years [35]. A cableless oximetry probe is a potential solution to remove hindrances to patient ambulation after surgery [33]. With the recent improvements in motion-resistant technology and algorithms, manufacturers are now using new techniques to address this issue [36].

Lau et al used the CA framework to evaluate the impact of electronic medical records postimplementation in an ambulatory care clinic [19,21]. Various evaluation studies, including systematic reviews, have used this framework to understand technology adoption in different clinical settings [19,37,38]. The CA framework offered a multilevel, interrelated view of the various issues impacting the VIGILANCE intervention implementation.

With future trends towards improvements in biosensors, wireless technology, Bluetooth and radio-frequency identification, more wireless devices capturing multiple physiological parameters are being developed and marketed [4,10,39]. Soon, these monitors will make it possible to monitor all the vital signs on regular hospital wards that are currently routinely monitored in the ICU. It will be important to evaluate this technology carefully to ensure it functions in a way that clinicians expect and in a reliable manner [31,32].

Limitations

A key limitation of this report is that the need to evaluate the impact of factors that might have affected the VIGILANCE study was conceived post study design, and thus we do not have event numbers for all the issues. As this report looked at issues impacting just a single intervention, they might not be generalizable to other eHealth interventions. Future evaluations could include formal evaluation throughout different phases of the project to enhance eHealth intervention implementation and stakeholder management.

Lessons Learned

The findings from this study support the significance of giving importance to not only health outcomes but also to evaluating the process and people aspects of eHealth research projects to overcome challenges and to optimize the use of eHealth intervention. The results from this study have key implications in a clinical setting. The assessment of challenges shows that it is essential for the originators of eHealth research projects to ensure that the stakeholders, such as nurses, other health care providers, and information and technology staff, are consulted in planning and implementing the intervention, establishing the workflow, and testing the intervention in the already existing hospital infrastructure. Identifying champions among the involved stakeholders and having them as leaders of a research project is crucial for better stakeholder engagement and successful eHealth project completion. Medical device manufacturers are encouraged to consider alarm fatigue while providing configuration and display features for their devices. The lessons learned from this study can help future eHealth research implementation projects. Key issues and potential solutions are summarized in (Table 3).

Table 3. Key issues and potential solutions for eHealth research projects.

Issues	Potential solutions
Issues with stakeholder engagement and change management	Involve key stakeholders in planning, establishing workflows, and user testing. Project originators should identify champions and involve them to lead the projects from front-line.
Monitoring technology issues	Usability testing in the actual hospital environment prior to project implementation.
Wireless connectivity	Test for interference and connectivity in the actual environment prior to procuring wireless medical devices.
False alarms	Medical device manufacturers are encouraged to consider alarm fatigue while providing configuration and display features.

Conclusion

Lau et al's CA framework was a useful tool for categorizing and understanding the impact of the issues that influenced the deployment of the intervention in the VIGILANCE study. The wireless network in the hospital was demonstrated to be a critical enabler for eHealth interventions. Devices should be chosen based on the available bandwidth and the ability of the device to coexist with other connected devices, and alarm fatigue should be considered while configuring medical devices. Managing change, establishing workflows, testing usability, and engaging stakeholders are key factors in deploying new digital health solutions aimed at improving the process of care and ultimately patient outcomes. The complexities surrounding the implementation of digital interventions should be taken into consideration along with the clinical outcomes while planning eHealth research studies.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank the nursing staff on the study wards (E4 and F4) and the HITS team at the Juravinski Hospital for their tremendous support. They would also like to thank the research nurses, assistants, and students who contributed towards VIGILANCE study. MC has received the Health Professional Student Research Award from the Canadian Institute of Health Research for his involvement with the VIGILANCE study.

Authors' Contributions

PH, JEP, and LT conceptualized the study; JEP and LT developed the study methodology; PH developed the study protocol; JEP, MAC, and NB were the study investigators; JEP, LT, MM, and NB supervised the study; AC, DB, and MAC provided training; ZS, JW, DB, and AC were responsible for equipment maintenance and troubleshooting; PH, MAC, AT, and MAC were responsible for data collection; PH was responsible for data extraction and verification; PH, TV were responsible for data analysis; JEP, LT, MM, JW, ZS, DB, AC, and AT were responsible for data interpretation; PH, JEP, and LT were responsible for writing the manuscript; and MM, JW, TV, ZS, DB, AC, AT, NB, MAC reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Evaluation of impact of issues on the measures of the clinical adoption framework [19].
[PDF File (Adobe PDF File), 146 KB - medinform_v7i4e14603_app0.pdf]

References

1. Shortliffe EH. Strategic action in health information technology: why the obvious has taken so long. Health Aff (Millwood) 2005;24(5):1222-1233. [doi: 10.1377/hlthaff.24.5.1222] [Medline: 16162567]
2. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
3. McGain F, Cretikos MA, Jones D, Van Dyk DS, Buist MD, Opdam H, et al. Documentation of clinical review and vital signs after major surgery. Med J Aust 2008 Oct 06;189(7):380-383. [Medline: 18837681]
4. McGillion MH, Duceppe E, Allan K, Marcucci M, Yang S, Johnson AP, PROTECT Network Investigators. Postoperative Remote Automated Monitoring: Need for and State of the Science. Can J Cardiol 2018 Jul;34(7):850-862 [FREE Full text] [doi: 10.1016/j.cjca.2018.04.021] [Medline: 29960614]
5. Hands C, Reid E, Meredith P, Smith GB, Prytherch DR, Schmidt PE, et al. Patterns in the recording of vital signs and early warning scores: compliance with a clinical escalation protocol. BMJ Qual Saf 2013 Sep;22(9):719-726. [doi: 10.1136/bmjqs-2013-001954] [Medline: 23603474]
6. Lee LA, Caplan RA, Stephens LS, Posner KL, Terman GW, Voepel-Lewis T, et al. Postoperative opioid-induced respiratory depression: a closed claims analysis. Anesthesiology 2015 Mar;122(3):659-665. [doi: 10.1097/ALN.0000000000000564] [Medline: 25536092]
7. Chung F, Liao P, Yegneswaran B, Shapiro CM, Kang W. Postoperative changes in sleep-disordered breathing and sleep architecture in patients with obstructive sleep apnea. Anesthesiology 2014 Feb;120(2):287-298. [doi: 10.1097/ALN.0000000000000040] [Medline: 24158049]
8. Paul JE, Chong MA, Buckley N, Harsha P, Shanthanna H, Tidy A, et al. Vital sign monitoring with continuous pulse oximetry and wireless clinical notification after surgery (the VIGILANCE pilot study)-a randomized controlled pilot trial. Pilot Feasibility Stud 2019;5:36 [FREE Full text] [doi: 10.1186/s40814-019-0415-8] [Medline: 30858986]
9. Lam T, Nagappa M, Wong J, Singh M, Wong D, Chung F. Continuous Pulse Oximetry and Capnography Monitoring for Postoperative Respiratory Depression and Adverse Events: A Systematic Review and Meta-analysis. Anesth Analg 2017 Dec;125(6):2019-2029. [doi: 10.1213/ANE.0000000000002557] [Medline: 29064874]
10. Nangalia V, Prytherch DR, Smith GB. Health technology assessment review: remote monitoring of vital signs--current status and future challenges. Crit Care 2010;14(5):233 [FREE Full text] [doi: 10.1186/cc9208] [Medline: 20875149]
11. Field MJ. Telemedicine: a guide to assessing telecommunications in healthcare. J Digit Imaging 1997 Aug;10(3 Suppl 1):28 [FREE Full text] [doi: 10.1007/bf03168648] [Medline: 9268830]
12. Nellcor PB. BioClinical Services. 2006. Nellcor Oxinet III Service Manual Internet URL: https://www.bioclinicalservices.com.au/nellcor/pulse-oximetry/oxinet-iii-operators-manual-rev-a-june-2006 [accessed 2019-02-21]
13. Catwell L, Sheikh A. Evaluating eHealth interventions: the need for continuous systemic evaluation. PLoS Med 2009 Aug;6(8):e1000126 [FREE Full text] [doi: 10.1371/journal.pmed.1000126] [Medline: 19688038]

XSL•FO
RenderX

14. Edmondson AC, Bohmer RM, Pisano GP. Disrupted Routines: Team Learning and New Technology Implementation in Hospitals. Administrative Science Quarterly 2001 Dec;46(4):685. [doi: 10.2307/3094828]
15. Covidien. Nellcor OxiMax N-600x Pulse Oximeter Service Manual. 2011. Service Manual Nellcor OxiMax N-600x Pulse Oximeter Internet URL: https://www.medtronic.com/content/dam/covidien/library/us/en/product/pulse-oximetry/N600X_OperatorsManual_EN_10071092A001.pdf [accessed 2019-02-11] [WebCite Cache ID 76JDq4k0M]
16. Soomro A, Cavalcanti D. Opportunities and challenges in using WPAN and WLAN technologies in medical environments [Accepted from Open Call]. IEEE Commun. Mag 2007 Feb;45(2):114-122. [doi: 10.1109/mcom.2007.313404]
17. Cisco. Cisco. Wireless Local Area Network (WLAN) - Cisco Internet URL: https://www.cisco.com/c/en/us/tech/wireless-2f-mobility/wireless-lan-wlan/index.html [accessed 2019-02-19] [WebCite Cache ID 76JEp3E98]
18. Lau F, Price M, Keshavjee K. From benefits evaluation to clinical adoption: making sense of health information system success in Canada. Healthc Q 2011;14(1):39-45. [Medline: 21301238]
19. Francis Lau and Craig Kuziemsky. Clinical Adoption Framework. In: Handbook of eHealth Evaluation: An Evidence-based Approach. Victoria: University of Victoria; Feb 27, 2017:55-76.
20. Lau F, Hagens S, Muttitt S. A proposed benefits evaluation framework for health information systems in Canada. Healthc Q 2007;10(1):112-6, 118. [Medline: 17326376]
21. Lau F, Partridge C, Randhawa G, Bowen M. Applying the clinical adoption framework to evaluate the impact of an ambulatory electronic medical record. Stud Health Technol Inform 2013;183:15-20. [Medline: 23388247]
22. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
23. Dropbox. Dropbox. 2018. What is Dropbox - Features URL: https://www.dropbox.com/features [accessed 2019-02-19] [WebCite Cache ID 1550608463488481]
24. Krishnamoorthy S, Reed J, Anderson C, Max RP, Srikanteswara S. Characterization of the 2.4 GHz ISM band electromagnetic interference in a hospital environment. : IEEE; 2003 Sep 21 Presented at: Proceedings of the 25th Annual International Conference of the IEEE EMBS Cancun; 2003-09-17 to 2003-09-21; Cancun, Mexico p. 3245-3248 URL: http://ieeexplore.ieee.org/document/1280835/ [doi: 10.1109/iembs.2003.1280835]
25. Alavikia Z, Khadivi P, Hashemi MR. A Model for QoS - Aware Wireless Communication in Hospitals. J Med Signals Sens 2012 Jan;2(1):1-10 [FREE Full text] [Medline: 23493832]
26. Yufei Wang, Qixin Wang, Guanbo Zheng, Zheng Zeng, Rong Zheng, Qian Zhang. WiCop: Engineering WiFi Temporal White-Spaces for Safe Operations of Wireless Personal Area Networks in Medical Applications. IEEE Trans. on Mobile Comput 2014 May;13(5):1145-1158. [doi: 10.1109/tmc.2013.31]
27. FDA. https://www.fda.gov/media/71975/download. 2013. US Food & Drug Administration URL: https://www.fda.gov/media/71975/download [accessed 2019-02-21] [WebCite Cache ID 76Mn55bZS]
28. Al Kalaa MO, Balid W, Refai HH, LaSorte NJ, Seidman SJ, Bassen HI, et al. Characterizing the 2.4 GHz Spectrum in a Hospital Environment: Modeling and Applicability to Coexistence Testing of Medical Devices. IEEE Trans. Electromagn. Compat 2017 Feb;59(1):58-66. [doi: 10.1109/TEMC.2016.2602083]
29. Continua Health Alliance. Continua Design Guidelines | Personal Connected Health Alliance Internet URL: http://www.pchalliance.org/continua-design-guidelines [accessed 2019-04-15] [WebCite Cache ID 1555295421613611]
30. Gagnon M, Desmartis M, Labrecque M, Car J, Pagliari C, Pluye P, et al. Systematic review of factors influencing the adoption of information and communication technologies by healthcare professionals. J Med Syst 2012 Feb;36(1):241-277 [FREE Full text] [doi: 10.1007/s10916-010-9473-4] [Medline: 20703721]
31. Jeskey M, Card E, Nelson D, Mercaldo ND, Sanders N, Higgins MS, et al. Nurse adoption of continuous patient monitoring on acute post-surgical units: managing technology implementation. J Nurs Manag 2011 Oct;19(7):863-875. [doi: 10.1111/j.1365-2834.2011.01295.x] [Medline: 21988434]
32. Granja C, Janssen W, Johansen MA. Factors Determining the Success and Failure of eHealth Interventions: Systematic Review of the Literature. J Med Internet Res 2018 May 01;20(5):e10235 [FREE Full text] [doi: 10.2196/10235] [Medline: 29716883]
33. McGillion M, Yost J, Turner A, Bender D, Scott T, Carroll S, et al. Technology-Enabled Remote Monitoring and Self-Management - Vision for Patient Empowerment Following Cardiac and Vascular Surgery: User Testing and Randomized Controlled Trial Protocol. JMIR Res Protoc 2016 Aug 01;5(3):e149 [FREE Full text] [doi: 10.2196/resprot.5763] [Medline: 27480247]
34. Winters BD, Cvach MM, Bonafide CP, Hu X, Konkani A, O'Connor MF, Society for Critical Care Medicine AlarmAlert Fatigue Task Force. Technological Distractions (Part 2): A Summary of Approaches to Manage Clinical Alarms With Intent to Reduce Alarm Fatigue. Crit Care Med 2018 Jan;46(1):130-137. [doi: 10.1097/CCM.0000000000002803] [Medline: 29112077]
35. Jubran A. Pulse oximetry. Crit Care 2015 Jul 16;19:272 [FREE Full text] [doi: 10.1186/s13054-015-0984-8] [Medline: 26179876]
36. Petterson MT, Begnoche VL, Graybeal JM. The effect of motion on pulse oximetry and its clinical significance. Anesth Analg 2007 Dec;105(6 Suppl):S78-S84. [doi: 10.1213/01.ane.0000278134.47777.a5] [Medline: 18048903]

XSL•FO
RenderX

37. Bassi J, Lau F, Lesperance M. Perceived impact of electronic medical records in physician office practices: a review of survey-based research. Interact J Med Res 2012 Jul 28;1(2):e3 [FREE Full text] [doi: 10.2196/ijmr.2113] [Medline: 23611832]
38. Lau F, Price M, Boyd J, Partridge C, Bell H, Raworth R. Impact of electronic medical record on physician practice in office settings: a systematic review. BMC Med Inform Decis Mak 2012 Feb 24;12:10 [FREE Full text] [doi: 10.1186/1472-6947-12-10] [Medline: 22364529]
39. Michard F, Gan TJ, Kehlet H. Digital innovations and emerging technologies for enhanced recovery programmes. Br J Anaesth 2017 Jul 01;119(1):31-39 [FREE Full text] [doi: 10.1093/bja/aex140] [Medline: 28605474]

Abbreviations

CA: clinical adoption
CPOX: continuous pulse oximetry
eHealth: electronic health
GHz: gigahertz
HITS: health information technology services
ICU: intensive care unit
IEEE: Institute of Electrical and Electronics Engineers
ISM: industrial, scientific, and medical
Mbps: megabits per second
QoS: quality of service
PR: pulse rate
REDCap: Research Electronic Data Capture
SpO_2: blood oxygen saturation
VIGILANCE: VItal siGns monItoring with continuous puLse oximetry And wireless cliNiCian notification aftEr surgery
WLAN: wireless local area network

Edited by C Lovis; submitted 06.05.19; peer-reviewed by M Andrews, W Habre, P Schoettker; comments to author 22.06.19; revised version received 05.07.19; accepted 25.07.19; published 28.10.19.

Please cite as:
Harsha P, Paul JE, Chong MA, Buckley N, Tidy A, Clarke A, Buckley D, Sirko Z, Vanniyasingam T, Walsh J, McGillion M, Thabane L
Challenges With Continuous Pulse Oximetry Monitoring and Wireless Clinician Notification Systems After Surgery: Reactive Analysis of a Randomized Controlled Trial
JMIR Med Inform 2019;7(4):e14603
URL: http://medinform.jmir.org/2019/4/e14603/
doi:10.2196/14603
PMID:31661079

©Prathiba Harsha, James E Paul, Matthew A Chong, Norm Buckley, Antonella Tidy, Anne Clarke, Diane Buckley, Zenon Sirko, Thuva Vanniyasingam, Jake Walsh, Michael McGillion, Lehana Thabane. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 28.10.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Usability Factors Associated With Physicians' Distress and Information System–Related Stress: Cross-Sectional Survey

Tarja Heponiemi[1], PhD; Sari Kujala[2], PhD; Suvi Vainiomäki[3], PhD; Tuulikki Vehko[1], PhD; Tinja Lääveri[4], MD; Jukka Vänskä[5], MSocSci; Eeva Ketola[1], PhD; Sampsa Puttonen[6], PhD; Hannele Hyppönen[1], PhD

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

[2]Aalto University, Espoo, Finland

[3]University of Turku, Turku, Finland

[4]Helsinki University Hospital and University of Helsinki, Helsinki, Finland

[5]Finnish Medical Association, Helsinki, Finland

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

Corresponding Author:
Tarja Heponiemi, PhD
National Institute for Health and Welfare
PoBox 30
Helsinki, 00271
Finland
Phone: 358 295247434
Email: tarja.heponiemi@thl.fi

Abstract

Background: Constantly changing and difficult-to-use information systems have arisen as a significant source of stress in physicians' work. Physicians have reported several usability problems, system failures, and a lack of integration between the systems and have experienced that systems poorly support the documentation and retrieval of patient data. This stress has kept rising in the 21st century, and it seems that it may also affect physicians' well-being.

Objective: This study aimed to examine the associations of (1) usability variables (perceived benefits, technical problems, support for feedback, and user-friendliness), (2) the number of systems in daily use, (3) experience of using information systems, and (4) participation in information systems development work with physicians' distress and levels of stress related to information systems (SRIS) levels.

Methods: A cross-sectional survey was conducted among 4018 Finnish physicians (64.82%, 2572 out of 3968 women) aged between 24 and 64 years (mean 46.8 years) in 2017. The analyses of covariance were used to examine the association of independent variables with SRIS and distress (using the General Health Questionnaire) adjusted for age, gender, employment sector, specialization status, and the electronic health record system in use.

Results: High levels of technical problems and a high number of systems in daily use were associated with high levels of SRIS, whereas high levels of user-friendliness, perceived benefits, and support for feedback were associated with low levels of SRIS. Moreover, high levels of technical problems were associated with high levels of psychological distress, whereas high levels of user-friendliness were associated with low distress levels. Those who considered themselves experienced users of information systems had low levels of both SRIS and distress.

Conclusions: It seems that by investing in user-friendly systems with better technical quality and good support for feedback that professionals perceive as being beneficial would improve the work-related well-being and overall well-being of physicians. Moreover, improving physicians' skills related to information systems by giving them training could help to lessen the stress that results from poorly functioning information systems and improve physicians' well-being.

(JMIR Med Inform 2019;7(4):e13466) doi:10.2196/13466

KEYWORDS

health information systems; physicians; electronic health records; computers, digital

XSL•FO
RenderX

Introduction

Background

The poor usability of information systems (IS)—such as problematic data entry and difficulties in use—has arisen as an important source of stress in physicians' work [1-3]. Moreover, a recent finding shows that the physicians' strain coming from the IS has kept rising in the 21st century [1]. Evidence implies that this strain may even affect the well-being of physicians [4,5].

Finnish physicians have given their electronic health record (EHR) systems rather critical ratings, depending on the working facility. When asked about the overall school grading for the EHR primarily in use, on a scale from 1 (*fail*) to 7 (*excellent*), the average ratings varied from 2.5 to 4.3 in 2010 and 3.2 to 4.4 in 2014 [6,7]. Recent findings from the United States showed that EHR design and use factors accounted for 12.5% of variance in measures of stress and 6.8% of variance in measures of burnout [8]. However, previous findings also showed that satisfied physicians who find their IS facilitate the continuity of care and make clinical information more accessible [9-11].

Usability problems with the current EHRs are common. Previous studies from the United States and Denmark showed over 100 usability problems, for example, related to consistency, user control, flexibility, and lack of support [12,13]. Poor IS usability and time-consuming data entry have been found as prominent sources of US physicians' professional dissatisfaction [5]. Moreover, technical problems in IS have been related to more experiences of time pressure and lower possibilities to control one's job [14]. Previous studies have also shown that physicians' stress emerging from the IS is related to cognitive workload and time pressures at work [1,15].

The use of the IS is further complicated by the multiplicity of screens and options and by the need to use many different systems. There are findings suggesting that a higher number of functions increase stress and job dissatisfaction [4]. In addition, the ever-changing new functionalities and systems need constant development of physicians' skills and time for orientation. Thus, being experienced in using different systems might help when facing challenges related to the IS.

Physicians' participation in the development work associated with the IS might help to tackle usability problems and improve physicians' attitudes toward the IS. However, physicians are dissatisfied with their impact possibilities and think that neither managers nor software providers are interested in end users' opinions [16].

Objectives

Thus, as mentioned above, the increasing use of IS in daily work is associated with many problems that stress physicians, and previous findings suggest that this might even have negative ramifications for physicians' general well-being. However, the evidence is still limited; there are no exact findings showing which factors related to IS and EHRs are the most stressful, and whether these problems are related to the actual well-being of physicians. Therefore, this study examined the associations of (1) usability variables (perceived benefits, technical problems, support for feedback, and user-friendliness), (2) the number of systems in daily use, (3) experience of using EHRs, and (4) participation in EHR development work with physicians' distress and stress related to IS (SRIS).

Methods

The Study Sample

The data were collected in April 2017 [17]. The addresses were obtained from the Finnish Medical Association's register. A link to the study was sent via email to the target group that was all physicians younger than 65 years who lived in Finland (N=19,627). Altogether 93.37% (18,326/19,627) had provided email addresses to which the survey could be sent. The questionnaires were sent to all working-aged physicians with a cover letter calling for responses from physicians in clinical work. This was done because the Finnish Medical Association's membership register did not allow us to select only physicians in clinical work as the target population. Thus, the sample also included physicians who were not practicing clinical patient work at the time of the data collection. Those who answered that they did not do clinical patient work (n=48) were coded as *missing*.

The representativeness of the sample was assessed by comparing the distributions of the background variables with the corresponding distribution of the target population. The respondents were slightly older than the eligible population (the percentage of those older than 54 years was 31.65% (1266/3999) in the respondents, whereas it was 26.90% (5280/19,627) in the eligible population), more often female (64.81% (2572/3968) in the respondents and 61.09% (11,992/19,627) in the eligible population), and more often specialized (67.44% (2710/4018) in the respondents and 59.90% (11,757/19,627) in the eligible population) [18]. There were no significant regional differences between the respondents and eligible population according to the place of work [18]. Due to incomplete data in some variables, the n varied between 3744 and 3780 in different analyses.

The Context

There have been multiple reforms in Finland lately regarding IS in the health care sector. The public sector EHR adoption in Finland reached 100% in 2010, and the private sector adoption rates of EHRs are also high [19]. Finland has launched the national digital repository for electronic patient data, Kanta, in phases during the period 2012 to 2017. Kanta is targeted to health care service providers, pharmacies, and citizens. Kanta services include electronic prescriptions, My Kanta pages for citizens, a patient data repository, and an electronic prescription database. It is mandatory for all public health care providers to join Kanta and also for those private service providers that use electronic archiving.

Measurements

The measure items used in this study can be seen in Multimedia Appendix 1.

SRIS was used as a dependent variable and measured with the mean of 2 items, framed in 1 question that asked how often

XSL•FO
RenderX

(during the past half-year period) the respondent had been distracted by, worried about, or stressed about (1) constantly changing IS and (2) difficult, poorly performing information technology (IT) equipment or software. The answers were rated on a 5-point Likert scale ranging from 1 (*never*) to 5 (*very often*). The scale's reliability (Cronbach alpha) was .66 in this sample. This measure has previously been used and associated with, for example, employees' distress, cognitive workload, and higher levels of on-call duties [15,20,21]. In Finland, in addition to EHRs, a large number of separate IS are also in physicians' use, such as laboratory and radiological data systems, clinical decision-making software, and systems related to quality, patient safety, and security [22]. The wording of this measure refers to all these systems, not only to EHRs. The reliability of this scale (.66) can be considered low but acceptable given that the scale only included 2 items [23].

Psychological distress was used as a dependent variable and measured with the 4 items (alpha=.84) from General Health Questionnaire-12 (GHQ-12) [24] that represent the anxiety/depression factor, as suggested by Graetz [25]. Graetz's 3-factor structure has been suggested to be the most preferable factor model for GHQ-12 [26]. The GHQ is one of the most popular and very widely used measures of mental health and minor psychiatric disorders. A variety of scoring methods can be used when using the GHQ. The bimodal scoring method allows identification of the threshold for pathological deviations. This study used Likert-scale answer options ranging from 1 to 4 with a continuous mean variable, higher scores indicating a higher level of distress. This scoring method was used to get more variation because we were interested in general well-being and distress levels (not in pathology) in the basically healthy working-aged physician population. We have previously associated this measure with, for example, physicians' collegial support, team climate, and patient-related stress [27,28].

The following variables were used as independent variables: *The number of systems in daily use* was assessed by asking about the number of clinical systems that the responder needed to log into on a daily basis when working with patients. The response options were 0/1/2/3/4/5 or "more"/"my work does not include clinical work" (coded as *missing*). For the analyses, this measure was coded as 0=1 to 2 systems in daily use (nobody answered that they had 0 systems in daily use) and 1=3 or more systems in daily use. *Experience of using EHRs* was assessed by asking how experienced the respondent was as an EHR user with a 5-point scale ranging from 1 (*beginner*) to 5 (*expert*). For the analyses, this variable was coded as 0=*beginner* (answer options 1-3) and 1=*expert* (answer options 4 and 5). *Participation in the development work of the IS* was assessed by asking whether respondent had participated in the development work of the IS. Answer options were as follows: *plenty*/*a little*/*no*. For the analyses, variable was coded as 0=*no* and 1=*yes* (answer options: *plenty* and *a little*).

The usability variables were used as independent variables in this study and represented the 4 strongest factors (perceived benefits, technical problems, feedback, and user-friendliness) with the highest loadings from a previous factor analysis that used 36 usability-related items among Finnish physicians [14]. These variables have previously been associated with physicians' time pressure and control [14]. *The perceived benefits of the EHRs* were assessed by 6 items (alpha=.79) asking, for example, how IS help to improve the quality of care. *Technical problems* was a topic assessed by 6 items (alpha=.81), for example, "Information entered/documented occasionally disappears from the information system." *Feedback* was assessed with 4 items (alpha=.78), such as "The system vendor implements corrections and change requests according to the suggestions of end users." *User-friendliness* was assessed with 9 items (alpha=.86) asking, for example, whether the arrangement of fields and functions is logical on the computer screen. These usability variables were rated on a 5-point Likert scale ranging from 1 (*fully disagree*) to 5 (*fully agree*). We analyzed technical quality and user-friendliness in separate analyses to avoid multicollinearity because these variables correlated (r=–0.65). However, a recent validation study showed that these dimensions are separate constructs and should be studied separately as well as that all these usability variables offer a useful tool to measure the usability of the health IS [29].

The adjustment variables used were as follows: *specialization status*, which was asked as *none/specialization is ongoing/specialist. Employment sector* was categorized into 3 groups: hospitals, primary care, and other sectors. Moreover, respondents were asked their age, gender, and which EHR system they mainly use.

Statistical Analysis

The association of independent variable levels with SRIS and distress was analyzed with analyses of covariance (in separate analyses). The analyses were conducted in 2 steps. In the first step, the analyses included adjustments variables (age, gender, employment sector, specialization status, and the EHR system in use), the number of systems in daily use, experience of using EHRs, and participation in IS-related development work. In the second step, usability variables (perceived benefits, feedback, and technical problems/user-friendliness) were added to the former model. The analyses were conducted in these 2 steps to find out whether usability variables would partly account for possible associations of the independent variables from the first step with SRIS or distress. User-friendliness and technical problems were analyzed in separate analyses to avoid multicollinearity.

Results

Characteristics of the Study Population

The characteristics of the study population can be seen in Table 1. The questionnaire was answered by 4018 physicians (64.82%, 2572/3968, women; response rate 21.9%) aged between 24 and 64 years (mean 46.8, SD 11.1). Almost half of the respondents worked in hospitals, and two-thirds were specialists. Over half of the respondents had 1 to 2 systems in their daily use and 71.82% (2886/4018) considered themselves as experienced in using EHRs.

XSL•FO
RenderX

Table 1. The characteristics of the study sample (N=4018).

Characteristic	Value
Gender, n (%)	
Men	1396 (35.18)
Women	2572 (64.82)
Employment sector, n (%)	
Hospital	1943 (48.59)
Primary health care	1070 (26.76)
Other	986 (24.65)
Specialist status, n (%)	
No	401 (10.00)
Specialization ongoing	907 (22.57)
Yes	2710 (67.43)
Systems in daily use, n (%)	
1–2	2375 (60.43)
≥3	1555 (39.57)
Experience in using EHRs[a], n (%)	
Beginner	1111 (27.80)
Experienced	2886 (72.20)
Participation in IS[b] development, n (%)	
Not at all	2045 (51.34)
Yes	1938 (48.66)
Age, mean (SD)	46.76 (11.05)
SRIS[c,d], mean (SD)	3.32 (0.92)
Psychological distress[e], mean (SD)	1.83 (0.66)
Perceived benefits[d], mean (SD)	2.77 (0.79)
Technical problems[d], mean (SD)	2.83 (0.86)
Feedback[d], mean (SD)	2.25 (0.91)
User-friendliness[d], mean (SD)	2.81 (0.81)

[a]EHRs: electronic health records.

[b]IS: information systems.

[c]SRIS: stress related to information systems.

[d]The scale ranged between 1 and 5.

[e]The scale ranged between 1 and 4.

Stress Related to Information Systems

Analyses of covariance showed that all the studied variables were significantly associated with SRIS (Table 2), but the association of participation in development with SRIS attenuated to nonsignificance after adjusting for usability factors. Those who had more than 3 systems in daily use (mean SRIS 3.47, SE 0.027) had higher levels of SRIS compared with those who had only 1 or 2 systems in daily use (mean SRIS 3.23, SE 0.022; the means shown here are estimated marginal means with all adjustments). Those who had longer experience in using EHRs (mean SRIS 3.30, SE 0.022) had lower levels of SRIS compared with those who were beginners (mean SRIS 3.40, SE 0.029). High levels of technical problems were associated with high levels of SRIS, whereas high levels of user-friendliness, perceived benefits, and feedback were associated with low levels of SRIS. The study variables were able to explain much of the variance in SRIS given the rather high adjusted R squared (0.35).

XSL•FO
RenderX

Table 2. The results of the analyses of covariance for stress related to information systems.

Studied variables[a]	Model A		Model B	
	F test *(df)*	*P* value	*F* test *(df)*	*P* value
Number of systems in daily use	145.70 *(1)*	<.001	52.32 *(1)*	<.001
Experience of using EHRs[b]	12.22 *(1)*	<.001	13.73 *(1)*	<.001
Participation in IS[c] development	15.76 *(1)*	<.001	3.54 *(1)*	.06
Perceived benefits	—[d]	—	95.13 *(1)*	<.001
Technical problems	—	—	719.50 *(1)*	<.001
Feedback	—	—	25.88 *(1)*	<.001
User-friendliness	—	—	376.86 *(1)*	<.001
R^2	0.082 *(1)*	—	0.349 *(1)*	—

[a]All analyses were adjusted for gender, age, employment sector, specialization status, and electronic health record in use.

[b]EHRs: electronic health records.

[c]IS: information systems.

[d]Not applicable.

Psychological Distress

The experience of using EHRs, technical problems, and user-friendliness were significantly associated with distress (Table 3). Those who were experienced users of EHRs (mean SRIS 1.82, SE 0.019) had lower levels of distress compared with those who were beginners (mean SRIS 1.92, SE 0.025). High levels of technical problems were associated with high levels of distress, whereas high levels of user-friendliness were associated with low levels of distress. Even though technical problems had a rather strong association with distress, the studied IS-related variables were not able to explain much of the variance in distress given the low adjusted R squared levels of the models.

Table 3. The results of the analyses of covariance for distress.

Variables[a]	Model A		Model B	
	F test *(df)*	*P* value	*F* test *(df)*	*P* value
Number of systems in daily use	3.61 *(1)*	.06	0.56 *(1)*	.46
Experience of using EHRs[b]	15.32 *(1)*	<.001	15.54 *(1)*	<.001
Participation in IS[c] development	0.11 *(1)*	.75	0.00 *(1)*	.99
Perceived benefits	—[d]	—	3.74 *(1)*	.05
Technical problems	—	—	21.05 *(1)*	<.001
Feedback	—	—	0.41 *(1)*	.52
User friendliness	—	—	6.77 *(1)*	.01
R^2	0.018 *(1)*	—	0.028 *(1)*	—

[a]All analyses were adjusted for gender, age, employment sector, specialization status, and the electronic health record in use.

[b]EHRs: electronic health records.

[c]IS: information systems.

[d]Not applicable.

Discussion

Principal Findings

This study found that high levels of technical problems and high number of systems in daily use were associated with high levels of IS-related stress, whereas high levels of user-friendliness, perceived benefits, and support for feedback were associated with lower levels of this stress. SRIS levels were also lower for those who considered themselves as experienced users of EHRs. Moreover, we found that IS-related variables were also associated with physicians' well-being. More specifically, we found that high levels of technical problems were associated with high levels of psychological distress, whereas high levels of user-friendliness were associated with low distress levels. Those who considered themselves as experienced users of EHRs had lower levels of distress.

XSL•FO
RenderX

Limitations

This study relied on self-reported measures, which may lead to problems associated with an inflation of the strengths of relationships and with common method variance. To minimize problems with self-reports, we used measures that showed good reliability and have been used in previous studies. Moreover, although we controlled for many factors—such as age, gender, employment sector, specialization status, and the EHR system in use—we cannot rule out the possibility of residual confounding. Finland is among the forerunners in the digitalization of health care [30], and tax-financed universal health care is provided for all residents; therefore, generalizing our findings to countries with other types of health care systems or IT systems should be done with caution. However, digitalization is increasing at a high pace in developed countries, and previous studies showed that IS cause stress to physicians, and all physicians have to face new challenges coming from IS [1].

The total number of respondents in the survey was rather large, about 4000. However, the response rate remained relatively low (21.92%; 4018/18,326), thus the generalizability of the findings to all physicians should be done with caution. The questionnaire was sent only electronically to physicians' emails; thus, it was not possible to answer by paper, which may have affected the response rate. Moreover, the survey was targeted to all physicians in clinical work, but the Finnish Medical Association's membership register did not allow us to select only physicians in clinical work as the target population. Therefore, the questionnaire was sent to all working-aged physicians with a cover letter calling for responses from physicians in clinical work. However, comparison with the target population showed good representativeness of the sample [18].

We found that IS-related variables were associated with stress levels and even well-being. However, according to our findings, it is not possible to clearly indicate whether the use of too many poorly functioning IS has extreme consequences and seriously impairs physicians' working life. Thus, it is difficult to define the clinical meaning of our findings. Future studies are needed in this regard.

Comparison With Previous Results

Our findings are congruent with previous findings showing that problems with IS may have negative ramifications for the well-being of physicians. For example, problems with IS have been associated with physicians' higher likelihood of burnout [31]. Poor EHR usability, time-consuming data entry, interference with face-to-face patient care, inefficient and less fulfilling work content, an inability to exchange health information between EHR products, and the degradation of clinical documentation have all been associated with physicians' professional dissatisfaction [5]. Moreover, technical problems in EHRs have been related to more experiences of time pressure and fewer possibilities to control one's job [14]. Previous studies have also shown that physicians' stress emerging from IS is related to cognitive workload, problems in teamwork, job dissatisfaction, and time pressures at work [1,15]. Moreover, IS have been associated with job dissatisfaction and intent to leave [4].

Technical problems appeared as the most important IS-related risk factor for both SRIS and psychological distress in our study. In addition, previous studies have shown the importance of the technical quality of the IS among physicians. For example, it has been shown that the technical characteristics of the IS, such as the reliability, response time, and functionality, emerged as the most important factor associated with user satisfaction [32]. Moreover, technical problems have been related to more experiences of time pressure and fewer possibilities to control one's job [14]. Technical problems have also been found as an important barrier to the uptake of a computerized decision-support system [33]. Moreover, technical problems have also previously been found to be one of the most important challenges for patients when using mobile intervention tools [34] and Web-based intervention tools [35]. Of the technical problems, system instability in particular has been a primary concern in previous studies [6,7]. The importance of technical problems is not a surprise given that system errors, instability, missing information, low speed, and unexpected reactions may seriously challenge the workflow, waste time, hinder the doctor-patient relationship, and cause danger to patient safety.

Experience in using EHRs seemed to be an important factor in our study. Years of experience in using laboratory IS have previously been associated with usability ratings [36]. Experience is important given that systems change often, and physicians have to learn to master the new systems and are required to constantly develop their skills. In Finland, it has been found that learning to use an EHR requires a lot of training, and the time needed for this learning has increased between the years 2010 and 2014. EHRs may be challenging to use because of the multiplicity of screens, options, and navigational aids [37]. The complexity and usability problem associated with EHRs demands that physicians allocate time and effort to mastering them. However, the demands and pressures of care may not afford them this time [38]. Physicians may also see being forced to learn how to use the EHR system effectively and efficiently as a burden.

SRIS was higher among those physicians who had a higher number of systems in daily use. This corresponds well with previous findings showing that the multiple sign-ins required for multiple systems and the use of several systems simultaneously caused stress among health care professionals; in addition, the need to use multiple views was perceived as disruptive [39]. It has also been found that using several clinical systems on a daily basis led to the experience of time pressure and lessened job control [14]. We found that approximately 40% (39.56%, 1555/3930) of our respondents used 3 or more clinical systems on a daily basis. These physicians might be a group at high risk of stress. Thus, decreasing the number of systems a physician needs to log in to could have a big effect on physicians' work-related stress levels. If it is not possible to decrease the number of systems in daily use, it might be useful to identify these physicians and offer them support or provide them with compensation for their efforts (such as extra time off).

XSL•FO
RenderX

In our study, participation in IS-related development work did not have an effect on SRIS or distress levels. Half of the participants had participated in development work, which can be considered as a big proportion. A previous study suggested that participation in development work may cause time pressure but gives an important perception of having opportunities to control one's job [14]. It has been suggested that physicians should be included more in the development of their IS [1,15]. Moreover, it has been shown that physicians are interested in participating in IS development [16] and physician-driven improvements to EHR systems have been found to be useful [40]. An alternative approach to physicians' participation in development work is to question why physicians should invest their time and be involved in developing the IS when their education is totally focused on another subject. On the contrary, perhaps IT professionals should invest more time and effort in understanding the needs of physicians, for example, by using robust heuristic methods and dedicated resources.

Conclusions

We found that the usability of the IS, the number of systems in daily use, and one's experience as a user are associated with how stressful a physician perceives the IS to be and, furthermore, to a smaller extent, associated with the physician's well-being. According to our results, it seems that by investing in user-friendly systems with better technical quality and good support for feedback that professionals perceive as being of benefit would improve the work-related well-being and overall well-being of physicians. In particular, preventing technical problems seems to be very important.

Organizations should pay much more attention to the usability of their systems. By offering easy-to-use systems without technical problems, organizations could promote the work of physicians and release time for patient work. Good systems support workflow instead of hindering it. Improving the stability of the IS, a single sign-on, better documentation and retrieval of patient data, a peaceful documentation environment, and better access to patient data from other organizations have all been suggested as tools for promoting the IS-related well-being of health care professionals [39]. However, it seems that the needs of documenting and billing are prioritized when designing the IS instead of focusing on the needs of doctors and patients [11]. Moreover, IT professionals and hospital administrators may have a stronger voice compared with end users in decisions about the IS because they are perceived more clearly by vendors as the buyers of their systems and given higher priority [41].

Moreover, organizations would benefit from offering training opportunities to physicians and minimizing the number of their systems in daily use. However, uncertainty seems to exist about whose responsibility this training would be [39], and it would be beneficial if organizations had some kind of strategy for improving their professionals' electronic skills. Time has been shown to be of great importance in relation to IS use [1,15]. Physicians also need some sort of support when they face IS-related problems. For instance, clerical support personnel have been found to lessen stress and fatigue after implementing new systems [42].

Acknowledgments

This study was supported by the Finnish Work Environment Fund (project 116104), the Strategic Research Council at the Academy of Finland (projects 303607 and 327145), and the Ministry of Social Affairs and Health (project 112241). The authors wish to thank all the physicians that responded to the questionnaires.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Measures used in the study.
[PDF File (Adobe PDF File), 365 KB - medinform_v7i4e13466_app1.pdf]

References

1. Heponiemi T, Hyppönen H, Vehko T, Kujala S, Aalto A, Vänskä J, et al. Finnish physicians' stress related to information systems keeps increasing: a longitudinal three-wave survey study. BMC Med Inform Decis Mak 2017 Oct 17;17(1):147 [FREE Full text] [doi: 10.1186/s12911-017-0545-y] [Medline: 29041971]
2. Shanafelt TD, Dyrbye LN, Sinsky C, Hasan O, Satele D, Sloan J, et al. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. Mayo Clin Proc 2016 Jul;91(7):836-848. [doi: 10.1016/j.mayocp.2016.05.007] [Medline: 27313121]
3. Vänskä J, Viitanen J, Hyppönen H, Elovainio M, Winblad I, Reponen J. Doctors critical of electronic patient record systems. Finn Med J 2010;50-52:4177-4183.
4. Babbott S, Manwell LB, Brown R, Montague E, Williams E, Schwartz M, et al. Electronic medical records and physician stress in primary care: results from the MEMO Study. J Am Med Inform Assoc 2014 Feb;21(e1):e100-e106 [FREE Full text] [doi: 10.1136/amiajnl-2013-001875] [Medline: 24005796]

XSL•FO
RenderX

5. Friedberg M, Chen P, Van Busum KR, Aunon F, Pham C, Caloyeras J. Factors Affecting Physician Professional Satisfaction and Their Implications for Patient Care. Washington, DC: Rand offices; 2013.
6. Kaipio J, Lääveri T, Hyppönen H, Vainiomäki S, Reponen J, Kushniruk A, et al. Usability problems do not heal by themselves: national survey on physicians' experiences with EHRs in Finland. Int J Med Inform 2017 Jan;97:266-281 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.10.010] [Medline: 27919385]
7. Viitanen J, Hyppönen H, Lääveri T, Vänskä J, Reponen J, Winblad I. National questionnaire study on clinical ICT systems proofs: physicians suffer from poor usability. Int J Med Inform 2011 Oct;80(10):708-725. [doi: 10.1016/j.ijmedinf.2011.06.010] [Medline: 21784701]
8. Kroth PJ, Morioka-Douglas N, Veres S, Babbott S, Poplau S, Qeadan F, et al. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Netw Open 2019 Aug 2;2(8):e199609 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9609] [Medline: 31418810]
9. Hellström L, Waern K, Montelius E, Astrand B, Rydberg T, Petersson G. Physicians' attitudes towards ePrescribing--evaluation of a Swedish full-scale implementation. BMC Med Inform Decis Mak 2009 Aug 7;9:37 [FREE Full text] [doi: 10.1186/1472-6947-9-37] [Medline: 19664219]
10. Bouamrane M, Mair FS. A study of general practitioners' perspectives on electronic medical records systems in NHSScotland. BMC Med Inform Decis Mak 2013 May 21;13:58 [FREE Full text] [doi: 10.1186/1472-6947-13-58] [Medline: 23688255]
11. O'Malley AS, Grossman JM, Cohen GR, Kemper NM, Pham HH. Are electronic medical records helpful for care coordination? Experiences of physician practices. J Gen Intern Med 2010 Mar;25(3):177-185 [FREE Full text] [doi: 10.1007/s11606-009-1195-2] [Medline: 20033621]
12. Edwards PJ, Moloney KP, Jacko JA, Sainfort F. Evaluating usability of a commercial electronic health record: a case study. Int J Hum-Comp Stud 2008 Oct;66(10):718-728. [doi: 10.1016/j.ijhcs.2008.06.002]
13. Kjeldskov J, Skov MB, Stage J. A longitudinal study of usability in health care: does time heal? Int J Med Inform 2010 Jun;79(6):e135-e143. [doi: 10.1016/j.ijmedinf.2008.07.008] [Medline: 18757234]
14. Vainiomäki S, Aalto A, Lääveri T, Sinervo T, Elovainio M, Mäntyselkä P, et al. Better usability and technical stability of EPRs could lead to better work-related well-being among physicians. Appl Clin Inform 2017 Oct;8(4):1057-1067 [FREE Full text] [doi: 10.4338/ACI-2017-06-RA-0094] [Medline: 29241245]
15. Heponiemi T, Hyppönen H, Kujala S, Aalto A, Vehko T, Vänskä J, et al. Predictors of physicians' stress related to information systems: a nine-year follow-up survey study. BMC Health Serv Res 2018 Apr 13;18(1):284 [FREE Full text] [doi: 10.1186/s12913-018-3094-x] [Medline: 29653530]
16. Martikainen S, Viitanen J, Korpela M, Lääveri T. Physicians' experiences of participation in healthcare IT development in Finland: willing but not able. Int J Med Inform 2012 Feb;81(2):98-113. [doi: 10.1016/j.ijmedinf.2011.08.014] [Medline: 21956004]
17. Hyppönen H, Lumme S, Reponen J, Vänskä J, Kaipio J, Heponiemi T, et al. Health information exchange in Finland: usage of different access types and predictors of paper use. Int J Med Inform 2019 Feb;122:1-6. [doi: 10.1016/j.ijmedinf.2018.11.005] [Medline: 30623778]
18. Saastamoinen A, Hyppönen H, Kaipio J, Lääveri T, Reponen J, Vainiomaki S, et al. Lääkärien arviot potilastietojärjestelmistä ovat parantuneet hieman [Doctors' assessments of patient information systems have improved slightly]. Finn Med J 2018;73(34):1814-1819 [FREE Full text]
19. Hyppönen H, Hämäläinen P, Reponen J. e-health and e-Welfare of Finland: check point 2015. In: National Institute for Health and Welfare (THL) Report 18/2015. Finland: National Institute for Health and Welfare; 2015.
20. Heponiemi T, Aalto A, Pekkarinen L, Siuvatti E, Elovainio M. Are there high-risk groups among physicians that are more vulnerable to on-call work? Am J Emerg Med 2015 May;33(5):614-619. [doi: 10.1016/j.ajem.2015.01.034] [Medline: 25680563]
21. Kuusio H, Heponiemi T, Aalto A, Sinervo T, Elovainio M. Differences in well-being between GPs, medical specialists, and private physicians: the role of psychosocial factors. Health Serv Res 2012 Feb;47(1 Pt 1):68-85 [FREE Full text] [doi: 10.1111/j.1475-6773.2011.01313.x] [Medline: 22091688]
22. Reponen J, Kangas M, Hämäläinen P, Keränen N, Haverinen J. Use of information and communications technology in Finnish health care in 2017. Current situation and trends. Terveyden ja hyvinvoinnin laitos (THL) National Institute for Health and Welfare (THL); Report 5/2018. Helsinki 2018:2018.
23. Hair J, Black W, Babin B, Anderson R, Tatham R. Multivariate Data Analysis. New Jersey: Pearson Educational Inc; 2006.
24. Goldberg D. The Detection of Psychiatric Illness by Questionnaire: A Technique for the Identification and Assessment of Non-psychotic Psychiatric Illness. Oxford, London: Oxford U Press; 1972.
25. Graetz B. Multidimensional properties of the General Health Questionnaire. Soc Psychiatry Psychiatr Epidemiol 1991 May;26(3):132-138. [doi: 10.1007/bf00782952] [Medline: 1887291]
26. Penninkilampi-Kerola V, Miettunen J, Ebeling H. A comparative assessment of the factor structures and psychometric properties of the GHQ-12 and the GHQ-20 based on data from a Finnish population-based sample. Scand J Psychol 2006 Oct;47(5):431-440. [doi: 10.1111/j.1467-9450.2006.00551.x] [Medline: 16987212]

XSL•FO
RenderX

27. Aalto A, Heponiemi T, Josefsson K, Arffman M, Elovainio M. Social relationships in physicians' work moderate relationship between workload and wellbeing-9-year follow-up study. Eur J Public Health 2018 Oct 1;28(5):798-804. [doi: 10.1093/eurpub/ckx232] [Medline: 29365062]
28. Heponiemi T, Aalto A, Puttonen S, Vänskä J, Elovainio M. Work-related stress, job resources, and well-being among psychiatrists and other medical specialists in Finland. Psychiatr Serv 2014 Jun 1;65(6):796-801. [doi: 10.1176/appi.ps.201300200] [Medline: 24585088]
29. Hyppönen H, Kaipio J, Heponiemi T, Lääveri T, Aalto A, Vänskä J, et al. Developing the national usability-focused health information system scale for physicians: validation study. J Med Internet Res 2019 May 16;21(5):e12875 [FREE Full text] [doi: 10.2196/12875] [Medline: 31099336]
30. Sabes-Figuera R, Maghiros I. European Commission. 2013. European Hospital Survey - Benchmarking Deployment of eHealth services (2012-2013) URL: https://ec.europa.eu/digital-single-market/en/news/european-hospital-survey-benchmarking-deployment-ehealth-services-2012-2013 [accessed 2019-09-26]
31. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, et al. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assoc 2019 Feb 1;26(2):106-114. [doi: 10.1093/jamia/ocy145] [Medline: 30517663]
32. Kuo K, Liu C, Talley PC, Pan S. Strategic improvement for quality and satisfaction of hospital information systems. J Healthc Eng 2018;2018:3689618 [FREE Full text] [doi: 10.1155/2018/3689618] [Medline: 30298099]
33. Liberati EG, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci 2017 Sep 15;12(1):113 [FREE Full text] [doi: 10.1186/s13012-017-0644-2] [Medline: 28915822]
34. Amoakoh-Coleman M, Borgstein AB, Sondaal SF, Grobbee DE, Miltenburg AS, Verwijs M, et al. Effectiveness of mHealth interventions targeting health care workers to improve pregnancy outcomes in low- and middle-income countries: a systematic review. J Med Internet Res 2016 Aug 19;18(8):e226 [FREE Full text] [doi: 10.2196/jmir.5533] [Medline: 27543152]
35. Bruggeman-Everts FZ, Wolvers MD, van de Schoot R, Vollenbroek-Hutten MM, van der Lee ML. Effectiveness of two web-based interventions for chronic cancer-related fatigue compared to an active control condition: results of the "Fitter na kanker" randomized controlled trial. J Med Internet Res 2017 Oct 19;19(10):e336 [FREE Full text] [doi: 10.2196/jmir.7180] [Medline: 29051138]
36. Mathews A, Marc D. Usability evaluation of laboratory information systems. J Pathol Inform 2017;8:40 [FREE Full text] [doi: 10.4103/jpi.jpi_24_17] [Medline: 29114434]
37. Ludwick DA, Doucette J. Primary care physicians' experience with electronic medical records: barriers to implementation in a fee-for-service environment. Int J Telemed Appl 2009;2009:853524 [FREE Full text] [doi: 10.1155/2009/853524] [Medline: 19081787]
38. Simon SR, Kaushal R, Cleary PD, Jenter CA, Volk LA, Orav EJ, et al. Physicians and electronic health records: a statewide survey. Arch Intern Med 2007 Mar 12;167(5):507-512. [doi: 10.1001/archinte.167.5.507] [Medline: 17353500]
39. Vehko T, Hyppönen H, Ryhänen M, Tuukkanen J, Ketola E, Heponiemi T. Health information systems and wellbeing ? health professionals? experiences. Fin J EHealth EWelfare 2018 Mar 8;10(1):143-163. [doi: 10.23996/fjhw.65387]
40. Guo U, Chen L, Mehta PH. Electronic health record innovations: helping physicians - One less click at a time. Health Inf Manag 2017 Sep;46(3):140-144. [doi: 10.1177/1833358316689481] [Medline: 28671038]
41. Shaha JS, El-Othmani MM, Saleh JK, Bozic KJ, Wright J, Tokish JM, et al. The growing gap in electronic medical record satisfaction between clinicians and information technology professionals: issues of most concern and suggested remediations. J Bone Joint Surg Am 2015 Dec 2;97(23):1979-1984. [doi: 10.2106/JBJS.N.01118] [Medline: 26632000]
42. Contratto E, Romp K, Estrada CA, Agne A, Willett LL. Physician order entry clerical support improves physician satisfaction and productivity. South Med J 2017 May;110(5):363-368. [doi: 10.14423/SMJ.0000000000000645] [Medline: 28464179]

Abbreviations

EHR: electronic health record
GHQ: General Health Questionnaire
IS: information systems
IT: information technology
SRIS: stress related to information systems

XSL•FO
RenderX

Edited by G Eysenbach; submitted 23.01.19; peer-reviewed by D Marc, L Moja, KM Kuo; comments to author 01.05.19; revised version received 14.06.19; accepted 31.08.19; published 05.11.19.

Please cite as:
Heponiemi T, Kujala S, Vainiomäki S, Vehko T, Lääveri T, Vänskä J, Ketola E, Puttonen S, Hyppönen H
Usability Factors Associated With Physicians' Distress and Information System–Related Stress: Cross-Sectional Survey
JMIR Med Inform 2019;7(4):e13466
URL: http://medinform.jmir.org/2019/4/e13466/
doi:10.2196/13466
PMID:31687938

©Tarja Heponiemi, Sari Kujala, Suvi Vainiomäki, Tuulikki Vehko, Tinja Lääveri, Jukka Vänskä, Eeva Ketola, Sampsa Puttonen, Hannele Hyppönen. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 05.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Cohort Selection for Clinical Trials From Longitudinal Patient Records: Text Mining Approach

Irena Spasic[1], PhD; Dominik Krzeminski[2], BSc; Padraig Corcoran[1], PhD; Alexander Balinsky[3], PhD

[1]School of Computer Science & Informatics, Cardiff University, Cardiff, United Kingdom

[2]School of Psychology, Cardiff University, Cardiff, United Kingdom

[3]School of Mathematics, Cardiff University, Cardiff, United Kingdom

Corresponding Author:
Irena Spasic, PhD
School of Computer Science & Informatics
Cardiff University
5 The Parade
Cardiff, CF24 3AA
United Kingdom
Phone: 44 02920870320
Email: spasici@cardiff.ac.uk

Abstract

Background: Clinical trials are an important step in introducing new interventions into clinical practice by generating data on their safety and efficacy. Clinical trials need to ensure that participants are similar so that the findings can be attributed to the interventions studied and not to some other factors. Therefore, each clinical trial defines eligibility criteria, which describe characteristics that must be shared by the participants. Unfortunately, the complexities of eligibility criteria may not allow them to be translated directly into readily executable database queries. Instead, they may require careful analysis of the narrative sections of medical records. Manual screening of medical records is time consuming, thus negatively affecting the timeliness of the recruitment process.

Objective: Track 1 of the 2018 National Natural Language Processing Clinical Challenge focused on the task of cohort selection for clinical trials, aiming to answer the following question: Can natural language processing be applied to narrative medical records to identify patients who meet eligibility criteria for clinical trials? The task required the participating systems to analyze longitudinal patient records to determine if the corresponding patients met the given eligibility criteria. We aimed to describe a system developed to address this task.

Methods: Our system consisted of 13 classifiers, one for each eligibility criterion. All classifiers used a bag-of-words document representation model. To prevent the loss of relevant contextual information associated with such representation, a pattern-matching approach was used to extract context-sensitive features. They were embedded back into the text as lexically distinguishable tokens, which were consequently featured in the bag-of-words representation. Supervised machine learning was chosen wherever a sufficient number of both positive and negative instances was available to learn from. A rule-based approach focusing on a small set of relevant features was chosen for the remaining criteria.

Results: The system was evaluated using microaveraged F measure. Overall, 4 machine algorithms, including support vector machine, logistic regression, naïve Bayesian classifier, and gradient tree boosting (GTB), were evaluated on the training data using 10–fold cross-validation. Overall, GTB demonstrated the most consistent performance. Its performance peaked when oversampling was used to balance the training data. The final evaluation was performed on previously unseen test data. On average, the F measure of 89.04% was comparable to 3 of the top ranked performances in the shared task (91.11%, 90.28%, and 90.21%). With an F measure of 88.14%, we significantly outperformed these systems (81.03%, 78.50%, and 70.81%) in identifying patients with advanced coronary artery disease.

Conclusions: The holdout evaluation provides evidence that our system was able to identify eligible patients for the given clinical trial with high accuracy. Our approach demonstrates how rule-based knowledge infusion can improve the performance of machine learning algorithms even when trained on a relatively small dataset.

(JMIR Med Inform 2019;7(4):e15980) doi:10.2196/15980

XSL•FO
RenderX

KEYWORDS

natural language processing; machine learning; electronic medical records; clinical trial; eligibility determination

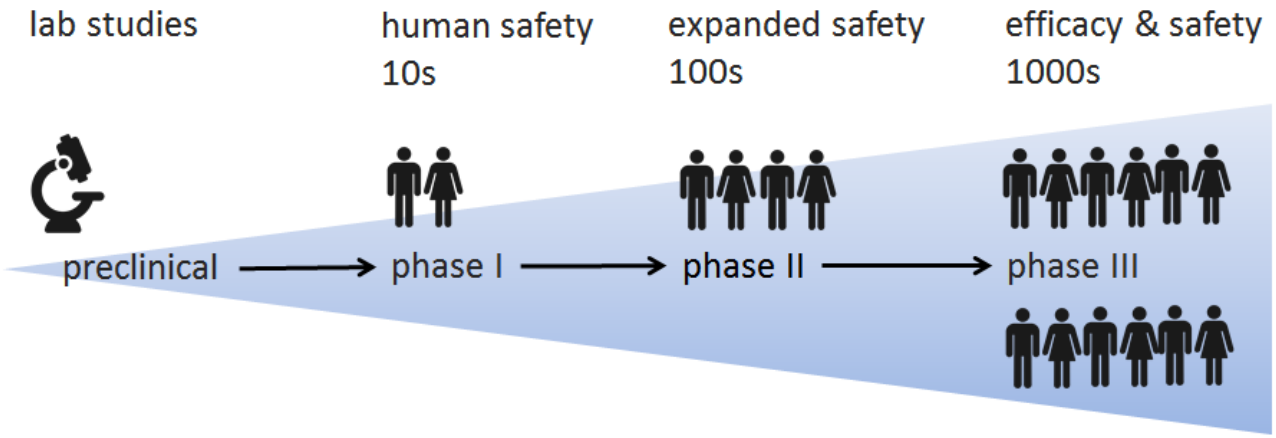
Introduction

Background

Clinical trials are medical research studies focusing on a specific health intervention. They involve human participants to generate data on safety and efficacy as any new health intervention needs to comply with the Hippocratic Oath: "First, do no harm!" With this principle in mind, clinical trials leading up to regulatory approval are typically divided into 3 phases, each involving a significantly higher number of patients (see Figure 1). Phase I aims to answer the following question: Is the intervention safe? The first few healthy participants are given very low doses of the treatment and are monitored closely. If there are no major side effects, the dose is iteratively increased until an effective dose whose possible side effects that are deemed acceptable is reached. Phase II involves patients to determine whether the new intervention works or not. In other words, it assesses its efficacy while continually monitoring the side effects. Finally, in addition to safety and efficacy, phase III also tests the efficiency of the intervention by comparing it with other available interventions. When introducing control groups of participants, clinical trials need to ensure that they are as similar as possible to be able to attribute any findings to the interventions studied and not some other factors. Therefore, each clinical trial defines the eligibility criteria that describe characteristics that must be shared by all participants.

Patient recruitment is universally recognized as a key determinant of success for clinical trials, yet they commonly fail to reach their recruitment goals [1]. Almost a fifth of trials were terminated because they failed to recruit enough participants [2], with less than one-fifth managing to reach their recruitment targets within the proposed time frames [3]. Eligibility represents a major clinical domain barrier to participation [4]. The eligibility criteria are often criticized for being too narrow, thus having a negative impact on recruitment rates and also the generalizability of findings. A stakeholder survey of the barriers to patient recruitment and possible solutions revealed identification of eligible patients using medical records and hospital-based registries and other databases as the key method to improve recruitment [5]. Unfortunately, the complexities of the eligibility criteria do not allow them to be translated directly into readily executable database queries. Instead, they require careful analysis of information contained in the narrative sections of medical records. Manual screening of medical records is time consuming, thus negatively affecting the timeliness of the recruitment process. Text mining has a potential to provide a technical means for unclogging this bottleneck.

Figure 1. The three premarketing phases of a clinical trial.



Related Work

The problem of matching the eligibility criteria against their electronic medical records (EMRs) can be framed using a variety of natural language processing (NLP) tasks depending on the type and level of automation expected. In the context of decision making, automation can be applied to 4 classes of functions: information acquisition, information analysis, decision selection, and decision implementation [6]. In our scenario, we focused on a clinician as a human operator who, given a collection of EMRs and a set of eligibility criteria, needs to decide which patients should be recruited to a given clinical trial. In this context, we can think of information acquisition as identification of information relevant to the eligibility criteria. This task can be automated by means of information retrieval (IR) or information extraction (IE).

IR can be applied to both structured and unstructured components of the EMRs to retrieve relevant records or their parts. The usability of any IR system depends on two key factors: system effectiveness and user utility [7]. A test collection of 56 topics based on patient statements (eg, signs, symptoms, and treatment) and inclusion/exclusion criteria (eg, patient's demographics, laboratory test, and diagnosis) can be used to evaluate the effectiveness of IR for cohort selection [8]. The utility of IR systems can be improved by designing an intuitive visual query interface easily used by clinical researchers [9]. Both utility and effectiveness depend on how well the system incorporates domain-specific knowledge. An ontology

XSL•FO
RenderX

can be used to support term disambiguation, term normalization, and subsumption reasoning. Most studies mapped textual elements to concepts in the Unified Medical Language System (UMLS) for normalization with few studies discussing the use of semantic Web technologies for phenotyping [10]. For instance, the UMLS hierarchy can be used to expand a query searching for cancer to other related terms (eg, neuroblastoma and glioma). However, using such a broad hierarchy for unsupervised expansion can introduce many irrelevant terms, which can be detrimental to eligibility-screening performance [11]. This problem can be reduced by using the UMLS to bootstrap creation of custom ontologies relevant to the problem at hand. For example, to identify patients with cerebral aneurysms, a domain-specific ontology was created by querying the UMLS for concepts related to the locations of aneurysms (eg, middle cerebral artery or anterior communicating artery), other clinical phenotypes related to cerebral aneurysms (eg, saccular aneurysm or subarachnoid hemorrhage), associated conditions (eg, polycystic kidney disease), and competing diagnoses (eg, arteriovenous malformation) [12]. Where available, other relevant systems can be used to inform the development of domain-specific ontologies. For instance, the Epilepsy Data Extraction and Annotation uses a novel Epilepsy and Seizure Ontology, which is based on the International League Against Epilepsy classification system as the core knowledge resource [9].

The complexity of clinical sublanguage may require new language modeling approaches to be able to formulate multilayered queries and customize the level of linguistic granularity [13]. This approach to IR incorporates the output of other NLP systems to represent a document or a query using multiple aligned layers consisting of tokens, their part of speech, named entities with mappings to external knowledge sources, and syntactic dependencies among these elements. Other IR efforts focused on directing a clinician's attention toward specific sentences that are relevant for eligibility determination [14]. This is achieved by segmenting the natural language description of eligibility criteria into individual sentences, analyzing them further to identify domain-specific concepts, and using them to identify sentences in the EMRs that make references to these concepts. This approach is designed to work with categorical data but falls short when numerical data need to be interpreted. For instance, 5 numerical values are needed to diagnose a metabolic syndrome [15]. Of these values, 3 (triglycerides, high-density lipoprotein cholesterol, and elevated fasting glucose) are stored in the laboratory information system, and as structured data are readily available for querying and comparison with referent values. However, in some systems, 2 values may be hidden in the narrative notes (elevated waist circumference and elevated blood pressure). Traditionally, IR approaches are based on the bag-of-words (BoW) model, which represents each document as an unordered collection of features that correspond to the words in a vocabulary for a given document collection. Therefore, by design, IR approaches will be ineffective when it comes to dealing with continuous variables. Conversely, IE based on simple regular expressions can be used to extract numerical values from text and make them amenable for further analysis and interpretation [15-18].

However, the technical feasibility of the IE process does not mean that all relevant attributes are necessarily documented in a single source as the previous example illustrates. For example, a study on case-finding algorithms for hepatocellular cancer discovered significant differences in performance between 2 types of documents (pathology and radiology reports) [19]. It also revealed a significant difference between the narrative reports and coded fields. This raises an important aspect of the completeness of information recorded in an EMR [15]. It has been established that case finding by the International Classification of Diseases, Ninth Revision (ICD-9) coding alone is not sufficient to reliably identify patients with a particular disease or risk factors [20-22]. A few studies contrasted the utility of structured and unstructured information, with the NLP approaches usually demonstrating better results [19,23-28]. In particular, the use of ICD-9 codes for patient phenotyping demonstrated markedly lower precision (or positive predictive value) [19,24,26]. This finding is compatible with a hypothesis that ICD-9 codes are designed for billing purposes and as such may not capture the nuances of phenotypic characteristics in terms of information completeness, expressiveness, and granularity [23].

The analysis of the strengths and weaknesses of both data sources together with practical experiments has led to a consensus that clinical narratives should be used in combination with structured data for eligibility screening [19,23,25,26,28]. Therefore, data fusion is a key component of the information acquisition step in eligibility screening. It should by no means be limited to these 2 modalities of data. For example, clinical electroencephalography (EEG) is the most important investigation in the diagnosis and management of epilepsies. A multimodal patient cohort retrieval system has been designed to leverage the heterogeneous nature of EEG data by integrating EEG reports with EEG signal data [29]. Though evidently important, data fusion techniques are beyond the scope of this study. Here, we focused exclusively on reviewing the methods used to mine clinical narratives for the purpose of eligibility screening. However, the awareness of the need for data fusion can help the reader realize the existence of an externally imposed upper bound on expected performance of text mining approaches.

We have thus far discussed the role of IR and IE in the context of information acquisition. The clinician is still expected to review the retrieved information to decide who satisfies the eligibility criteria. Text mining can be used to support this process by automating information analysis and decision selection by means of feature extraction and text classification, respectively. Two NLP systems tailored to the clinical domain are most often used to extract rich linguistic and semantic features from the narrative found in EMRs: Medical Language Extraction and Encoding (MedLEE [30]) [16,23,25] and clinical Text Analysis and Knowledge Extraction System (cTAKES [31]) [9,11,12,16,18,19,32,33]. They model the semantics by mapping text to the UMLS or a custom dictionary if required. Clinical text analysis needs to make fine-grained semantic distinctions as medical concepts may be negated, may describe someone other than the patient, and may be referring to time other than the present [13]. MedLEE and cTAKES can not only

XSL•FO
RenderX

identify concepts of interest but can also interpret their meaning in the context of negation, hedging, and specific sections. Both systems can also perform syntactic analysis to extract linguistic features such as part of speech and syntactic dependencies. Abbreviations are some of the most prominent features of clinical narratives. Unfortunately, both MedLEE and cTAKES demonstrated suboptimal performance in abbreviation recognition [34], which may require development of bespoke solutions [16,35].

Once the pertinent features have been extracted, they can be exploited by rule-based or machine learning approaches. A review of approaches to identifying patient cohorts using EMRs revealed that out of 97 studies, 24 described rule-based systems; 41 used statistical analyses, data mining, or machine learning; and 22 described hybrid systems [10]. A minimal set of rules is sufficient to accurately extract highly standardized information from the narratives [15]. Their development requires iterative consultation with a clinical expert [26]. Nonetheless, a well-designed rule-based system can achieve good performance on cohort selection even with a small training dataset [36], which remains a problem associated with supervised machine learning approaches. When relevant concepts can be accurately identified from clinical text, both rule-based and machine learning approaches demonstrate good performance, albeit it is slightly in favor of machine learning [25,33].

A variety of supervised machine learning approaches have been used to support cohort selection, including support vector machines (SVMs) [22,25], decision trees [22], Repeated Incremental Pruning to Produce Error Reduction, random forests [25], C4.5 [33], logistic regression (LR) [25,28], naïve Bayesian (NB) learning [22,37], perceptron [37], conditional random fields [19], and deep learning [29,38]. Unfortunately, few studies report systematic evaluation of a wide range of machine learning algorithms, thus offering little insight into the optimal performance of machine learning for cohort selection [39]. Another issue associated with supervised learning is that of imbalanced data. The number of positive examples will typically vary significantly across the eligibility criteria. The data used for the 2018 National Natural Language Processing Clinical Challenge (n2c2) shared task on cohort selection for clinical trials provide a perfect illustration of this problem [18,36,38]. Yet, few approaches tackled this issue with different sampling approaches. Instead, they may resort to using machine learning approaches generally perceived to be the most robust for imbalanced data, for example, SVMs [40,41].

Our review of related work illustrates the ways in which the eligibility screening process can be automated. One study reported that the time for cohort identification was reduced significantly from a few weeks to a few seconds [16]. Others reported the workload reduction with automated eligibility screening around 90% [42] achieved a 450% increase in trial screening efficiency [11]. Most recently, the patient screening time was reduced by 34%, allowing for the saved time to be redirected to activities that further streamlined teamwork among the clinical research coordinators [43]. The same study showed that the numbers of subjects screened, approached, and enrolled were increased by 14.7%, 11.1%, and 11.1%, respectively. In this study, we aimed to illustrate the complexity of the eligibility screening problem and propose a way in which this task can be automated.

Methods

System Overview

In this paper, we describe Cardiff Cohort Selection System (c2s2) [44], an open-source NLP system that, given a longitudinal patient record, performs binary classification against 13 eligibility criteria for a clinical trial. For each criterion in turn, the system determines whether a patient meets or does not meet a given criterion. The eligibility criteria were predefined by the organizers of the 2018 n2c2 shared task (see Table 1) that aimed to answer the question whether NLP systems can use narrative medical records to identify patients eligible for clinical trials.

For the majority of criteria, a record needs to contain the supporting evidence for the corresponding patient to meet a given criterion, otherwise the criterion is considered *not met* (eg, if glycated hemoglobin [HbA_{1c}] value is 4.7 or missing, then the criterion HBA_{1c} is not met). The only 2 exceptions are the criteria concerning a patient's ability to speak English and make their own medical decisions, which are assumed to be *met*, that is, the evidence of the contrary needs to be identified to overturn this assumption. Our system is designed to find and tag such evidence in text using a rule-based approach. A text classifier was trained on the tagged text for each criterion that had a sufficient number of both positive and negative representatives to learn from. Overall, the system consists of 5 modules whose functionality is outlined in Figure 2.

The input to the system is a longitudinal patient record distributed as a single UTF-encoded text file, which contains multiple records generated across various health care encounters. Each individual record represents either a discharge summary or a correspondence between health care professionals [45,46]. Their content may cover patient demographics, progress notes, problems, prescribed medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports. Individual records start with a line formatted as *Record date: YYYY-MM-DD* and are arranged in the ascending order by the record date. Other than that, there are no other restrictions on the format of individual records. Indeed, they may reflect a variety of different styles.

XSL•FO
RenderX

Table 1. Description of the eligibility criteria, as provided in the annotation guidelines used for the National Natural Language Processing Clinical Challenge shared task.

ID	Criterion	Time period	Default
ABDOMINAL	Intra-abdominal surgery, small or large intestine resection, or small bowel obstruction	Any	Not met
ADVANCED-CAD	Advanced artery disease	Present	Not met
ALCOHOL-ABUSE	Alcohol use exceeds weekly recommended limits	Present	Not met
ASP-FOR-MI	Use of aspirin to prevent myocardial infarction	Any	Not met
CREATININE	Serum creatinine is above the upper limit of normal	Any	Not met
DIETSUPP-2MOS	Use of dietary supplements (excluding vitamin D)	Past 2 months	Not met
DRUG-ABUSE	Drug abuse	Any	Not met
ENGLISH	Speaks English	Any	Met
HBA_{1c}	Glycated hemoglobin value is between 6.5 and 9.5	Any	Not met
KETO-1YR	Diagnosed with ketoacidosis	Past year	Not met
MAJOR-DIABETES	Major diabetes-related complication	Any	Not met
MAKES-DECISIONS	Able to make decisions for themselves	Present	Met
MI-6MOS	Myocardial infarction	Past 6 months	Not met

Figure 2. System architecture.

Preprocessing

In addition to standard preprocessing operations (see Figure 2), special consideration is given to punctuation. Its use in clinical narratives proved to affect the results of text segmentation algorithms developed for general language [47]. On one hand, clinical narratives commonly use punctuation as means of abbreviation (see Table 2 for examples). Such use of punctuation may easily be misinterpreted as a sentence terminator. For instance, phrases such as “q. Sunday,” “vit. D,” and “Dr. Harold Nutter” feature a period followed by an uppercase letter, a pattern that is commonly exploited in both rule-based and machine learning approaches to split sentences. Segmentation errors can propagate onto the subsequent stages of text processing, resulting in the loss of syntactic dependencies between related words and consequently their contribution to

XSL•FO
RenderX

the overall semantics. For example, incorrectly splitting a sentence within the phrase "vit. D" would effectively erase this mention of "vitamin D," a named entity of direct relevance to the eligibility criterion DIETSUPP-2MOS (see Table 1). To prevent parsing errors of this type, pattern-matching rules were developed to identify and remove punctuation used in such contexts before performing sentence segmentation (see Table 2 for examples).

Table 2. A selection of rule-based punctuation removal examples.

Rule target	Input	Output
Prescription	• q. a.m. • q. Sunday • tab.	• qam • q Sunday • tab
Vitamin	• vit. D • MVit.	• vit D • MVit
Personal title	• Dr. Harold Nutter • Harold Nutter, Ph.D.	• Dr Harold Nutter • Harold Nutter, PhD
Shorthand x	• hx. of migraines • sx. of depression • Rx. for cpap	• hx of migraines • sx of depression • Rx for cpap
Species name	• E. coli • C. diff • H. pylori	• E coli • C diff • H pylori

Clinical narratives also feature prevalent use of short formulaic statements such as field:value combinations (eg, *Substance abuse: none*) and itemized lists (see Textbox 1 for an example).

Such statements are not commonly terminated by means of punctuation. When used consecutively, this can often result in independent statements being incorrectly grouped together in a single sentence. Their intersentential co-occurrence may later be easily confused with relatedness. Consider, for instance, amalgamating the above itemized list into a continuous sequence "*s/p cerebral infarction myocardial scan normal blood pressure today 190/108.*" It could lead to incorrectly recognizing *infarction* as a *myocardial* one and the *blood pressure* as *normal*, when in fact, the *infarction* is *cerebral*, and the *blood pressure* is *abnormally high*. Acting preemptively, we perform document layout analysis to identify itemized lists and insert punctuation where appropriate before performing sentence segmentation. Consequently, this will enforce independent fragments to be interpreted as separate sentences.

Finally, to streamline subsequent text analysis, we use pattern-matching rules to fully expand enclitics and special characters. For example, *couldn't* is expanded to *could not*, whereas *con't* is expanded to *continue*. This will later simplify identification of negated expressions. Similarly, to prune the number of IE rules, we lexicalized a relevant set of special characters. For example, *BUN/Cr ratio is >20* would become *BUN/Cr ratio is greater than 20*.

Textbox 1. An example of assessment recorded as an itemized list.

- *s/p cerebral infarction*
- *myocardial scan normal*
- *blood pressure today 190/108*

Normalization

Text normalization is performed with a similar intent: to simplify subsequent text analysis. It involves mapping of a selected subset of words and phrases onto their representatives, which can be either a preferred synonym or a hypernym (see Table 3 for examples). Special consideration is given to acronyms and abbreviations as they are known to have a major impact on retrieval of relevant information. First, disambiguation is performed for a small subset of abbreviations of direct relevance for the given classification tasks. Examples include *ca* (*calcium* vs *cancer*), *mg* (*magnesium* vs *milligram*), and *CR* (*creatinine* vs *controlled release*). A context-sensitive approach is used to select an appropriate interpretation. For example, if CR is used in combination with words such as *tablet* or *capsule*, then it is assumed to refer to *controlled release*.

XSL•FO
RenderX

Table 3. Examples of text normalization.

Example	Surface forms	Normalized form	Relevance
1	mom, father, sister	family member	filtering
2	FH, FHx, FamHx	family history	filtering
3	whiskey, vodka, beer	alcohol	ALCOHOL-ABUSE
4	Lantus, Humalog, NPH	insulin	MAJOR-DIABETES
5	DM2, DMII, NIDDM	diabetes mellitus 2	MAJOR-DIABETES
6	CRRT, CRRTX	continuous renal replacement therapy	MAJOR-DIABETES
7	ARF	acute renal failure	MAJOR-DIABETES
8	CKD	chronic kidney disease	MAJOR-DIABETES
9	BB, bblocker, betablocker	beta blocker	ADVANCED-CAD
10	ECG, EKG	electrocardiogram	ADVANCED-CAD
11	ICD	implantable cardioverter defibrillator	ADVANCED-CAD
12	CVD	cardiovascular disease	ADVANCED-CAD
13	MI, heart attack	myocardial infarction	MI-6MOS, ASP-FOR-MI, ADVANCED-CAD
14	STEMI	ST elevation myocardial infarction	MI-6MOS, ASP-FOR-MI, ADVANCED-CAD
15	ASA, ECASA	aspirin	ASP-FOR-MI

Other acronyms and abbreviations of interest are then expanded using a bespoke lexicon (>500 entries) developed specifically for this task. To bootstrap the lexicon construction, the raw training data were used to analyze frequently occurring words. Orthographic features (uppercase typeset, eg, *STEMI*, or the use of punctuation, eg, *q.a.m.* or *r/o*) and spelling checker (eg, *inpt*) were used to identify potential acronyms and abbreviations as *unknown* words that are also relatively short. Medical expertise was used to identify the corresponding full forms. Simple Concordance Program [48] was used to verify manually whether the proposed full forms apply across the majority of contexts within the training data to enable the use a context-free approach for acronym and abbreviation expansion.

The only acronym exempt from expansion was *CCB*. In fact, all occurrences of *calcium channel blocker* were replaced by the corresponding acronym. The reason behind this decision is the fact that both *calcium* as a supplement and *calcium channel blocker* often occur in similar context (eg, medication list). As one of the eligibility criteria was concerned with dietary supplementation (see DIETSUPP-2MOS in Table 1), this reduced the risk of interpreting the latter mention of *calcium* as a supplement.

To illustrate the extent to which text normalization can simplify its subsequent analysis, we can use examples provided in Table 3. For example, by replacing the surface forms in Example 1 by their hypernym and expanding abbreviations in Example 2, we can simply use the occurrence of the word *family* to filter out sentences or the whole sections that refer to family members. Consider, for example, the original text given in Textbox 2 and its normalized counterpart in Textbox 3.

Textbox 2. An original example of family history.

FH: Mom w/ PM at age 50, died of MI at 71. Father w/ EtOH, HTN. Sister w/ 4 miscarriages.

Textbox 3. A normalized example of family history.

Family history: Family member with pacemaker at age 50, died of myocardial infarction at 71. Family member with alcohol abuse, hypertension. Family member with 4 miscarriages.

By filtering out references to family members, we are effectively removing the mentions of *myocardial infarction* and *alcohol abuse* that do not apply to the given patient. Consequently, we can use the remaining references to *myocardial infarction* and *alcohol abuse*, if any, as evidence for eligibility criteria MI-6MOS and ALCOHOL-ABUSE (see Table 1). Similarly, by mapping alcoholic beverages in Example 3 to their hypernym, the subsequent analysis related to the eligibility criterion ALCOHOL-ABUSE (see Table 1) can simply focus on any mention of the word *alcohol*. Examples 4 and 5 show that 2 keywords, *insulin* and *diabetes*, can be used to look for evidence of diabetes. Once unpacked from the corresponding acronyms (Example 5), the word *diabetes* becomes accessible to text analysis. Similarly, words *renal* and *kidney* become visible after expanding acronyms in Examples 5-7. Knowing that diabetes is a major risk factor for kidney disease, we can subsequently use close occurrences of the word *diabetes* to either of the words *renal* or *kidney* as evidence for the eligibility

XSL•FO
RenderX

criterion MAJOR-DIABETES (see Table 1). Similar to lexical analysis, morphological analysis can be used to identify features relevant to the given eligibility criteria. Normalized forms in Examples 10-14 related to ADVANCED-CAD (see Table 1) incorporate a morpheme *cardi(o)*, which signifies that these medical concepts are related to the heart, which can be affected by coronary artery disease.

Filtering

Once the text has been regularized by means of preprocessing and normalization, information not directly relevant to the given classification tasks is filtered out. We focus on 4 types of such information:

1. negation, for example, *ruled out for MI by enzymes*
2. family history, for example, *mother died at age 62 of a heart attack*
3. allergies, for example, *Allergies: aspirin—GI upset*
4. time window, for example, *records older than the last 6 months*

Removal of such information simplifies subsequent classification by allowing the use of a BoW approach. For example, by not considering the first 2 examples, the risk of misclassifying a patient as having a *myocardial infarction* is reduced. Similarly, by removing the third example from consideration, the risk of misclassifying a patient as one taking *aspirin* to prevent *myocardial infarction* is also reduced. Finally, as some of the eligibility criteria were time dependent (namely, ALCOHOL-ABUSE, DIETSUPP-2MOS, KETO-1YR, MAKES-DECISIONS, and MI-6MOS—see Table 1 for definitions), we identified dates of individual medical records to extract the ones relevant to the given time windows and stored them separately for use by the corresponding classifiers.

We used a set of regular expressions, which are available from the c2s2 GitHub repository [42], to identify the 4 types of information considered. Regular expressions used to identify negation are based on the NegEx algorithm for identifying negated concepts in clinical notes [49].

Feature Extraction

Thus far, we reduced the noise and lexical variability in the data by means of filtering and normalization. This is expected to improve the performance of a supervised classifier. Another action that stands to improve the classification performance when trained on a relatively small dataset is that of reducing dimensionality of a BoW representation by aggregating related features into a single representative. In its simplest form, feature aggregation can be achieved by abstracting words into semantic classes. Where domain ontology is available, such abstraction can be automated by exploiting its taxonomic structure. The Semantic Network of the UMLS can be used to automatically abstract words into semantic types. However, as examples given in Table 4 illustrate, the UMLS semantic types are too broad in the context of eligibility criteria described in Table 1. For example, abstracting Examples 1-4 into pharmacologic substance would dilute rather than distil relevant information. A finer-grained abstraction tuned for the given eligibility criteria would be more appropriate (see the last 2 columns in Table 4), but it would also incur some knowledge engineering overhead. However, the widespread availability of Web resources that summarize information pertaining to health and well-being can greatly reduce such overhead. We defined a total of 8 abstraction categories and assembled the corresponding lexica using online resources (see Table 5).

Table 4. Examples of word abstraction.

Example	Surface forms	Semantic type	Abstraction	Relevance
1	marijuana, heroin, ecstasy	Pharmacologic substance	Illicit drug	DRUG-ABUSE
2	beta blocker, nitroglycerin, CCB	Pharmacologic substance	Heart medication	ADVANCED-CAD
3	crestor, advicor, compactin	Pharmacologic substance	Statin	ADVANCED-CAD
4	vitamin C, calcium, primrose oil	Pharmacologic substance	Supplement	DIETSUPP-2MOS
5	turmeric, green tea, cinnamon	Food	Supplement	DIETSUPP-2MOS
6	vodka, beer, wine	Food	Alcohol	ALCOHOL-ABUSE

XSL•FO
RenderX

Table 5. Rule-based feature extraction.

Tag	Feature	Extraction[a]	Examples[b]
MEDRX	Prescription instructions	Regular expressions	po q4h prn
KIDMED	Kidney medication	Lexicon (221 entries)[c]	Thymoglobulin
BRPMED	Blood pressure medication	—[d]	Avapro
HRTMED	Heart medication	—	Plavix
HRTTRT	Heart treatment	Regular expressions	Re*catheteriz*ation
HRTISC	Heart ischemia	Regular expressions	Electro*cardio*gram demonstrated *ischemic* changes
HRTANG	Angina	Regular expressions	Chest wall heaviness
HRTCAD	Any of the HRT tags above + explicit references to CAD	Regular expressions	Given his extensive cardiac history
ASPFMI	Aspirin for heart problems	Regular expressions	Start on heparin *HRT*MED and *aspirin* and take to *HRT*TRT catheterization laboratory
SPLMNT	Supplement (strong evidence)	Lexicon (67 entries) + regular expressions	Ibuprofen 800 mg *MED*RX *potassium* chloride 10 meq *MED*RX lasix 20 mg *MED*RX
DFCNCY	Supplement (weak evidence)	Lexicon (27 entries) + regular expressions	*Iron deficien*cy anemia
MNTCAP	Mental capacity	Regular expressions	Increasing *disorientat*ion and visual *hallucinat*ions
DRGADD	Substance abuse	Lexicon (17 entries) + regular expressions	History of *cocaine* abuse
NOENGL	Does not speak English	Lexicon (66 entries) + regular expressions	An *Indonesian speaking* 85-year-old male
ALCABS	Alcohol abuse	Lexicon (7 entries) + regular expressions	*Alcoholism* 10 years ago
ALCSTP	Stopped drinking alcohol	Regular expressions	Alcoholism 10 *years ago*
KETACD	Ketoacidosis	Regular expressions	Ketones positive
KIDDAM	Kidney problems	Regular expressions	*Worse*ning *renal* dys*function*
DMCMPL	Diabetic complications	Regular expressions	*Diabet*es mellitus related retino*pathy*/neuro*pathy*
ABDMNL	Abdominal surgery or small bowel obstruction	Regular expressions	Gastric *lapar*oscopic bypass surgery
HIGHCRT	High creatinine	Regular expressions + information extraction	Blood urea nitrogen/*creatinine* of 21/*1.7*
GLYHMG	Glycated hemoglobin in a given interval	Information extraction	*Hb*A_{1c} one month ago was *6.7*

[a]All lexicons and regular expressions are available from the c2s2 GitHub repository [44].

[b]Italic typeset is used to indicate the types of text features targeted by lexicons and regular expressions.

[c]KIDMED, BRPMED, HRTMED are organized into a single lexicon of 221 entries.

[d]Not applicable.

Once the BoW representation is passed onto a supervised classifier, the context of individual words will be lost. For instance, blood tests frequently feature essential minerals such as calcium, potassium, and iron, which can also be prescribed under the same names as supplements. The BoW approach will take these names out of context, keeping their frequency as the only information about them. Conversely, simple pattern analysis can be used to differentiate between the 2 types of context. For example, we can model prescription instructions using regular expressions (see Table 5) and tag this information in text in the form of a token (eg, MEDRX) that is lexically distinguishable from other tokens. We can subsequently apply another regular expression to find mentions of essential minerals in the close proximity to the MEDRX token and tag such mentions using another special-purpose tag (eg, SPLMNT). When we now apply the BoW approach, the token SPLMNT, treated as any other text token, will represent a feature that preserves relevant contextual information. Supervised machine learning algorithms can then take advantage of such a feature in combination with the standard BoW features. Regular expressions are used to embed a total of 18 context-sensitive features into text (see Table 5).

Regular expressions can be used to model categorical references to information relevant to the given eligibility criteria. For

XSL•FO
RenderX

example, regular expressions can be used to link the word *creatinine* with a stem *elev*-in the phrase *a mildly elevated creatinine* and use it as an indication for meeting the eligibility criterion CREATININE (see Table 1). However, knowing whether serum creatinine is above the upper limit of normal in a phrase such as "blood urea nitrogen and creatinine ratio of 40 and 1.0 respectively" requires not only extracting the correct numerical value (1.0) but also comparing it with the reference value (1.5). Two eligibility criteria, CREATININE and HBA_{1c}, require extraction of numerical information and its subsequent analysis, as indicated in Table 5. As before, the outcome of such context-sensitive analysis is embedded back into the text for further exploitation by supervised machine learning.

Overall, a total of 22 tags described in Table 5 were chosen so that they can be lexically and orthographically distinguishable from other words upon their imputation into the processed text. The corresponding features are extracted incrementally in the order given in Table 5 and, when appropriate, used to support extraction of other features. For example, knowing that heparin is a heart medication (indicated by the tag HRTMED—see Table 5) can be used to infer that, when aspirin is taken together with heparin, it is likely to be used as prophylaxis for the prevention of cardiovascular events such as myocardial infarction (indicated by the tag ASPFMI—see Table 5).

Classification

This module consists of 13 binary classifiers, 1 for each eligibility criterion (see Table 1). The distribution of class labels in the training data informed the choice of a classification method. Supervised machine learning was chosen wherever a sufficient number of both positive and negative instances were available to learn from (see Figure 3). A rule-based approach focusing on a small set of relevant features was chosen for the remaining criteria (see Table 6). The corresponding classification rules were based on a relevant set of manually engineered features described earlier in Table 5. Each rule was defined as a function of these features and a threshold value that maximizes the class separation, both chosen manually. The only exception was associated with the criterion MI-6MOS, where the final rule was induced from the training data in the form of a decision tree using a manually selected set of features.

Figure 3. Distribution of class labels.

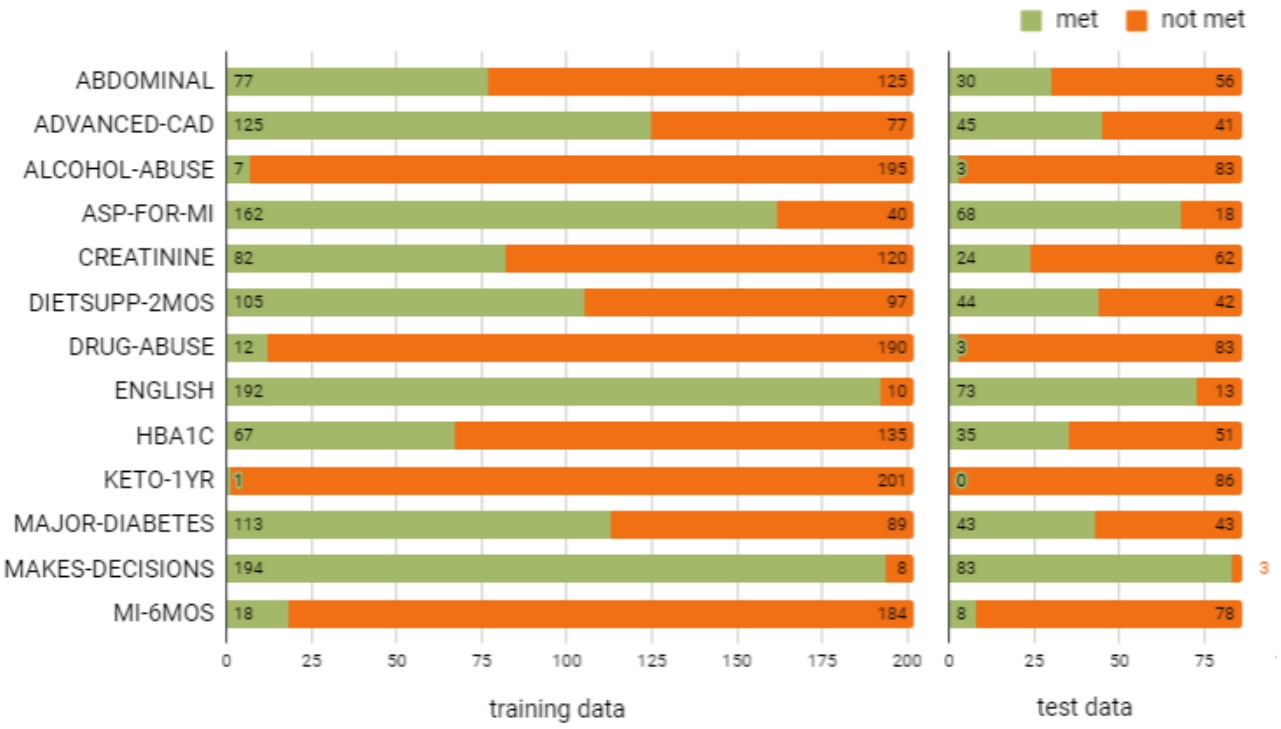


Table 6. Features used in rule-based classification.

ID	Features
ALCOHOL-ABUSE	ALCABS, ALCSTP
DRUG-ABUSE	DRGADD
ENGLISH	NOENGL
KETO-1YR	KETACD
MAKES-DECISIONS	MNTCAP
MI-6MOS	BRPMED, HRTMED, HRTTRT, HRTISC, HRTANG, HRTCAD, ASPFMI

XSL•FO
RenderX

Note that the numerical values used in criteria CREATININE and HBA_{1c} were also extracted using a rule-based approach. However, in a longitudinal report, different values may be reported at different time points. In the absence of clear guidelines, we used machine learning on top of IE to determine automatically from the training data how to deal with such cases.

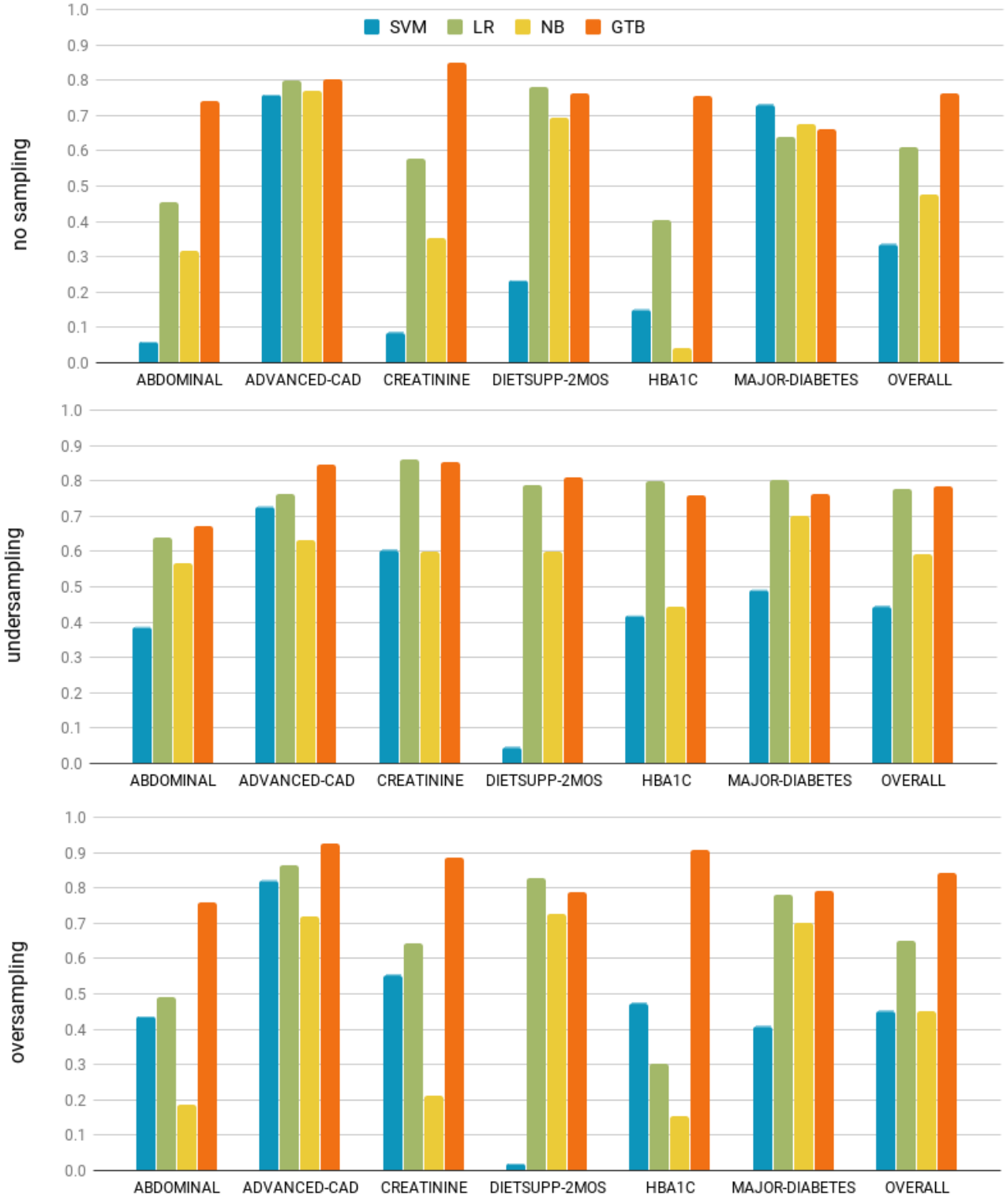
A machine learning approach was used for all other criteria. According to the *no free lunch* theorem [39], there is no universally best learning algorithm. In other words, the performance of machine algorithms depends not only on a specific computational task at hand but also on the properties of the data that characterize the problem. To compare the performance of different algorithms, we used 10–fold cross-validation experiments. We chose a representative algorithm from 4 major categories: function-based learning, regression analysis, probabilistic learning, and ensemble learning. Specific algorithms chosen were SVM with radial basis function kernel, LR, NB classifier, and gradient tree boosting (GTB), respectively. In our experiments, we used implementations of the first 3 algorithms in scikit-learn, an open-source Python library for data analysis and modeling [50]. Experiments with GTB were performed using XGBoost, an open-source software library that implements a gradient boosting framework for Python [51]. All experiments were performed with the default parameter values.

We trained all classifiers using single words and/or bigrams as features with and without feature selection based on L1 regularized linear SVM. The overall performance was statistically indistinguishable across different types of features used. Therefore, we opted for a simple BoW approach with feature selection for efficiency reasons. To evaluate the impact of the class imbalance on the classification performance, we balanced the training data using random undersampling and oversampling with default parameters from scikit-learn [50].

Figure 4 summarizes the performance in terms of microaveraged *F* measure. Overall, GTB demonstrated the most consistent performance. Its performance peaked when oversampling was used to balance the training data. GTB is an ensemble classifier over a set of simple decision trees, which are varied according to specific parameter settings (learning rate and maximum tree depth). Having chosen GTB as the learning method, we optimized its parameters by performing grid search on learning rate (0.001-0.5) and maximum tree depth (2-10) using the oversampled training data. The learning rate of 0.02 and maximum depth of 10 were chosen for the holdout evaluation described in the next section.

XSL•FO
RenderX

Figure 4. Summary of cross-validation results. SVM:support vector machines; LR: logistic regression; NB: naïve Bayesian; GTB: gradient tree boosting; HBA_{1c}:glycated hemoglobin.



Results

The results of classification experiments on previously unseen test data are summarized in Table 7. The evaluation results were calculated using a script released by the organizers of the 2018 n2c2 shared task. We used the best results from 3 related studies as the baseline. They used rule-based [36], hybrid [18], and hierarchical neural network (HNN) [38] approaches. We interpret the results for each classifier separately.

XSL•FO
RenderX

Table 7. Detailed holdout test results.

ID	Met[a]			Not met[a]			Overall	Baseline[b]		c2s2[c]
	P[d] (%)	R[e] (%)	F[f] (%)	P (%)	R (%)	F (%)	F (%)	F (%)	System	Rank
ABDOMINAL	64.86	80.00	71.64	87.76	76.79	81.90	76.77	90.64	Rules	4
ADVANCED-CAD	83.02	97.78	89.80	96.97	78.05	86.49	88.14	88.14	c2s2	1
ALCOHOL-ABUSE	22.22	66.67	33.33	98.70	91.57	95.00	64.17	*89.70*	Hybrid	2
ASP-FOR-MI	87.67	94.12	90.78	69.23	50.00	58.06	74.42	77.34	HNN[g]	2
CREATININE	80.00	83.33	81.63	93.44	91.94	92.68	87.16	89.75	Rules	2
DIETSUPP-2MOS	78.85	93.18	85.42	91.18	73.81	81.58	83.50	89.53	Hybrid	4
DRUG-ABUSE	40.00	66.67	50.00	98.77	96.39	97.56	73.78	*92.55*	Hybrid	2
ENGLISH	91.25	100.00	95.42	100.00	46.15	63.16	79.29	97.66	Hybrid	4
HBA_{1c}	100.00	82.86	90.62	89.47	100.00	94.44	92.53	93.82	Rules	2
KETO-1YR	0.00	0.00	0.00	100.00	100.00	100.00	50.00	*50.00*	All	1
MAJOR-DIABETES	85.00	79.07	81.93	80.43	86.05	83.15	82.54	86.02	Hybrid	2
MAKES-DECISIONS	97.62	98.80	98.20	50.00	33.33	40.00	69.10	*74.40*	HNN	2
MI-6MOS	33.33	50.00	40.00	94.59	89.74	92.11	66.05	*87.59*	Rules	4
Overall[h] (microaveraged)	83.97	91.29	87.47	93.54	87.86	90.61	89.04	91.11	Hybrid	4

[a]The binary classification task involves 2 classes (*met* and *not met*). The results are provided for each class separately and then combined into the overall F value.

[b]The best results from 3 related studies are used as the baseline. They are named after the approach they used: rules [34], hybrid [17], and HNN [36]. The baseline results in italics were calculated on the basis of at most eight positive examples, which account for less than 10% of the test data.

[c]c2s2: Cardiff Cohort Selection System.

[d]P: precision.

[e]R: recall.

[f]F: *F* measure.

[g]HNN: hierarchical neural network.

[h]The overall values provided in the bottom row have been microaveraged across the 13 classifiers.

The best results marked with an asterisk in Table 7 were calculated on the basis of at most 8 positive examples, which account for less than 10% of the test data. This makes it impossible to differentiate between random and statistically significant outcomes, thus making it difficult to generalize the findings. The most extreme example is that of KETO-1YR, which had no positive examples in the test data. The results of all 4 systems were identical with no classification errors. Again, given that the training data contained only 1 positive example, the best classification strategy would be the majority rule, which would achieve the same result. Similarly, ALCOHOL-ABUSE, DRUG-ABUSE, and MAKES-DECISIONS had only 3 positive examples in the test data. On these classes, the 4 systems achieved average precision, recall, and *F* measure of 58.43%, 65.46%, and 59.38% with standard deviations of 40.11%, 36.51%, and 37.48%, respectively, again illustrating the difficulty of generalizing these findings. Finally, MI-6MOS had 8 positive examples. The rule-based system achieved the best performance followed by HNN. At 40.00%, the remaining 2 systems achieved a modest *F* measure on the *met* class, but they did differ in the way they balanced precision and recall. Overall, no obvious pattern could be noticed in the classification performance on this class.

All 4 systems achieved similar performance for HBA_{1c} and ASP-FOR-MI. On the *met* class, all 4 systems achieved maximal precision on HBA_{1c} with recall in the 80s, resulting in an *F* measure just below or just above 90%. Conversely, on the *met* class, all 4 systems achieved almost perfect recall on ASP-FOR-MI with precision in the high 80s, resulting in an *F* measure over 90%. Given the consistently high performance, we infer that the 2 eligibility criteria are semantically tractable in the sense that they lend themselves to being modeled computationally.

The rule-based approach performed best against the following eligibility criteria: ABDOMINAL and CREATININE. For ABDOMINAL, recall was in the 80s on the *met* class with no significant variation across the systems. However, the 2 machine learning approaches demonstrated markedly lower precision than the rule-based approach: 60s versus 90s. Further experiments are needed to determine whether more training data would help reduce the number of false positives. In reality, the cost and time associated with data annotation imposes an upper bound on the amount of training data available. Given the *F* measure is in high 80s, rule-based approaches could be a preferred option for narrowly defined eligibility criteria, which can be mapped to explicit references in text. We can observe

XSL•FO
RenderX

similar results for CREATININE. The rule-based approach performed best with an *F* measure in the 80s on the *met* class, followed by our own approach with comparable performance. Although we used machine learning, the key feature used by the classifier was in fact extracted using a rule-based approach. This is consistent with our previous recommendation.

Conversely, broader eligibility criteria, which require some reasoning over multiple references made across the discourse, may require a machine learning approach to model the complexities of target classification problems. MAJOR-DIABETES is one such example where major complications may not be restricted to a finite class of signs and symptoms. In addition, such complications may be mentioned without an explicit reference to diabetes. This requires complex analysis of the wider context. Neural networks can be used to model nonlinearity in text. Not surprisingly, the HNN approach achieved the best results in this case. In particular, the robustness of this approach is reflected in achieving a recall of over 90% on the *met* class. The rule-based approaches demonstrated lower recall. Our own approach demonstrated the lowest recall as we also used a rule-based approach to extract pertinent features. However, our use of machine learning on top of such features resulted in the second highest precision on the *met* class.

Another example of this type of problem is ADVANCED-CAD. As expected, both machine learning approaches performed better than the other 2, with overall *F* measure in the 80s and 70s, respectively. In particular, our approach significantly outperformed all others in both precision and recall (see Table 8). We attribute such a performance to a suitable combination of rule-based feature extraction and supervised classification. By examining Table 5, we can see that the majority of features are related to advanced cardiovascular disease either directly (eg, HRTMED, HRTTRT, HRTISC, HRTANG, HRTCAD, and ASPFMI) or indirectly (eg, BRPMED and DMCMPL). Our approach demonstrates the degree to which domain knowledge infusion can improve the performance of machine learning when trained on a relatively small dataset. However, it does not require comprehensive knowledge elicitation. We simply used online resources and simple corpus analysis to inform the development of the corresponding lexica and regular expressions following the same approach used successfully in previous shared tasks [52,53].

Table 8. Detailed holdout test results for ADVANCED-CAD.

System	Met			Not met			Overall
	P[a] (%)	R[b] (%)	F[c] (%)	P (%)	R (%)	F (%)	F (%)
c2s2[d]	83.02	97.78	89.80	96.97	78.05	86.49	88.14
Hybrid	74.55	91.11	82.00	87.10	65.85	75.00	78.50
Rules	67.80	88.89	76.92	81.48	53.66	64.71	70.81
HNN[e]	77.36	91.11	83.67	87.88	70.73	78.38	81.03

[a]P: precision.

[b]R: recall.

[c]F: *F* measure.

[d]c2s2: Cardiff Cohort Selection System.

[e]HNN: hierarchical neural network.

Discussion

Ideally, supervised learning performs best when large training datasets with a reasonable class balance are available to extrapolate a classification model while minimizing overfitting. As we can see from the data (see Figure 3), this was not the case in this particular study. This is likely to be the norm in practice rather than the exception. When structured data are available to support certain eligibility criteria, there is no need for analyzing the unstructured text data. When such a need does exist, the use of supervised learning requires manual annotation of text data, which requires clinical expertise. The cost and time associated with this activity naturally imposes an upper bound on the amount of training data available. This limited amount of training data will immediately exclude approaches such as deep learning, which, in theory, could be used to extract complex relationships between words using long- and short-term memory. Therefore, the remaining choices include rule-based classification and supervised learning. Clinical trials are plagued by insufficient recruitment rates. On average, 86% of trials fail to recruit a sufficient number of patients, 85% of trials overrun because of insufficient recruitment, 37% of sites do not meet their recruitment targets, and 20% fail to recruit any patients [54]. Even when sufficient numbers are initially recruited, the problem of 30% dropout rate remains. Not surprisingly, 30% of phase III trial terminations are because of recruitment failures. Owing to these recruitment concerns, one would naturally opt for supervised learning approaches as they are more robust than rule-based approaches in terms of recall. In other words, it would help identify a much larger pool of patients to potentially recruit. However, the limited amount of training data will prevent the use of longer n-grams as it would lead to document representation vectors that are long and sparse, a combination prone to overfitting. This leaves the BoW approach as the most plausible option. To compensate for the loss of context, manual feature engineering can be used to model complex relationships between words. This represents a practical compromise between rule-based and machine learning approaches. This study provides a practical example of such a hybrid approach. The

XSL•FO
RenderX

development of our system incurred less than 2 person-months, while achieving performance that could boost the recruitment. The system is expected to reduce clinicians' workload in line with the estimates reported by other studies [11,16,42,43].

Acknowledgments

The authors gratefully thank Nikola Cihoric, MD, for sharing his medical expertise, which partly informed the development of the preprocessing module.

Authors' Contributions

IS designed the system. IS and PC implemented the following modules: preprocessing, normalization, filtering, and feature extraction. DK and AB implemented the classification module. All authors were involved in the evaluation and interpretation of the results. IS drafted the manuscript. All authors reviewed and approved the manuscript for publication.

Conflicts of Interest

None declared.

References

1. Huang G, Bull J, McKee KJ, Mahon E, Harper B, Roberts J, CTTI Recruitment Project Team. Clinical trials recruitment planning: a proposed framework from the Clinical Trials Transformation initiative. Contemp Clin Trials 2018 Mar;66:74-79 [FREE Full text] [doi: 10.1016/j.cct.2018.01.003] [Medline: 29330082]
2. Carlisle B, Kimmelman J, Ramsay T, MacKinnon N. Unsuccessful trial accrual and human subjects protections: an empirical analysis of recently closed trials. Clin Trials 2015 Feb;12(1):77-83 [FREE Full text] [doi: 10.1177/1740774514558307] [Medline: 25475878]
3. Treweek S, Lockhart P, Pitkethly M, Cook J, Kjeldstrøm M, Johansen M, et al. Methods to improve recruitment to randomised controlled trials: Cochrane systematic review and meta-analysis. BMJ Open 2013;3(2):pii: e002360 [FREE Full text] [doi: 10.1136/bmjopen-2012-002360] [Medline: 23396504]
4. Unger J, Vaidya R, Hershman D, Minasian L, Fleury M. Systematic review and meta-analysis of the magnitude of structural, clinical, and physician and patient barriers to cancer clinical trial participation. J Natl Cancer Inst 2019 Mar 1;111(3):245-255 [FREE Full text] [doi: 10.1093/jnci/djy221] [Medline: 30856272]
5. Mahon E, Roberts J, Furlong P, Uhlenbrauck G, Bull J. Barriers to trial recruitment and possible solutions. Applied Clinical Trials 2016;25(2/3):20 [FREE Full text]
6. Parasuraman R, Sheridan T, Wickens C. A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern A 2000;30(3):286-297. [doi: 10.1109/3468.844354]
7. Clough P, Sanderson M. Evaluating the performance of information retrieval systems using test collections. Inform Res 2013;18(2) [FREE Full text]
8. Wang Y, Wen A, Liu S, Hersh W, Bedrick S, Liu H. Test collections for electronic health record-based clinical information retrieval. JAMIA Open 2019:ooz016. [doi: 10.1093/jamiaopen/ooz016]
9. Cui L, Bozorgi A, Lhatoo S, Zhang G, Sahoo S. EpiDEA: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification. AMIA Annu Symp Proc 2012;2012:1191-1200 [FREE Full text] [Medline: 23304396]
10. Shivade C, Raghavan P, Fosler-Lussier E, Embi P, Elhadad N, Johnson S, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014;21(2):221-230 [FREE Full text] [doi: 10.1136/amiajnl-2013-001935] [Medline: 24201027]
11. Ni Y, Kennebeck S, Dexheimer J, McAneney C, Tang H, Lingren T, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc 2015 Jan;22(1):166-178 [FREE Full text] [doi: 10.1136/amiajnl-2014-002887] [Medline: 25030032]
12. Castro V, Dligach D, Finan S, Yu S, Can A, Abd-El-Barr M, et al. Large-scale identification of patients with cerebral aneurysms using natural language processing. Neurology 2017 Jan 10;88(2):164-168 [FREE Full text] [doi: 10.1212/WNL.0000000000003490] [Medline: 27927935]
13. Wu S, Wen A, Wang Y, Liu S, Liu H. Aligned-layer text search in clinical notes. Stud Health Technol Inform 2017;245:629-633. [doi: 10.3233/978-1-61499-830-3-629] [Medline: 29295172]
14. Shivade C, Hebert C, Lopetegui M, de Marneffe MC, Fosler-Lussier E, Lai A. Textual inference for eligibility criteria resolution in clinical trials. J Biomed Inform 2015 Dec(58 Suppl):S211-S218 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.008] [Medline: 26376462]
15. Kreuzthaler M, Schulz S, Berghold A. Secondary use of electronic health records for building cohort studies through top-down information extraction. J Biomed Inform 2015 Feb;53:188-195 [FREE Full text] [doi: 10.1016/j.jbi.2014.10.010] [Medline: 25451102]

XSL•FO
RenderX

16. Chen W, Kowatch R, Lin S, Splaingard M, Huang Y. Interactive cohort identification of sleep disorder patients using natural language processing and i2b2. Appl Clin Inform 2015;6(2):345-363 [FREE Full text] [doi: 10.4338/ACI-2014-11-RA-0106] [Medline: 26171080]
17. Jonnalagadda S, Adupa A, Garg R, Corona-Cox J, Shah S. Text mining of the electronic health record: an information extraction approach for automated identification and subphenotyping of HFpEF patients for clinical trials. J Cardiovasc Transl Res 2017 Jun;10(3):313-321. [doi: 10.1007/s12265-017-9752-2] [Medline: 28585184]
18. Vydiswaran V, Strayhorn A, Zhao X, Robinson P, Agarwal M, Bagazinski E, et al. Hybrid bag of approaches to characterize selection criteria for cohort identification. J Am Med Inform Assoc 2019 Nov 1;26(11):1172-1180. [doi: 10.1093/jamia/ocz079] [Medline: 31197354]
19. Sada Y, Hou J, Richardson P, El-Serag H, Davila J. Validation of case finding algorithms for hepatocellular cancer from administrative data and electronic health records using natural language processing. Med Care 2016 Feb;54(2):e9-14 [FREE Full text] [doi: 10.1097/MLR.0b013e3182a30373] [Medline: 23929403]
20. Birman-Deych E, Waterman A, Yan Y, Nilasena D, Radford M, Gage B. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. Med Care 2005 May;43(5):480-485. [doi: 10.1097/01.mlr.0000160417.39497.a9] [Medline: 15838413]
21. Gundlapalli A, South B, Phansalkar S, Kinney A, Shen S, Delisle S, et al. Application of natural language processing to VA electronic health records to identify phenotypic characteristics for clinical and research purposes. Summit Transl Bioinform 2008 Mar 1;2008:36-40 [FREE Full text] [Medline: 21347124]
22. Zheng C, Rashid N, Wu Y, Koblick R, Lin AT, Levy GD, et al. Using natural language processing and machine learning to identify gout flares from electronic clinical notes. Arthritis Care Res (Hoboken) 2014 Nov;66(11):1740-1748 [FREE Full text] [doi: 10.1002/acr.22324] [Medline: 24664671]
23. Li L, Chase H, Patel C, Friedman C, Weng C. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. AMIA Annu Symp Proc 2008 Nov 6:404-408 [FREE Full text] [Medline: 18999285]
24. Friedlin J, Overhage M, Al-Haddad M, Waters J, Aguilar-Saavedra J, Kesterson J, et al. Comparing methods for identifying pancreatic cancer patients using electronic data sources. AMIA Annu Symp Proc 2010 Nov 13;2010:237-241 [FREE Full text] [Medline: 21346976]
25. Xu H, Fu Z, Shah A, Chen Y, Peterson N, Chen Q, et al. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. AMIA Annu Symp Proc 2011;2011:1564-1572 [FREE Full text] [Medline: 22195222]
26. Danforth K, Early M, Ngan S, Kosco A, Zheng C, Gould M. Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. J Thorac Oncol 2012 Aug;7(8):1257-1262 [FREE Full text] [doi: 10.1097/JTO.0b013e31825bd9f5] [Medline: 22627647]
27. Bielinski S, Pathak J, Carrell D, Takahashi P, Olson J, Larson N, et al. A robust e-Epidemiology tool in phenotyping heart failure with differentiation for preserved and reduced ejection fraction: the electronic medical records and genomics (emerge) network. J Cardiovasc Transl Res 2015 Nov;8(8):475-483 [FREE Full text] [doi: 10.1007/s12265-015-9644-2] [Medline: 26195183]
28. Corey K, Kartoun U, Zheng H, Shaw S. Development and validation of an algorithm to identify nonalcoholic fatty liver disease in the electronic medical record. Dig Dis Sci 2016 Mar;61(3):913-919 [FREE Full text] [doi: 10.1007/s10620-015-3952-x] [Medline: 26537487]
29. Goodwin T, Harabagiu S. Multi-modal patient cohort identification from EEG report and signal data. AMIA Annu Symp Proc 2016;2016:1794-1803 [FREE Full text] [Medline: 28269938]
30. Friedman C, Alderson P, Austin J, Cimino J, Johnson S. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161-174 [FREE Full text] [doi: 10.1136/jamia.1994.95236146] [Medline: 7719797]
31. Savova G, Masanz J, Ogren P, Zheng J, Sohn S, Kipper-Schuler K, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
32. Pathak J, Bailey K, Beebe C, Bethard S, Carrell D, Chen P, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. J Am Med Inform Assoc 2013 Dec;20(e2):e341-e348 [FREE Full text] [doi: 10.1136/amiajnl-2013-001939] [Medline: 24190931]
33. Wu S, Sohn S, Ravikumar K, Wagholikar K, Jonnalagadda S, Liu H, et al. Automated chart review for asthma cohort identification using natural language processing: an exploratory study. Ann Allergy Asthma Immunol 2013 Nov;111(5):364-369 [FREE Full text] [doi: 10.1016/j.anai.2013.07.022] [Medline: 24125142]
34. Wu Y, Denny J, Rosenbloom T, Miller R, Giuse D, Xu H. A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries. AMIA Annu Symp Proc 2012;2012:997-1003 [FREE Full text] [Medline: 23304375]
35. Chopard D, Spasic I. A deep learning approach to self-expansion of abbreviations based on morphology and context distance. In: Martín-Vide C, Purver M, Pollak S, editors. Statistical Language and Speech Processing. Cham: Springer; 2019:71-82.

XSL•FO
RenderX

36. Chen L, Gu Y, Ji X, Lou C, Sun Z, Li H, et al. Clinical trial cohort selection based on multi-level rule-based natural language processing system. J Am Med Inform Assoc 2019 Nov 1;26(11):1218-1226. [doi: 10.1093/jamia/ocz109] [Medline: 31300825]
37. Pakhomov S, Buntrock J, Chute C. Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier. J Biomed Inform 2005 Apr;38(2):145-153 [FREE Full text] [doi: 10.1016/j.jbi.2004.11.016] [Medline: 15797003]
38. Xiong Y, Shi X, Chen S, Jiang D, Tang B, Wang X, et al. Cohort selection for clinical trials using hierarchical neural network. J Am Med Inform Assoc 2019 Nov 1;26(11):1203-1208. [doi: 10.1093/jamia/ocz099] [Medline: 31305921]
39. Wolpert D. The lack of a priori distinctions between learning algorithms. Neural Comput 1996;8(7):1341-1390. [doi: 10.1162/neco.1996.8.7.1341]
40. Maguen S, Madden E, Patterson O, DuVall S, Goldstein L, Burkman K, et al. Measuring use of evidence based psychotherapy for posttraumatic stress disorder in a large national healthcare system. Adm Policy Ment Health 2018 Jul;45(4):519-529. [doi: 10.1007/s10488-018-0850-5] [Medline: 29450781]
41. Kotfila C, Uzuner O. A systematic comparison of feature space effects on disease classifier performance for phenotype identification of five diseases. J Biomed Inform 2015 Dec(58 Suppl):S92-102 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.016] [Medline: 26241355]
42. Ni Y, Wright J, Perentesis J, Lingren T, Deleger L, Kaiser M, et al. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. BMC Med Inform Decis Mak 2015 Apr 14;15:28 [FREE Full text] [doi: 10.1186/s12911-015-0149-3] [Medline: 25881112]
43. Ni Y, Bermudez M, Kennebeck S, Liddy-Hicks S, Dexheimer J. A real-time automated patient screening system for clinical trials eligibility in an emergency department: design and evaluation. JMIR Med Inform 2019 Jul 24;7(3):e14185 [FREE Full text] [doi: 10.2196/14185] [Medline: 31342909]
44. GitHub. 2019. c2s2 URL: https://github.com/dokato/c2s2 [accessed 2019-10-15]
45. Kumar V, Stubbs A, Shaw S, Uzuner O. Creation of a new longitudinal corpus of clinical narratives. J Biomed Inform 2015 Dec(58 Suppl):S6-10 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.018] [Medline: 26433122]
46. Stubbs A, Kotfila C, Xu H, Uzuner O. Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task Track 2. J Biomed Inform 2015 Dec(58 Suppl):S67-S77 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.001] [Medline: 26210362]
47. Griffis D, Shivade C, Fosler-Lussier E, Lai A. A quantitative and qualitative evaluation of sentence boundary detection for the clinical domain. AMIA Jt Summits Transl Sci Proc 2016;2016:88-97 [FREE Full text] [Medline: 27570656]
48. TextWorld. 2019. Simple Concordance Program URL: http://www.textworld.com/scp/ [accessed 2019-10-15]
49. Chapman W, Bridewell W, Hanbury P, Cooper G, Buchanan B. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001 Oct;34(5):301-310 [FREE Full text] [doi: 10.1006/jbin.2001.1029] [Medline: 12123149]
50. scikit-learn. 2019. URL: https://scikit-learn.org/ [accessed 2019-10-15]
51. XGBoost Documentation. 2019. URL: https://xgboost.readthedocs.io/ [accessed 2019-10-15]
52. Spasic I, Sarafraz F, Keane J, Nenadic G. Medication information extraction with linguistic pattern matching and semantic rules. J Am Med Inform Assoc 2010;17(5):532-535 [FREE Full text] [doi: 10.1136/jamia.2010.003657] [Medline: 20819858]
53. Spasic I, Burnap P, Greenwood M, Arribas-Ayllon M. A naïve bayes approach to classifying topics in suicide notes. Biomed Inform Insights 2012;5(Suppl 1):87-97 [FREE Full text] [doi: 10.4137/BII.S8945] [Medline: 22879764]
54. Nuttall A. Vert Asset Management. 2012. Considerations For Improving Patient Recruitment Into Clinical Trials URL: http://vertassets.blob.core.windows.net/download/64c39d7e/64c39d7e-c643-457b-aec2-9ff7b65b3ad2/rdprecruitmentwhitepaper.pdf [accessed 2019-10-16]

Abbreviations

BoW: bag-of-words
c2s2: Cardiff Cohort Selection System
cTAKES: clinical Text Analysis and Knowledge Extraction System
EEG: electroencephalography
EMR: electronic medical record
GTB: gradient tree boosting
HBA_{1c}: glycated hemoglobin
HNN: hierarchical neural network
ICD-9: International Classification of Diseases, Ninth Revision
IE: information extraction
IR: information retrieval
LR: logistic regression
MedLEE: Medical Language Extraction and Encoding
n2c2: National natural language processing Clinical Challenge

XSL•FO
RenderX

NB: naïve Bayesian
NLP: natural language processing
SVM: support vector machine
UMLS: Unified Medical Language System

Edited by G Eysenbach; submitted 28.08.19; peer-reviewed by G Karystianis, HJ Dai, T Basu, C Zheng, JT Pollettini, Y Wang; comments to author 19.09.19; revised version received 29.09.19; accepted 02.10.19; published 31.10.19.

Please cite as:
Spasic I, Krzeminski D, Corcoran P, Balinsky A
Cohort Selection for Clinical Trials From Longitudinal Patient Records: Text Mining Approach
JMIR Med Inform 2019;7(4):e15980
URL: http://medinform.jmir.org/2019/4/e15980/
doi:10.2196/15980
PMID:31674914

©Irena Spasic, Dominik Krzeminski, Padraig Corcoran, Alexander Balinsky. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 31.10.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Extracting Clinical Features From Dictated Ambulatory Consult Notes Using a Commercially Available Natural Language Processing Tool: Pilot, Retrospective, Cross-Sectional Validation Study

Jeremy Petch[1,2*], BA, MA, PhD; Jane Batt[3,4,5*], MD; Joshua Murray[6,7*], MSc; Muhammad Mamdani[1,6,8,9*], PharmD, MA, MPH

[1]Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[2]Centre for Data Science and Digital Health, Hamilton Health Sciences, Hamilton, ON, Canada
[3]Division of Respirology, Department of Medicine, University of Toronto, Toronto, ON, Canada
[4]Keenan Research Centre for Biomedical Science, St. Michael's Hospital, Toronto, ON, Canada
[5]Department of Medicine, St. Michael's Hospital, Toronto, ON, Canada
[6]Li Ka Shing Centre for Healthcare Analytics Research and Training, St. Michael's Hospital, Toronto, ON, Canada
[7]Department of Statistical Sciences, Faculty of Arts and Sciences, University of Toronto, Toronto, ON, Canada
[8]Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON, Canada
[9]Department of Medicine, Faculty of Medicine, University of Toronto, Toronto, ON, Canada
[*]all authors contributed equally

Corresponding Author:
Jeremy Petch, BA, MA, PhD
Centre for Data Science and Digital Health
Hamilton Health Sciences
293 Wellington St North
Hamilton, ON, L8L 8E7
Canada
Phone: 1 905 521 2100 ext 47579
Email: jeremy.petch@utoronto.ca

Abstract

Background: The increasing adoption of electronic health records (EHRs) in clinical practice holds the promise of improving care and advancing research by serving as a rich source of data, but most EHRs allow clinicians to enter data in a text format without much structure. Natural language processing (NLP) may reduce reliance on manual abstraction of these text data by extracting clinical features directly from unstructured clinical digital text data and converting them into structured data.

Objective: This study aimed to assess the performance of a commercially available NLP tool for extracting clinical features from free-text consult notes.

Methods: We conducted a pilot, retrospective, cross-sectional study of the accuracy of NLP from dictated consult notes from our tuberculosis clinic with manual chart abstraction as the reference standard. Consult notes for 130 patients were extracted and processed using NLP. We extracted 15 clinical features from these consult notes and grouped them a priori into categories of simple, moderate, and complex for analysis.

Results: For the primary outcome of overall accuracy, NLP performed best for features classified as simple, achieving an overall accuracy of 96% (95% CI 94.3-97.6). Performance was slightly lower for features of moderate clinical and linguistic complexity at 93% (95% CI 91.1-94.4), and lowest for complex features at 91% (95% CI 87.3-93.1).

Conclusions: The findings of this study support the use of NLP for extracting clinical features from dictated consult notes in the setting of a tuberculosis clinic. Further research is needed to fully establish the validity of NLP for this and other purposes.

(JMIR Med Inform 2019;7(4):e12575) doi:10.2196/12575

XSL•FO
RenderX

KEYWORDS

natural language processing; electronic health record; tuberculosis

Introduction

Background

In recent years, the use of electronic health records (EHRs) in office-based clinical practices in the United States has more than doubled, from approximately 40% in 2008 to nearly 90% in 2015 [1]. This rise has been even sharper in hospitals, where EHR adoption has increased from about 10% in 2008 to nearly 85% in 2015 [2]. The increasing adoption of EHRs in clinical practice holds the promise of improving care and advancing research by serving as a rich source of data. However, gleaning useful information from EHR data can be challenging, and the use of such data for research purposes varies considerably across jurisdictions [3].

One challenge relates to EHRs allowing clinicians to enter data in text format without much structure. Although this enhances clinical usability, it often requires costly and time-consuming manual chart abstraction processes to extract useful information in a structured manner. These challenges have sparked an increasing interest in the potential for natural language processing (NLP) approaches to process unstructured clinical digital text data, extract clinical features, and convert them into structured data.

Although NLP approaches for processing radiological reports are now well established [4], the practice of using NLP for processing more general clinical documentation, especially consult notes, is still developing. Research to date has explored several applications of NLP to general clinical documentation, including identification of breast cancer recurrence [5], social isolation [6], falls risk [7], depression [8], homelessness [9], intraductal papillary mucinous neoplasms [10], and new clinically relevant information for organ transplant patients [11]. One common feature of much of the research to date is that studies have tended to leverage open-source and academic tools for NLP. Although these tools can be highly effective, most are available as libraries for programing languages such as Python and R, which can pose a barrier for health care organizations that lack robust digital capacity or academic partnerships. However, there are an increasing number of commercially available NLP tools, such as Linguimatics I2E and Google Cloud's AutoML, that promise to make NLP significantly more accessible for general users, but to date, there have been relatively fewer studies that have evaluated the validity of these tools for clinical feature extraction [6,7,12].

Objective

We conducted a pilot study to examine the accuracy of a commercially available NLP tool relative to manual chart abstraction in capturing useful information from free-text consult notes in an outpatient tuberculosis (TB) clinic.

Methods

Study Setting

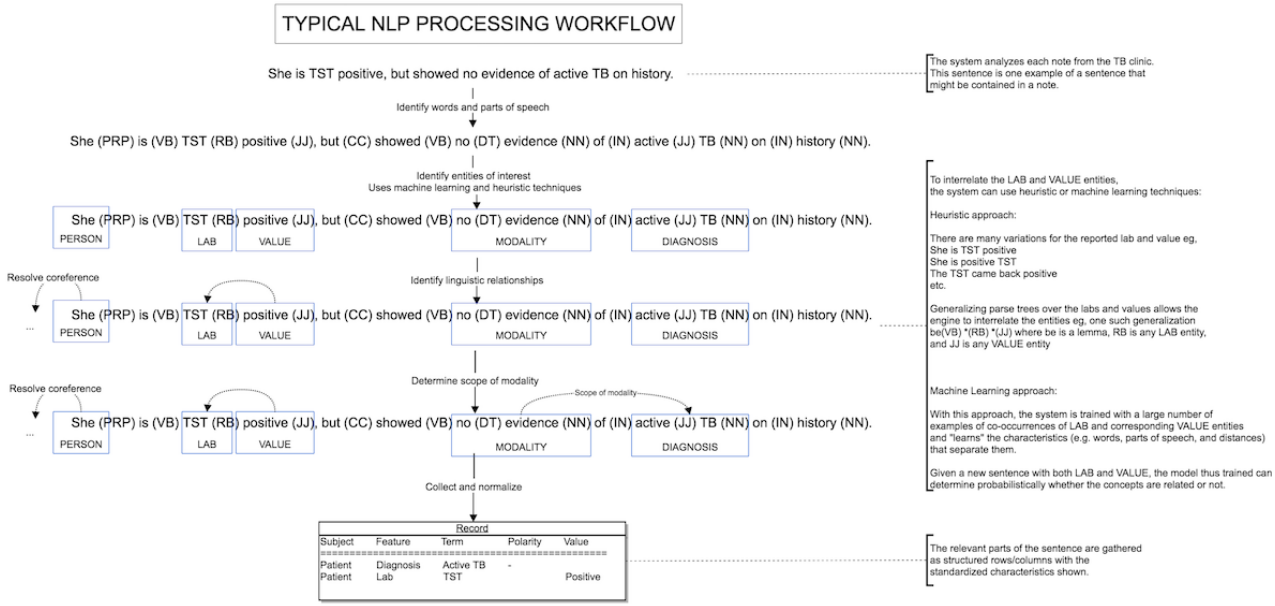
We conducted a pilot, retrospective, cross-sectional study of feature extraction accuracy using NLP, with manual chart abstraction as the reference standard. The study setting was St. Michael's Hospital, which is a 450-bed urban academic hospital affiliated with the University of Toronto. The St. Michael's TB program serves as a tertiary referral center for patients with active TB and latent TB infection, managing patients in both inpatient and outpatient settings. The program is staffed by a rotating roster of 8 physicians (6 respirologists and 2 infectious disease physicians) and 1 TB nurse practitioner and has a volume of approximately 2000 outpatient encounters annually. This study was approved by the St. Michael's Hospital Research Ethics Board and conducted in accordance with its policies.

Natural Language Processing Approach

We conducted our NLP analysis using a commercial NLP engine (Pentavere's DARWEN), which integrates 3 primary approaches to extract clinical features: (1) manually prepared natural language extraction rules that describe the general syntax and lexicon of each feature (both custom and internationally recognized ontologies such as Medical Subject Headings and Systematized Nomenclature of Medicine-Clinical Terms are utilized as an initial source of synonyms for common clinical terms), (2) machine-learned inferred rules that are designed to complement and reduce the extraction error rate of the manually prepared rules (the usage of machine learning in DARWEN is directed to improve the quality of the clinical natural language extraction rather than to predict or infer clinical features based on other features, as is the case with many competing systems), and (3) heuristic rules that encapsulate overarching clinical knowledge that must be respected when considering the clinical features holistically. This workflow is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. Natural language processing (NLP) workflow using the DARWEN tool. PRP: pronoun; VB: verb; RB: adverb; JJ: adjective; CC: coordinating conjunction; DT: determiner; NN: noun; IN: preposition; TB: tuberculosis; TST: Tuberculin Skin Test.



We followed the standard process for employing DARWEN, which involves tuning, testing, and retuning against a reference standard, together with clinician consultation to resolve any semantic issues as well as to develop the heuristic rules. Tuning refers to the process of refining NLP extraction rules based on manual analysis of text and is an essential step to successfully account for the variability in terminology and documentation structure between clinicians. Generating rules during the tuning process is an iterative, feature-by-feature, semisupervised process. First, we focused on recognizing the key entities associated with any feature, such as comorbidities. Given the low volume of data in the training set, we started with recurrent neural network-based named entity recognition (NER) models, which were pretrained for recognizing drugs, diagnosis, medical risk factors, and adverse drug reactions on Pentavere's proprietary clinical dataset (Pentavere's proprietary corpus includes over 100,000 patients, with an average of 50 clinical notes per patient); discussed the match results with the clinician; and supplemented the NER model with heuristics to accommodate any discrepancies. For clinical features not appropriate for NER models, we employed a purely heuristic approach. For example, for a feature such as smoking status, we developed an initial set of rules to cover 3 straightforward cases: explicit mention of nonsmoker (eg, "She never smokes"), explicit mention of former smoker (eg, "she is a former light smoker"), and qualified mention of former smoker (eg, "She is a smoker who gave up 2 years ago"). Although these captured many cases of smoking found in the text, the tuning process revealed many more subtle cases that require further development of rules, such as a smoker who quit and then started again, handling of indeterminant language (eg, "She has a 20 pack year smoking history" in which it is not clear whether the patient still smokes or has quit), oblique mentions (eg, "She uses marijuana"), and second-hand smoker (eg, "Her former roommate was a smoker, but she was not.") In this case, we developed rules to label token sequences (spans) into each of the different cases of smoker, former smoker, and nonsmoker. These rules are a combination of syntactic and lexical patterns, sometimes manually specified and sometimes induced from the data itself.

We then turned our attention to modeling the relationships between entities using a constituent parse tree kernel–based induction semisupervised machine learning technique, Pentavere's proprietary algorithm inspired by the Dual Iterative Pattern Relation Expansion algorithm [13]. For training data, the algorithm uses a few starting phrases or sentences that provide a valid relationship and a few that provide an invalid relationship. Given some initial examples of related entities, the algorithm finds generalizations of parse trees that define those known relationships. These syntactic rules/patterns were then applied to find other entities that appear to be in similar relationships. We also leveraged features of the tool that support several contextual states, including polarity (negation), certainty/uncertainty, hypothesis (if… then…), historical context (history of…), and experiencer (patient and family member). This contextualization uses constituent and dependency parse trees to describe different types of relationships between tokens in text and thus determine the scope of the context, for example, to restrict a context to only apply to entities contained in specific sub (constituent) trees of the context and/or require a specific dependency relationship between the entities in context. For a case such as, "She has no apparent rash causing her pruritus," this approach recognizes that rash is negated but pruritus is not negated.

Sampling Approach

To create our corpus, we randomly sampled 130 patient records from a total pool of 351 records from our hospital's outpatient TB clinic without exclusion and extracted their consult notes from their EHR. Consult notes for all outpatient encounters in the TB clinic are dictated by the attending physician or resident, followed by review and electronic sign-off by the attending physician. Dictations are free format, with no standardized template. They contain detailed clinical information about

XSL•FO
RenderX

patients' demographics, diagnosis, treatment course (including medications), and progress. Given that these notes contain personal health information, we are not able to share the corpus, but we have included synthetic samples of both assessment and follow-up notes in Multimedia Appendix 1, which are representative of the corpus.

We randomly divided our sample into 3 parts to support the tuning process described above, a tuning sample (n=30), a first-round testing/retuning sample (n=50), and a final testing sample (n=50). A single patient record allotted to the final testing sample contained corrupted data, reducing the final testing sample size to 49.

Feature Identification

The following features were selected for extraction: country of birth, date of immigration to Canada, HIV status, known TB exposure, previous TB, smoking status, diagnosis, method of diagnosis, TB sensitivities, sputum culture conversion date, drug treatments, adverse drug reactions, medical risk factors for TB acquisition, social risk factors for TB acquisition, and disease extent (Table 1).

Table 1. Feature categorization based on a priori assessment of clinical and linguistic complexity.

Feature complexity and feature	Type	Examples
Simple		
Country of birth	Country	India; Indonesia
Date of immigration	Date	30/06/2013
Smoking status	Categorical	Current smoker; former smoker
Drug treatment	Text mapped to drug list	Isoniazid; rifampin
Moderate		
HIV status	Binary	Positive/negative
Known TB[a] exposure	Binary	Yes/no
Previous TB	Binary	Yes/no
Method of diagnosis	Categorical	Culture positive; polymerase chain reaction positive
TB sensitivities	Categorical	Fully sensitive; isoniazid resistant
Complex		
Diagnosis	Categorical	Active TB; latent TB infection
Sputum conversion date	Date	22/07/2016
Adverse drug reactions	Categorical	Peripheral neuropathy; rash
Medical risk factors	Categorical	Chemotherapy; renal failure
Social risk factors	Categorical	Refugee camp resident; jail inmate
Disease extent	Categorical	Pulmonary acid fast bacilli smear positive; disseminated

[a]TB: tuberculosis.

For each feature where a patient could have multiple observations, a series of dichotomous indicator features were created. For example, for drug treatment, patients could be on multiple medications, so dichotomous features were created for each relevant medication.

For analysis, we pooled these features into 3 categories—simple, moderate, and complex—based on an a priori assessment by a clinical expert of the relative clinical and linguistic complexity of each feature, based upon their clinical judgment (Table 1). Complex features were typically those where NLP would have to go well beyond simply categorizing terms based on a reference dictionary but would instead have to successfully process rich language with significant clinical context. For example, adverse drug reactions are particularly challenging as we may see the mention of a *rash* in the text, but this does not determine whether there was in fact a rash or whether a rash was the result of an adverse drug reaction. To determine whether there was a rash, we have to be able to rule out cases with the physician dictating "no evidence of rash," patient complaining of rash but not diagnosed as such by the physician, and the physician dictating that she discussed rashes as possible side effects of the medication. Once it has been determined that a rash is present, we must first determine whether a rash is in fact a possible side effect of a drug the patient had been prescribed and then identify if the rash started when the drug was administered, which unless explicitly dictated, requires the solution to process the patient encounters longitudinally.

The reference standard was created by manually extracting features from patient records using a standardized data extraction form by a trained chart reviewer to serve as the *reference standard analysis*. One of the coauthors (JB) trained both the chart reviewer and the NLP engineer on how to perform chart abstraction to ensure the same clinical criteria would be used by both. This coauthor (JB) performed arbitration in cases of disagreement between the chart abstractor and the NLP tool's

XSL•FO
RenderX

output. Arbitrated results were used to retune the model on the training dataset before the final testing phase.

Statistical Analysis

The primary outcome of our study was overall accuracy, defined as the number of correctly classified observations divided by the total number of observations [14]. Secondary outcomes were sensitivity (recall), specificity, positive predictive value (PPV; precision), and negative predictive value (NPV) [15,16]. NLP-abstracted data were treated as the *index analysis*, with manual chart review acting as the *reference standard analysis*.

Analysis was divided into 2 stages. The first stage was conducted after a single round of tuning of the NLP algorithms (n=50). The results of this stage were used to retune the semantic and heuristic rules used by the NLP tool to improve accuracy. The final analysis stage was conducted on the remaining records (n=49).

For the primary outcome, within each feature category, we calculated the accuracy and a 95% CI using standard methods for continuous features and proportions [17]. For secondary outcomes, we calculated the average and standard deviation within each category. For example, for the simple category, we calculated secondary outcomes for each feature within the category, averaged them, and calculated the standard deviation. This is a way of illustrating the average sensitivity, specificity, PPV and NPV, and spread across all classes of a multicategorical feature. All analyses were conducted using R (v 3.3.0).

Results

Overview

The study sample of 129 subjects included 71 females (55.0%, 71/129) with a mean age of 36.51 years and 58 males (45%) with a mean age of 46.74 years. Consult notes from 9 clinicians (8 physicians and 1 nurse practitioner) were included in the sample. A total of 138 points of discrepancy between the NLP process and the reference standard chart abstraction were identified.

Natural Language Processing Performance

For the primary outcome (Table 2), NLP performed best for features classified as simple, achieving an overall accuracy of 96% (95% CI 94.3-97.6). Performance was slightly lower for features of moderate clinical and linguistic complexity at 93% (95% CI 91.1-94.4) and lowest for complex features at 91% (95% CI 87.3-93.1).

For secondary outcomes (Table 2), NLP achieved a sensitivity of 94% (SD 7.7) for simple, 60% (SD 38.6) for moderate, and 74% (SD 45.7) for complex features and PPV of 96% (SD 6.4) for simple, 70% (SD 33.7) for moderate, and 54% (SD 37.4) for complex features. The relatively low sensitivity and PPV for moderate and complex features is in contrast to its specificity of 99% (SD 0.5) for simple, 94% (SD 5.0) for moderate, and 89% (SD 8.3) for complex features and NPV of 99% (SD 1.7) for simple, 96% (SD 6.6) for moderate, and 98% (SD 2.9) for complex features.

Unsurprisingly, we saw considerable variation in NLP's performance at the clinical feature level (Table 3). NLP performed extremely well for detecting drug prescriptions, achieving 100% for all primary and secondary outcomes for moxifloxacin, rifampin, ethambutol, and isoniazid. In contrast, NLP did not perform well at the feature level when measuring disease extent, with a sensitivity of only 25% for pulmonary acid fast bacilli (AFB) positive and 0% for extra pulmonary cases because of a very low number of these cases in our sample (4 pulmonary AFB-positive cases and 2 extrapulmonary cases).

Table 2. Primary and secondary outcomes for natural language processing (index analysis) compared with manual chart review (reference standard analysis).

Feature complexity	Primary outcome, overall accuracy (95% CI)	Secondary outcomes			
		Sensitivity/recall (SD)	Specificity (SD)	Positive predictive value/precision (SD)	Negative predictive value (SD)
Simple	96.3 (94.3-97.6)	93.8 (7.7)	99.7 (0.5)	96.4 (6.4)	99.0 (1.7)
Moderate	92.9 (91.1-94.4)	60.2 (38.6)	94.2 (5.0)	70.2 (33.7)	95.6 (6.6)
Complex	90.6 (87.3-93.1)	73.8 (45.7)	89.2 (8.3)	53.6 (37.4)	98.4 (2.9)

XSL•FO
RenderX

Table 3. Primary and secondary outcomes for natural language processing (index analysis) compared with manual chart review (reference standard analysis) at the clinical feature level.

Feature	Primary outcome, overall accuracy (95% CI)	Secondary outcomes[a]			
		Sensitivity/recall (SD)	Specificity (SD)	Positive predictive value/precision (SD)	Negative predictive value (SD)
Simple features					
Country of birth	0.92 (0.80-0.98)	0.88 (0.32)	0.99 (0.01)	0.97 (0.11)	0.99 (0.01)
Year of immigration	0.90 (0.78-0.97)	0.89 (0.29)	0.99 (0.02)	0.98 (0.08)	0.99 (0.01)
Smoking status	0.94 (0.83-0.99)	0.92 (0.08)	0.98 (0.03)	0.85 (0.30)	0.97 (0.02)
Sputum conversion date	0.98 (0.89-0.99)	0.80 (0.45)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
Pyrazinamide	0.96 (0.86-0.99)	1.00	0.85	0.95	1.00
Moxifloxacin	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Vitamin B6	0.92 (0.80-0.98)	1.00	0.86	0.84	1.00
Rifampin	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Ethambutol	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Isoniazid	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Levofloxacin	0.98 (0.89-0.99)	N/A[b]	0.98	N/A	N/A
Moderate features					
HIV status	0.94 (0.83-0.99)	0.94	0.94	0.89	0.97
TB[c] contact	0.82 (0.68-0.91)	0.80	0.82	0.67	0.90
Old TB	0.94 (0.83-0.99)	0.71	0.98	0.83	0.95
Culture positive	0.88 (0.75-0.95)	0.33	1.00	1.00	0.87
Polymerase chain reaction positive	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Clinical diagnosis	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Drug sensitivity	0.92 (0.80-0.98)	0.81 (0.27)	0.97 (0.04)	0.73 (0.25)	0.91 (0.14)
Corticosteroids	0.98 (0.89-0.99)	N/A	0.98	N/A	N/A
Chemotherapy	0.94 (0.83-0.99)	0.50	0.96	0.33	0.98
Other immunosuppressive drugs	0.76 (0.61-0.87)	0.08	0.97	0.50	0.77
Cancer	0.92 (0.80-0.98)	1.00	0.91	0.33	1.00
Diabetes	0.98 (0.89-0.99)	0.86	1.00	1.00	0.98
Malnutrition	0.94 (0.83-0.99)	0.00	0.98	0.00	0.96
Other immunosuppressive conditions	0.82 (0.68-0.91)	0.10	1.00	1.00	0.81
Marginalized	0.96 (0.86-0.99)	0.66 (0.57)	0.93 (0.12)	0.99 (0.02)	0.91 (0.14)
Health care facility	0.90 (0.78-0.97)	0.38 (0.48)	0.95 (0.08)	0.95 (0.08)	0.97 (0.03)
Pulmonary acid fast bacilli					
Positive	0.92 (0.80-0.98)	0.25	0.98	0.50	0.93
Negative	0.96 (0.86-0.99)	1.00	0.96	0.67	1.00
Extrapulmonary (other than lymphadenitis)	0.88 (0.75-0.96)	0.00	0.96	0.00	0.91
Lymphadenitis	0.94 (0.83-0.99)	N/A	0.94	N/A	N/A
Disseminated	0.96 (0.86-0.99)	0.00	1.00	N/A	0.96
Complex features					

XSL•FO
RenderX

Feature	Primary outcome, overall accuracy (95% CI)	Secondary outcomes[a]			
		Sensitivity/recall (SD)	Specificity (SD)	Positive predictive value/precision (SD)	Negative predictive value (SD)
Active TB disease	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Latent TB infection	0.84 (0.70-0.93)	0.90	0.79	0.76	0.92
Pulmonary nontuberculous mycobacteria	0.88 (0.75-0.95)	1.00	0.87	0.25	1.00
Adverse drug reaction					
Gastrointestinal	0.84 (0.70-0.93)	1.00	0.76	0.65	1.00
Peripheral neuropathy	0.96 (0.86-0.99)	1.00	0.95	0.78	1.00
Rash	0.90 (0.78-0.97)	1.00	0.89	0.50	1.00
Other	0.94 (0.83-0.99)	0.00	0.98	0.00	0.96
Ocular toxicity	0.90 (0.75-0.97)	0.00	0.92	0.00	0.98

[a]Values within parenthesis are standard deviation values.

[b]N/A: not applicable.

[c]TB: tuberculosis.

Natural Language Processing Performance Adjusted for Adjudication

To understand whether NLP's relatively low sensitivity and PPV for moderate and complex features might be driven by errors in the manual chart review, rather than errors in NLP, we conducted a post hoc analysis in which all 138 points of discrepancy between the reference standard and index analysis were arbitrated by a clinical expert. The expert found the results to be in favor of NLP in 51.4% (71/138) of cases and chart review in 45.6% (63/138) of cases and found that both were incorrect in 2.8% (4/138) of cases.

After adjusting for the results of adjudication, results for our primary outcome of overall accuracy increased modestly to 98% (95% CI 96.1-98.7) for simple, 96% (95% CI 94.8-97.3) for moderate, and 94% (95% CI 91.3-96.1) for complex features. The sensitivity increased to 78% (SD 25.0) for moderate and 86% (SD 35.0) for complex features, and PPV increased to 93% (SD 14.7) for moderate and 70% (SD 34.2) for complex features (Table 4).

At the feature level (Table 5), adjustment for adjudication resulted in several dramatic improvements, particularly in the area of immunosuppressive drugs and conditions. For example, PPV for both cancer and chemotherapy was only 33% before adjudication but increased to 100% following adjudication. Similarly, for other immunosuppressive drugs, sensitivity was only 8% and PPV was only 50% initially, but it increased to 67% and 100%, respectively, after adjudication.

Table 4. Primary and secondary outcomes for natural language processing compared with manual chart review, adjusted for results of adjudication.

Feature complexity	Primary outcome, overall accuracy (95% CI)	Secondary outcomes			
		Sensitivity/recall (SD)	Specificity (SD)	Positive predictive value/precision (SD)	Negative predictive value (SD)
Simple	97.8 (96.1-98.7)	96.4 (5.4)	99.8 (0.5)	98.3 (4.5)	99.2 (1.7)
Moderate	96.2 (94.8-97.3)	78.2 (25.0)	93.3 (4.7)	92.7 (14.7)	97.2 (3.2)
Complex	94.1 (91.3-96.1)	86.3 (35.0)	92.8 (8.2)	70.5 (34.2)	98.7 (2.9)

XSL•FO
RenderX

Table 5. Primary and secondary outcomes for natural language processing compared with manual chart review, adjusted for results of adjudication at the clinical feature level.

Feature	Primary outcome, overall accuracy (95% CI)	Secondary outcomes[a]			
		Sensitivity/recall (SD)	Specificity (SD)	Positive predictive value/precision (SD)	Negative predictive value (SD)
Simple features					
Country of birth	0.94 (0.83-0.99)	0.91 (0.28)	0.99 (0.01)	0.98 (0.10)	0.99 (0.01)
Year of immigration	0.92 (0.80-0.98)	0.92 (0.23)	0.99 (0.02)	0.99 (0.06)	0.99 (0.01)
Smoking status	0.94 (0.83-0.99)	0.92 (0.08)	0.98 (0.03)	0.85 (0.30)	0.97 (0.02)
Sputum year	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Pyrazinamide	0.96 (0.86-0.99)	1.00	0.85	0.95	1.00
Moxifloxacin	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Vitamin B6	0.92 (0.80-0.98)	1.00	0.86	0.84	1.00
Rifampin	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Ethambutol	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Isoniazid	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Levofloxacin	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Moderate features					
HIV status	0.98 (0.89-0.99)	0.95	1.00	1.00	0.97
TB[b] contact	0.86 (0.73-0.94)	0.92	0.83	0.67	0.97
Old TB	0.96 (0.86-0.99)	0.75	1.00	1.00	0.95
Culture positive	0.88 (0.75-0.95)	0.33	1.00	1.00	0.87
Polymerase chain reaction positive	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Clinical diagnosis	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Drug sensitivity	0.96 (0.86-0.99)	0.98 (0.03)	0.99 (0.01)	0.80 (0.26)	0.94 (0.10)
Corticosteroids	1.00 (0.93, 1.00)	1.00	1.00	1.00	1.00
Chemotherapy	0.98 (0.89-0.99)	0.75	1.00	1.00	0.98
Other immunosuppressive drugs	0.98 (0.89-0.99)	0.67	1.00	1.00	0.98
Cancer	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Diabetes	0.98 (0.89-0.99)	0.86	1.00	1.00	0.98
Malnutrition	0.94 (0.83-0.99)	0.00	0.98	0.00	0.96
Other immunosuppressive conditions	0.98 (0.89-0.99)	0.5	1.00	1.00	0.98
Marginalized	0.98 (0.89-0.99)	0.75 (0.50)	0.95 (0.10)	0.99 (0.01)	0.99 (0.01)
Health care facility	0.92 (0.80-0.97)	0.50 (0.50)	0.86 (0.29)	0.95 (0.06)	0.97 (0.03)
Pulmonary acid fast bacillus					
Positive	0.92 (0.80-0.98)	0.25	0.98	0.50	0.93
Negative	0.96 (0.86-0.99)	1.00	0.95	0.67	1.00
Extrapulmonary (other than lymphadenitis)	0.96 (0.86-0.99)	0.50	1.00	1.00	0.95
Lymphadenitis	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Disseminated	1.00 (0.93-1.00)	N/A[c]	1.00	N/A	N/A
Complex features					

XSL•FO
RenderX

Feature	Primary outcome, overall accuracy (95% CI)	Secondary outcomes[a]			
		Sensitivity/recall (SD)	Specificity (SD)	Positive predictive value/precision (SD)	Negative predictive value (SD)
Active TB disease	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Latent TB infection	0.84 (0.70-0.93)	0.90	0.79	0.76	0.92
Pulmonary nontuberculous mycobacteria	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Adverse drug reaction					
Gastrointestinal	0.90 (0.78-0.97)	1.00	0.84	0.78	1.00
Peripheral neuropathy	1.00 (0.93-1.00)	1.00	1.00	1.00	1.00
Rash	0.90 (0.78-0.97)	1.00	0.89	0.50	1.00
Other	0.97 (0.89-0.99)	1.00	0.98	0.50	1.00
Ocular toxicity	0.90 (0.75-0.97)	0.00	0.92	0.00	0.98

[a]Values within parenthesis are standard deviation values.

[b]TB: tuberculosis.

[c]N/A: not applicable.

Discussion

Principal Findings

The findings of this study suggest that a commercially available NLP tool can perform very well when compared with the reference standard of manual chart review in extracting useful clinical information from digital text notes in our TB clinic with limited training. This was especially true in the case of straightforward findings, such as prescribed medications, smoking status, country of birth, year of immigration, and sputum conversion date. Unsurprisingly, accuracy decreased slightly as clinical features became more complex, but it remained over 90% for complex features.

One notable finding is that although NLP performed extremely well with respect to specificity and NPV for moderate and complex findings, sensitivity and PPV were considerably lower. These results are in keeping with other studies using free-format clinical notes for complex feature extraction, such as the study by Perlis et al, who reported a sensitivity of 42% and PPV of 78% for the detection of depression [8]. However, these findings are in contrast to the high sensitivity and PPV reported in studies looking at radiology reports, such as the study by Al-Haddad et al, who demonstrated a sensitivity of 97% and PPV of 95% in the detection of intraductal papillary mucinous neoplasms [10]. This discrepancy may be either because of differences in complexity of features or because of differences inherent between radiology reports, which are relatively structured, often with minimal variability from practitioner to practitioner, versus free-format clinical notes, which have less structure and greater variability across practitioners.

In terms of ease of use of a commercially available tool, deploying Pentavere's DARWEN in our environment was a straightforward installation of their application on a desktop computer. The iterative tuning and relationship modeling for all clinical features took our NLP engineer roughly 4 weeks to complete. The tuning required roughly 6 hours of clinician time to provide clinical context for the NLP engineer, confirm clinical validity of heuristic rules, and perform arbitration of discrepancies between chart review and NLP.

Strengths and Limitations

Our study is novel in several ways. First, to our knowledge, this is only the third study to explore the validity of NLP for the identification of TB patients and the first to examine dictated consult notes versus radiological reports and structured laboratory results for this purpose [18,19]. Second, research on NLP applications in medicine tend to focus on only a single clinical condition such as the presence of a tumor [10], a diagnosis such as depression [8], or a social condition such as homelessness [9]. In contrast, our study is substantially broader compared with other more commonly published studies, looking at 15 distinct medical and social features. Finally, our study is one of the few to evaluate the performance of a commercially available NLP tool [6,7,12].

Our study has several limitations. First, review of the feature-level analysis reveals that some dichotomous features had very low incidence, making sensitivity and PPV very sensitive to error. Second, our choice to randomly sample for our initial training dataset (n=30) resulted in an undersampling of cases of ocular toxicity because of adverse drug reaction. As a result, the NLP tool was never trained on this feature and subsequently performed poorly for this feature during the final testing set, potentially underestimating the effectiveness of a properly trained tool. This suggests that a real-world application of this technology may require a more purposive sampling strategy than our random sampling approach. Third, our study employed only a single chart abstractor and a single adjudicator. Finally, this study was conducted at a single center, in a focused clinical area, and with a relatively small final test sample (n=49), which may limit the generalizability of our findings. However, the goal of this pilot study was to establish the feasibility of using NLP to extract clinical features from dictated consult notes and to inform the approach to larger future studies.

XSL•FO
RenderX

Conclusions

NLP technology has been advancing quickly in recent years, and the potential clinical applications are numerous. The findings of this study support the application of extracting clinical features from dictated consult notes in the setting of a TB clinic. Further research is needed to fully establish the validity of NLP for this and other purposes. However, its application to free-format consult notes may be of particular benefit, as it offers a course whereby clinicians can document in their preferred method of narrative free text, with data still available for applications such as research and program quality control initiatives, for example, without the cost and effort of manual chart review.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Corpus sample (synthetic consult notes).
[PDF File (Adobe PDF File), 219 KB - medinform_v7i4e12575_app1.pdf]

References

1. Health IT Dashboard. 2017. Office-Based Physician Electronic Health Record Adoption URL: https://dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php [accessed 2018-08-12]
2. Henry J, Pylypchuk Y, Searcy T, Patel V. Health IT Dashboard. 2016. Adoption of Electronic Health Record Systems among US Non-Federal Acute Care Hospitals: 2008-2015 URL: https://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php [accessed 2018-08-12]
3. van Velthoven MH, Mastellos N, Majeed A, O'Donoghue J, Car J. Feasibility of extracting data from electronic medical records for research: an international comparative study. BMC Med Inform Decis Mak 2016 Jul 13;16:90 [FREE Full text] [doi: 10.1186/s12911-016-0332-1] [Medline: 27411943]
4. Pons E, Braun LM, Hunink MG, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016 May;279(2):329-343. [doi: 10.1148/radiol.16142770] [Medline: 27089187]
5. Zeng Z, Espino S, Roy A, Li X, Khan SA, Clare SE, et al. Using natural language processing and machine learning to identify breast cancer local recurrence. BMC Bioinformatics 2018 Dec 28;19(Suppl 17):498 [FREE Full text] [doi: 10.1186/s12859-018-2466-x] [Medline: 30591037]
6. Zhu VJ, Lenert LA, Bunnell BE, Obeid JS, Jefferson M, Halbert CH. Automatically identifying social isolation from clinical narratives for patients with prostate cancer. BMC Med Inform Decis Mak 2019 Mar 14;19(1):43 [FREE Full text] [doi: 10.1186/s12911-019-0795-y] [Medline: 30871518]
7. Zhu VJ, Walker TD, Warren RW, Jenny PB, Meystre S, Lenert LA. Identifying falls risk screenings not documented with administrative codes using natural language processing. AMIA Annu Symp Proc 2017;2017:1923-1930 [FREE Full text] [Medline: 29854264]
8. Perlis RH, Iosifescu DV, Castro VM, Murphy SN, Gainer VS, Minnier J, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol Med 2012 Jan;42(1):41-50 [FREE Full text] [doi: 10.1017/S0033291711000997] [Medline: 21682950]
9. Gundlapalli AV, Carter ME, Divita G, Shen S, Palmer M, South B, et al. Extracting concepts related to homelessness from the free text of VA electronic medical records. AMIA Annu Symp Proc 2014;2014:589-598 [FREE Full text] [Medline: 25954364]
10. Al-Haddad MA, Friedlin J, Kesterson J, Waters JA, Aguilar-Saavedra JR, Schmidt CM. Natural language processing for the development of a clinical registry: a validation study in intraductal papillary mucinous neoplasms. HPB (Oxford) 2010 Dec;12(10):688-695 [FREE Full text] [doi: 10.1111/j.1477-2574.2010.00235.x] [Medline: 21083794]
11. Zhang R, Pakhomov SV, Arsoniadis EG, Lee JT, Wang Y, Melton GB. Detecting clinically relevant new information in clinical notes across specialties and settings. BMC Med Inform Decis Mak 2017 Jul 5;17(Suppl 2):68 [FREE Full text] [doi: 10.1186/s12911-017-0464-y] [Medline: 28699564]
12. Levine MN, Alexander G, Sathiyapalan A, Agrawal A, Pond G. Learning health system for breast cancer: pilot project experience. JCO Clin Cancer Inform 2019 Aug;3:1-11 [FREE Full text] [doi: 10.1200/CCI.19.00032] [Medline: 31369338]
13. Brin S. Extracting Patterns and Relations from the World Wide Web. In: Proceedings of the International Workshop on the World Wide Web and Databases. 1998 Presented at: WebDB'98; March 27-28, 1998; Valencia, Spain p. 172-183 URL: https://doi.org/10.1007/10704656_11 [doi: 10.1007/10704656_11]
14. Shapiro DE. The interpretation of diagnostic tests. Stat Methods Med Res 1999 Jun;8(2):113-134. [doi: 10.1177/096228029900800203] [Medline: 10501649]
15. Altman DG, Bland JM. Diagnostic tests. 1: sensitivity and specificity. Br Med J 1994 Jun 11;308(6943):1552 [FREE Full text] [doi: 10.1136/bmj.308.6943.1552] [Medline: 8019315]

XSL•FO
RenderX

16. Altman DG, Bland JM. Diagnostic tests 2: predictive values. Br Med J 1994 Jul 9;309(6947):102 [FREE Full text] [doi: 10.1136/bmj.309.6947.102] [Medline: 8038641]
17. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed) 1986 Mar 15;292(6522):746-750 [FREE Full text] [doi: 10.1136/bmj.292.6522.746] [Medline: 3082422]
18. Jain NL, Knirsch CA, Friedman C, Hripcsak G. Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. Proc AMIA Annu Fall Symp 1996:542-546 [FREE Full text] [Medline: 8947725]
19. Hripcsak G, Knirsch CA, Jain NL, Pablos-Mendez A. Automated tuberculosis detection. J Am Med Inform Assoc 1997;4(5):376-381 [FREE Full text] [doi: 10.1136/jamia.1997.0040376] [Medline: 9292843]

Abbreviations

AFB: acid fast bacilli
EHR: electronic health record
NER: named entity recognition
NLP: natural language processing
NPV: negative predictive value
PPV: positive predictive value
TB: tuberculosis

Edited by CL Parra-Calderón, C Lovis; submitted 22.10.18; peer-reviewed by B Yu, M Torii, L Ferreira, S Zheng; comments to author 26.01.19; revised version received 12.05.19; accepted 29.08.19; published 01.11.19.

Please cite as:
Petch J, Batt J, Murray J, Mamdani M
Extracting Clinical Features From Dictated Ambulatory Consult Notes Using a Commercially Available Natural Language Processing Tool: Pilot, Retrospective, Cross-Sectional Validation Study
JMIR Med Inform 2019;7(4):e12575
URL: http://medinform.jmir.org/2019/4/e12575/
doi:10.2196/12575
PMID:31682579

©Jeremy Petch, Jane Batt, Joshua Murray, Muhammad Mamdani. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 01.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Combining Contextualized Embeddings and Prior Knowledge for Clinical Named Entity Recognition: Evaluation Study

Min Jiang[1], PhD; Todd Sanger[1], PhD; Xiong Liu[1], PhD

Eli Lilly and Company, Indianapolis, IN, United States

Corresponding Author:
Min Jiang, PhD
Eli Lilly and Company
893 Delaware St
Indianapolis, IN,
United States
Phone: 1 615 926 8277
Email: jiang_min@lilly.com

Abstract

Background: Named entity recognition (NER) is a key step in clinical natural language processing (NLP). Traditionally, rule-based systems leverage prior knowledge to define rules to identify named entities. Recently, deep learning–based NER systems have become more and more popular. Contextualized word embedding, as a new type of representation of the word, has been proposed to dynamically capture word sense using context information and has proven successful in many deep learning–based systems in either general domain or medical domain. However, there are very few studies that investigate the effects of combining multiple contextualized embeddings and prior knowledge on the clinical NER task.

Objective: This study aims to improve the performance of NER in clinical text by combining multiple contextual embeddings and prior knowledge.

Methods: In this study, we investigate the effects of combining multiple contextualized word embeddings with classic word embedding in deep neural networks to predict named entities in clinical text. We also investigate whether using a semantic lexicon could further improve the performance of the clinical NER system.

Results: By combining contextualized embeddings such as ELMo and Flair, our system achieves the F-1 score of 87.30% when only training based on a portion of the 2010 Informatics for Integrating Biology and the Bedside NER task dataset. After incorporating the medical lexicon into the word embedding, the F-1 score was further increased to 87.44%. Another finding was that our system still could achieve an F-1 score of 85.36% when the size of the training data was reduced to 40%.

Conclusions: Combined contextualized embedding could be beneficial for the clinical NER task. Moreover, the semantic lexicon could be used to further improve the performance of the clinical NER system.

(JMIR Med Inform 2019;7(4):e14850) doi:10.2196/14850

KEYWORDS

natural language processing; named entity recognition; deep learning; contextualized word embedding; semantic embedding; prior knowledge

Introduction

History of Clinical Named Entity Recognition

Clinical named entity recognition (NER), an important clinical natural language processing (NLP) task, has been explored for several decades. In the early stage, most NER systems leverage rules and dictionaries to represent linguistic features and domain knowledge to identify clinical entities, such as MedLEE [1], SymText/MPlus [2,3], MetaMap [4], KnowledgeMap [5], cTAKES [6], and HiTEX [7]. To promote the development of machine learning–based system, many publicly available corpora have been developed by organizers of some clinical NLP challenges such as the Informatics for Integrating Biology and the Bedside (i2b2) 2009 [8], 2010 [9-13], 2012 [14-18], 2014 [19-23], ShARe/CLEF eHealth Evaluation Lab 2013 dataset [24], and Semantic Evaluation 2014 task 7 [25], 2015 task 6 [26], 2015 task 14 [27], and 2016 task 12 [28] datasets. Many machine learning–based clinical NER systems have been proposed, and they greatly improved performance compared with the early rule-based systems [13,29,30]. Most systems are implemented based on two types of supervised machine learning

XSL•FO
RenderX

algorithms: (1) classification algorithms such as support vector machines (SVMs) and (2) sequence labeling algorithms such as conditional random fields (CRFs), hidden Markov models (HMMs), and structural support vector machines (SSVMs). Among all of the algorithms, CRFs play the leading roles due to the advantage of the sequence labeling algorithms over classification algorithms in considering context information when making the prediction; CRFs, as one type of discriminative model, tend to achieve better performance for the same source of testing data compared with generative model-based algorithms such as HMMs. Even though CRFs have achieved a huge success in the clinical NER area, they have some obvious limitations: CRF-based systems lie in manually crafted features, which are time consuming, and their ability to capture context in a large window is limited.

Deep Neural Network–Based Named Entity Recognition Algorithms

In recent years, deep neural network–based NER algorithms have been extensively studied, and many deep learning–based clinical NER systems have been proposed. They have an obvious advantage over traditional machine learning algorithms since they do not require feature engineering, which is the most difficult part of designing machine learning–based systems. They also improve the ability to leverage the context information. Initially, word embedding [31] is proposed as a method to represent the word in a continuous way to better support neural network structure. Then several new neural network structures including recurrent neural networks (RNNs) and long short-term memory (LSTM) [32] have been introduced to better represent sequence-based input and overcome long-term dependency issues. Recently, contextual word representations generated from pretrained bidirectional language models (biLMs) have been shown to significantly improve the performance of state-of-the-art NER systems [33].

In biLMs, the language model (LM) can be described as: given a sequence of N tokens, (t_1, t_2, ..., t_N), the probability of token t_k can be calculated given the history (t_1, ..., t_{k-1}), and the sequence probability can be computed as seen in Figure 1.

Recent neural LMs usually include one layer of token input, which is represented by word embedding or a CNN over characters, followed by L layers of forward LSTMs. On the top layer, the SoftMax layer is added to generate a prediction score for the next token [33]. The biLM combines two such neural LMs: the forward LM and backward LM; the backward LM is similar to the forward LM, except it runs over the reverse sequence. As a whole, the biLM tries to maximize the log-likelihood of the forward and backward directions as seen in Figure 2.

Figure 1. Sequence probability in bidirectional language models.

$$(1) \qquad p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_1, t_2, \ldots, t_{k-1})$$

Figure 2. Log-likelihood of the forward and backward directions language models.

$$(2) \qquad \sum_{k=1}^{N} (\log p(t_k \mid t_1, \ldots, t_{k-1}; \theta_x, \vec{\theta}_{LSTM}, \theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_N; \theta_x, \overleftarrow{\theta}_{LSTM}, \theta_s))$$

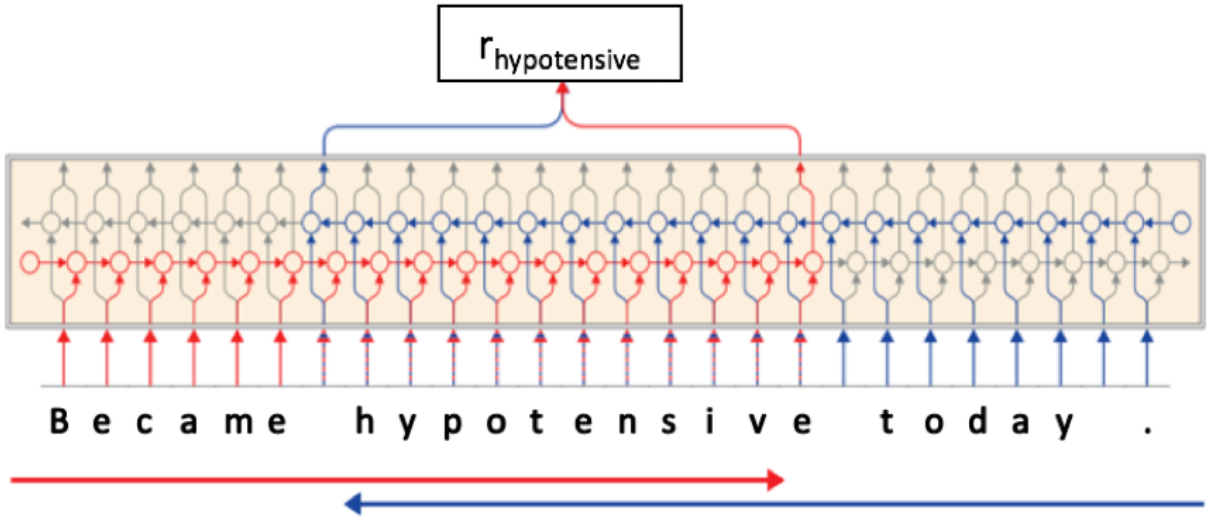
Where θ_x represents the token representation layer, θ_s represents the Softmax layer, and $\vec{\theta}_{LSTM}$ and $\overleftarrow{\theta}_{LSTM}$ represent the forward and backward directions of the LSTM layer.

In 2017, Peters et al [34] introduced a sequence tagger called TagLM that combines pretrained word embeddings and biLM embeddings as the representation of the word to improve the performance of the NER system. Since the output of each layer of the biLM represents a different type of contextual information [35], Peters et al [33] proposed another embedding, a deep contexualized word representation, ELMo, by concatenating all the biLM layer outputs into the biLM embedding with a weighted average pooling operation. The ELMo embedding adds CNN and highway networks over the character for each token as the input. ELMo has been proven to enhance the performance of different NLP tasks such as semantic role labeling and question answering [33].

Similar to Peters' ELMo, Akbik et al [36] introduced contextual string embeddings for sequence labeling, which leverages neural character-level language modeling to generate a contextualized embedding for each word input within a sentence. The principle of the character-level LM is that it is the same as biLMs except that it runs on the sequences of characters instead of tokens. Figure 3 shows the architecture of extracting a contextual string embedding for the word "hypotensive" in a sentence. We can see that instead of generating a fixed representation of the embedding for each word, the embedding of each token is composed of pretrained character embeddings from surrounding text, meaning the same token has dynamic representation depending on its context.

XSL•FO
RenderX

Figure 3. Architecture of extracting a contextual string embedding.



Deep Neural Network–Based Clinical Named Entity Recognition Systems

In the clinical domain, researchers investigated the performance of clinical NER tasks on various types of deep neural network structures. In 2015, researchers showed it is beneficial to use the large clinical corpus to generate word embeddings for clinical NER systems, and they comparatively investigated the different ways of generating word embeddings in the clinical domain [37]. In 2017, Wu et al [38] produced state-of-the-art results on the i2b2 2010 NER task dataset by employing the LSTM-CRF structure. Liu et al [39] investigated the effects of two types of character word embeddings on LSTM-based systems on multiple i2b2/Veterans Administration (VA) NER task datasets. In 2018, Zhu et al [40] employed a contextualized LM embedding on clinical data and boosted the state-of-the-art performance by 3.4% on the i2b2/VA 2010 NER dataset. The above studies show that, with the development of methods in text representation learning, especially contextual word embedding, more and more hidden knowledge can be learned from a large unannotated clinical corpus, which is beneficial for clinical NER tasks. According to the study by Peters et al [35], contextual word representations derived from pretrained biLMs can learn different levels of information that vary with the depth of the network, from local syntactic information to long-range dependent semantic information. Even without leveraging traditional domain knowledge such as lexicon and ontology, deep learning–based NER systems can achieve better performance than traditional machine learning–based systems.

Besides using pretrained representation from large unlabeled corpora, researchers started to integrate prior knowledge into deep learning frameworks to improve the performance of the NER system. For example, in the general domain, Yu and Dredze [41] created a semantic word embedding based on WordNet and evaluated the performance on language modeling, semantic similarity, and human judgment prediction. In another example, Weston et al [42] leveraged a CNN to generate a semantic embedding based on hashtags to improve the performance of the document recommendation task. In the clinical domain, Wu et al [43] compared two types of methods to inject medical knowledge into deep learning–based clinical NER solutions and found that the RNN-based system combining medical knowledge as embeddings achieved the best performance on the i2b2 2010 dataset. In 2019, Wang et al [44] explored two different architectures that extend the bidirectional LSTM (biLSTM) neural network and five different feature representation schemes to incorporate the medical dictionaries. In addition, other studies also use prior knowledge to generate embeddings [45-49].

To date, no detailed analysis has been published to investigate the value of combining different types of word embeddings and prior knowledge for clinical NER. In this study, we made the following contributions: (1) we proposed an innovative method to combine two types of contextualized embeddings to study their effects on the clinical NLP challenge dataset, (2) we incorporated prior knowledge from semantic resources such as medical lexicon to evaluate if it could further improve the performance of the clinical NER system, and (3) we conducted a thorough evaluation on our models with different sizes of data to gain knowledge on how much data are needed to train a high-performance clinical NER system.

Methods

Datasets

For this study, we used two datasets, the 2010 i2b2/VA concept extraction track dataset and the Medical Information Mart for Intensive Care III (MIMIC-III) corpus. The 2010 i2b2/VA challenge dataset is annotated with named entities, while the MIMIC-III corpus is unannotated data.

2010 i2b2/VA Concept Extraction Track Dataset

The goal of the 2010 i2b2/VA concept extraction task is to identify three types of clinical named entities including problem, treatment, and test from clinical notes. The original dataset includes 349 notes in the training set and 477 notes in the testing set, which include discharge summaries and progress notes from three institutions: Partners HealthCare, Beth Israel Deaconess

XSL•FO
RenderX

Medical Center, and University of Pittsburgh Medical Center. Since the University of Pittsburgh Medical Center's data have been removed from the original data set, the portion of discharge summaries that is available contains 170 notes for training and 256 for testing. In total, the training set contains 16,523 concepts including 7073 problems, 4844 treatments, and 4606 tests. The test set contains 31,161 concepts including 12,592 problems, 9344 treatments, and 9225 tests.

Medical Information Mart for Intensive Care III Corpus

The MIMIC-III corpus [50] is from MIMIC-III database, which is a large, freely available de-identified health-related dataset that integrates de-identified, comprehensive clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts.

The dataset comprises 2,083,180 notes from 15 different note types including "rehab services," "case management," "general," "discharge summary," "consult," "radiology," "electrocardiography," "nutrition," "social work," "pharmacy," "echocardiography," "physician," "nursing," "nursing/other," and "respiratory."

Embedding Generation

In order to fit our text input into the deep neural network structure, we generated three types of embeddings: classic word embeddings, (2) contextualized LM–based word embeddings, and semantic word embeddings.

Training Classic Word Embeddings

We generated two types of word embeddings based on the MIMIC-III corpus and a medical lexicon: MIMIC-III corpus-based embeddings and tagged MIMIC-III corpus-based embeddings. We adopted the Word2Vec implementation database from Github [51] to train word embeddings based on the MIMIC-III corpus. We used a continuous bag-of-words architecture with negative sampling. In accordance with the results from the study by Xu et al [52], we set the dimension of embedding as 50.

Training Contextual Language Model–Based Embeddings

Besides the word embeddings, we employed two recently proposed methods to generate contextual LM-based embeddings: ELMo embeddings and (2) contextual string embeddings for sequence labeling (Flair).

Training ELMo Embeddings

We followed the method introduced by Zhu et al [40] that uses a partial MIMIC-III corpus combined with a certain portion of Wikipedia pages as a training corpus to train the ELMo contextual LM in the clinical domain. In more detail, it combines discharge summaries and radiology reports from the MIMIC-III corpus and all the Wikipedia pages with titles that are items from the Systematized Nomenclature of Medicine–Clinical Terms. Such a corpus is trained on a deep neural network that contains a character-based CNN embedding layer followed by a two-layer biLSTM. Details have been published elsewhere [40].

Training Contextual String Embeddings for Sequence Labeling

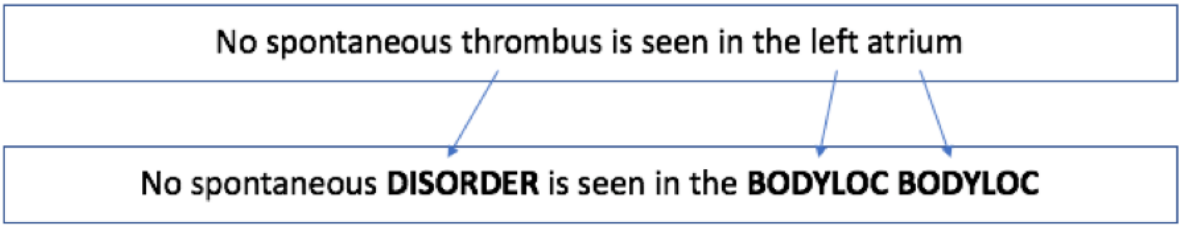
Akbik et al [36] proposed a new method to generate a neural character-level LM. The paper shows the state-of-the-art performance on the Conference on Computational Natural Language Learning 2003 NER task dataset. The LM for the general domain is publicly accessible. The author also integrates all the codes into an NLP framework called Flair. It achieved great success on the data in the general domain. However, according to the research by Friedman et al [53], clinical language has unique linguistic characteristics compared with general English, which make models generated from the public domain poorly adaptable to clinical narratives. It is demanding to train the LM on the clinical corpus to better support the clinical NER task. For training corpus preparation, we first did sentence segmentation on the entire corpus, then we randomly selected 1500 sentences as the testing set and another 1500 sentences for the validation set. The remaining part serves as the training set. For the hyperparameters, we kept the default setting: learning rate as 20.0, batch size as 32, anneal factor as 0.25, patience as 10, clip as 0.25, and hidden size as 1024.

Training Semantic Word Embeddings

Injecting domain knowledge into the deep learning model is a potential way to further improve the performance of the NER system. According to the results by Wu et al [43], combining medical knowledge into the embedding outperforms the method of representing it as a one-hot vector. Therefore, we similarly created the embedding to represent medical lexicon and fed it into the deep learning framework in our study. More specifically, we initially generated a lexicon dictionary based on a subset of semantic categories in the Unified Medical Language System. We then identified all the lexicon occurrences in the corpus using the dictionary and replaced them with semantic categories. Figure 4 shows an example of the conversion. In the example sentence of "No spontaneous thrombus is seen in the left atrium," "thrombus" is replaced with the tag "DISORDER" and "left atrium" is replaced with two "BODYLOC" tags. In this way, we can integrate semantic information into the word embeddings. For the embedding generation, we use the same setting as in the previous section.

XSL•FO
RenderX

Figure 4. One example of converting the sentence into the tagged sentence.

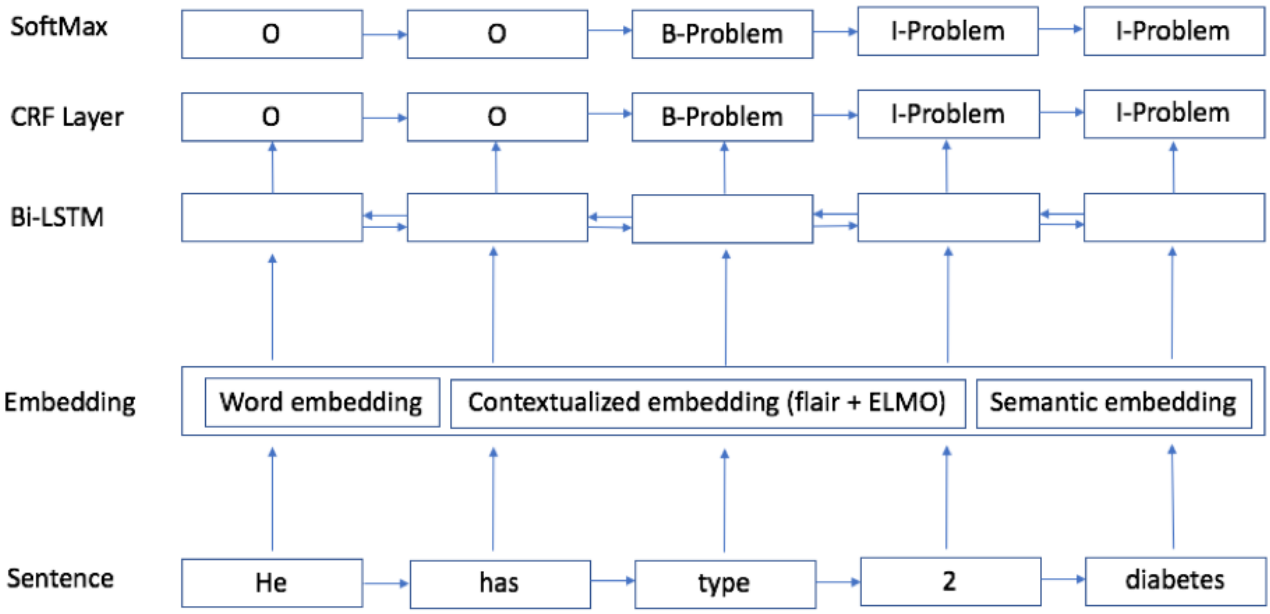


Deep Neural Network Architecture

After we generated all the embeddings, we started to fit them as the input into our deep neural network for the supervised training stage. Since each type of embedding is generated using one method, meaning each represents different aspects of knowledge from the large corpus, combining them is an obvious solution to potentially further improve the performance, which has also been proven by clinical NER studies [40,43]. Although there are many options to combine multiple embeddings in the deep neural network system such as weighting [54] and ensemble [55], in this study, we adopted the most straightforward way, which is simply concatenating them as the input.

We used the biLSTM-CRF sequence labeling module proposed by Huang et al [56]. Figure 5 shows the architecture of the whole deep neural network structure; the input is the embedding layer, which is concatenated by different types of embeddings as described in the previous section. Before we extracted embeddings for tagged word embedding, we used the same medical lexicon–based tagger to replace the tokens with the semantic tags. All the embedding inputs went through the biLSTM layer to generate forward and backward output, which was used to calculate the probability score by CRF layers. On the top, the prediction was given by a SoftMax layer.

Figure 5. Deep neural network structure with combined embeddings. Bi-LSTM: bidirectional long short-term memory; CRF: conditional random field.



Training the Deep Neural Network–Based Sequence Tagger

For the implementation, we employed Flair [57], which is a simple framework for NLP tasks including NER and text classification. We used the default hyperparameter setting in Flair, and we used the following configuration: learning rate as 0.1, batch size as 32, dropout probability as 0.5, and maximum epoch as 500. The learning rate annealing method is basically the same as the default: we half the learning rate if the training loss does not fall for the consecutive "patience" number of epochs. We set the patience number to 12 in this study. A TITAN V (NVIDIA Corporation) graphics processing unit was used to train the model. We took about 4 hours to train our model each time.

Evaluation

In order to get more reliable results, we ran each model three times. For the measurement of each running, we used precision, recall, and F-1 score.

XSL•FO
RenderX

Results

Table 1 shows the performance of the challenge winner system and different deep neural network systems. We used four benchmarks as our baseline systems, and then we reported the performance of the systems when adding ELMo embeddings, Flair embeddings, and tagged embeddings one at a time. All evaluation scores were based on exact matching. For the baseline systems, the first one is the semi-Markov model, developed by Debruijn et al [13], which reported an F-1 score of 85.23%. The second and third baselines are both based on the LSTM model, and they reported F-1 scores of 85.78% and 85.94%, respectively. The last baseline is the best result for the nonensemble models from Zhu et al [40], which used ELMo embedding. The three baseline systems used the original corpus (training: 349 notes; test: 477 notes), all other systems are based on the existing modified corpus (training: 170 notes; test: 256 notes). To start, we combined word embeddings with ELMo and Flair embeddings, respectively. Both models achieved an F-1 score of 87.01%, which is a little bit higher than what was reported by Zhu et al [40]. After combining word embeddings with ELMo and Flair embeddings, the F-1 score increased to 87.30%. When the word embedding on the tagged corpus was incorporated, the performance was further improved to 87.44% for the F-1 score.

In order to test if the improvement between different results is statistically significant, we conducted a statistical test based on results from bootstrapping. From the prediction result of the test set, we randomly selected 1000 sentences with replacement for 100 times and generated 100 bootstrap data sets. For each bootstrap data set, we evaluated F-measures for three pairs of results: (1) "biLSTM + ELMo" and "biLSTM + ELMo + Flair," (2) "biLSTM + ELMo + Flair" and "biLSTM + ELMo + Flair + semantic embedding," and (3) "biLSTM + ELMo by Zhu et al [40]" and "biLSTM + ELMo + Flair + semantic embedding." After that, we adopted a Wilcoxon signed rank test [58] to determine if the differences between F-measures from the three pairs were statistically significant. The results show that the improvement of F-measures for all three pairs were statistically significant (*P* values were .01, .02, and .03, respectively).

Table 1. Performance of all the models on the 2010 i2b2/VA dataset.

Model	F-1 (%)	Precision (%)	Recall (%)
Hidden semi-Markov[a]	85.23	86.88	83.64
LSTM[b] by Liu et al [39][a]	85.78	—[c]	—[c]
LSTM by Wu et al [43][a]	85.94	85.33	86.56
BiLSTM[d] + ELMo by Zhu et al [40][a]	86.84 (0.16)	87.44 (0.27)	86.25 (0.26)
BiLSTM + Flair	87.01 (0.18)	87.54 (0.15)	86.49 (0.21)
BiLSTM + ELMo	87.01 (0.24)	87.64 (0.19)	86.40 (0.30)
BiLSTM + ELMo + Flair	87.30 (0.06)	87.78 (0.09)	86.85 (0.07)
BiLSTM + ELMo + Flair + semantic embedding	87.44 (0.07)	88.03 (0.14)	86.91 (0.10)

[a]Model is trained using the complete dataset of i2b2 2010, which contains 349 notes in the training set and 477 notes in the test set.

[b]LSTM: long short-term memory.

[c]Not reported.

[d]BiLSTM: bidirectional LSTM.

Discussion

Principal Findings

NER is a fundamental task in the clinical NLP domain. In this study, we investigated the effects of combinations of different types of embeddings on the NER task. We also explored how to use medical lexicon to further improve performance. Based on the result, we found that either ELMo or Flair embeddings could boost the system's performance, and combining both embeddings could further improve the performance. Although both ELMo and Flair embeddings use biLM to train the LM on MIMIC-III corpus, they actually generate the contextualized word embeddings in different ways. ELMo concatenates all the biLM layers to represent all different levels of the knowledge, while Flair embedding is generated by a character-level LM. Character-level LM is different from character-aware LM [59] since it actually uses word-level LM while leveraging character-level features through a CNN encoding step. It was composed by the surrounding text's embedding in the character-level. The difference between ELMo and Flair embeddings could explain the reason why they can play complementary roles in the model.

The results show that adding semantic embeddings could further improve performance. According to the study by Peters et al [35], the lower biLM layer specializes in local syntactic relationships, while the higher layers focus on modeling longer range relationships. Those relationships are learned from the pure clinical corpus without any resources from outside such as medical lexicons and ontologies. This study shows an effective way to incorporate domain knowledge into the deep neural network–based NER system.

A large amount of training data is required to achieve success when applying deep learning algorithms [60]. Within the general domain, it is more difficult to accumulate a large size of the annotated corpus for most of the clinical NLP tasks since it

XSL•FO
RenderX

usually requires the annotator to have in-depth domain knowledge. Contextualized word embeddings, as an effective way of transferring the knowledge from the large unlabeled corpus, could address the issue of lack of training data. According to the results, by only using the small size of the training corpus (170 notes), contextualized word embedding–based models could achieve better performance than the models that use the large size training corpus (349 notes). To further investigate the effectiveness of transfer learning in our proposed models, we compared the performance of our best model generated from different sizes of the training data. Table 2 shows the F-1 score for the model "biLSTM + ELMo + Flair + semantic embedding" on randomly selected 80%, 60%, 40%, 20%, and 10% of the training data. Surprisingly, we found that using only 40% of the training corpus could achieve comparable performance as the original state-of-the-art traditional machine learning–based system. Even using 20% of the training corpus, the model's F-1 score is still more than 80%. This result indicates that contextualized word representation could potentially be an effective way to reduce the size of the training corpus, which could significantly improve the feasibility of applying deep learning to real practice.

Besides the performance reported in the Results section, we also recorded the change of performance for our proposed models during the fine-tuning stage. Table 3 shows the F-1 score on 1, 20, 40, and 60 epochs for our three models. On epoch 1, comparing to only word embeddings, any contextualized word embedding boosts the F-1 score. This is mostly because pretraining on contextualized word embeddings is very beneficial for the task of named entity recognition. This proves that the LM is a good way for pretraining that can be adapted to different downstream NLP tasks. Another interesting finding is that even though the model ELMo achieved the best performance among our three models, it was surpassed by the other two models on later epochs, which indicates that during the optimization process, the best starting point does not necessarily lead to the best local optimal solution.

Table 2. Performance of the best model training, BiLSTM[a] + ELMo + Flair + semantic embedding, on different sizes of the training corpus.

Amount of training data (%)	F-1 (%)	Prec (%)	Rec (%)
10	71.13	69.59	72.74
20	82.05	81.92	82.18
40	85.36	85.83	84.90
60	86.33	86.81	85.86
80	86.92	87.42	86.43

[a]BiLSTM: bidirectional long short-term memory.

Table 3. F-1 score for our proposed models on different epochs.

Model	1 epoch (%)	20 epochs (%)	40 epochs (%)	60 epochs (%)
Classic word embedding	61.23	75.67	78.11	79.52
Classic word embedding + ELMo	76.18	85.64	85.68	86.63
Classic word embedding + ELMo + Flair	73.28	85.33	85.97	86.96
Classic word embedding + ELMo + Flair + semantic embedding	74.38	85.85	86.46	87.13

Limitations

This study has some limitations. For contextualized embedding generation, we followed others' research methods and didn't test different configurations for LM training. For example, for ELMo embeddings, we followed the work of Zhu et al [40] for Flair embedding generation and kept the same configuration as seen in the work by Akbik et al [36]. For the fine-tuning stage, we only fine-tuned a limited set of hyperparameters including learning rate and patience. For domain knowledge integration, there are a lot of options that could be explored to merge the lexicon information into the input of the deep neural network structure. In this study, we only tried one way to represent it in the form of word embeddings. In this paper, we studied two contextualized embeddings: ELMo and Flair. In the future, we plan to test our framework by adding bidirectional encoder representations from transformers, which is another popular contextualized embedding [61].

Conclusions

In this study, we investigated the effects of the combination of two contextualized word embeddings including ELMo and Flair and clinical knowledge for the clinical NER task. Our evaluation on the 2010 i2b2/VA challenge dataset shows that using both ELMo and Flair embeddings outperforms using only ELMo embeddings, which indicates its great potential for the clinical NLP research. Furthermore, we demonstrate that incorporating the medical lexicon into the word representation could further improve the performance. Finally, we found that adopting our best model would be an effective way to reduce the size of the required training corpus for the clinical NER task.

XSL•FO
RenderX

Acknowledgments

This research was supported by the Advanced Analytics and Data Science organization at Eli Lilly and Company.

Conflicts of Interest

None declared.

References

1. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161-174 [FREE Full text] [Medline: 7719797]
2. Christensen L, Haug P, Fiszman M. MPLUS: a probabilistic medical language understanding system. 2002 Presented at: Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain; 2002; Stroudsburg. [doi: 10.3115/1118149.1118154]
3. Koehler SB. SymText: A Natural Language Understanding System for Encoding Free Text Medical Data. Provo: University of Utah; 1999.
4. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
5. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A. The KnowledgeMap project: development of a concept-based medical school curriculum database. AMIA Annu Symp Proc 2003:195-199 [FREE Full text] [Medline: 14728161]
6. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
7. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006 Jul 26;6(1):30. [doi: 10.1186/1472-6947-6-30]
8. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514-518 [FREE Full text] [doi: 10.1136/jamia.2010.003947] [Medline: 20819854]
9. Kim Y, Riloff E, Hurdle JF. A study of concept extraction across different types of clinical notes. AMIA Annu Symp Proc 2015;2015:737-746 [FREE Full text] [Medline: 26958209]
10. Tang B, Cao H, Wu Y, Jiang M, Xu H. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. BMC Med Inform Decis Mak 2013;13 Suppl 1:S1 [FREE Full text] [doi: 10.1186/1472-6947-13-S1-S1] [Medline: 23566040]
11. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552-556 [FREE Full text] [doi: 10.1136/amiajnl-2011-000203] [Medline: 21685143]
12. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21 [FREE Full text] [Medline: 11825149]
13. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J Am Med Inform Assoc 2011;18(5):557-562 [FREE Full text] [doi: 10.1136/amiajnl-2011-000150] [Medline: 21565856]
14. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc 2013 Sep 01;20(5):806-813. [doi: 10.1136/amiajnl-2013-001628]
15. Xu Y, Wang Y, Liu T, Tsujii J, Chang EI. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. J Am Med Inform Assoc 2013;20(5):849-858 [FREE Full text] [doi: 10.1136/amiajnl-2012-001607] [Medline: 23467472]
16. Tang B, Wu Y, Jiang M, Chen Y, Denny JC, Xu H. A hybrid system for temporal information extraction from clinical text. J Am Med Inform Assoc 2013;20(5):828-835 [FREE Full text] [doi: 10.1136/amiajnl-2013-001635] [Medline: 23571849]
17. Sohn S, Wagholikar KB, Li D, Jonnalagadda SR, Tao C, Komandur Elayavilli R, et al. Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification. J Am Med Inform Assoc 2013 Sep 01;20(5):836-842. [doi: 10.1136/amiajnl-2013-001622]
18. Kovačević A, Dehghan A, Filannino M, Keane JA, Nenadic G. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. J Am Med Inform Assoc 2013 Sep 01;20(5):859-866. [doi: 10.1136/amiajnl-2013-001625]
19. Stubbs A, Kotfila C, Uzuner O. Automated systems for the de-identification of longitudinal clinical narratives: overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform 2015 Dec;58 Suppl:S11-S19 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.007] [Medline: 26225918]
20. Yang H, Garibaldi JM. Automatic detection of protected health information from clinic narratives. J Biomed Inform 2015 Dec;58 Suppl:S30-S38 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.015] [Medline: 26231070]

XSL•FO
RenderX

21. Liu Z, Chen Y, Tang B, Wang X, Chen Q, Li H, et al. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. J Biomed Inform 2015 Dec;58 Suppl:S47-S52 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.009] [Medline: 26122526]
22. He B, Guan Y, Cheng J, Cen K, Hua W. CRFs based de-identification of medical records. J Biomed Inform 2015 Dec;58 Suppl:S39-S46 [FREE Full text] [doi: 10.1016/j.jbi.2015.08.012] [Medline: 26315662]
23. Dehghan A, Kovacevic A, Karystianis G, Keane JA, Nenadic G. Combining knowledge- and data-driven methods for de-identification of clinical narratives. J Biomed Inform 2015 Dec;58 Suppl:S53-S59 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.029] [Medline: 26210359]
24. Suominen H, Salanterä S, Velupillai S, Chapman W, Savova G, Elhadad N. Overview of the ShARe/CLEF eHealth evaluation lab. 2013 Presented at: International Conference of the Cross-Language Evaluation Forum for European Languages; 2013; Valencia. [doi: 10.1007/978-3-642-40802-1_24]
25. Pradhan S, Elhadad N, Chapman W, Manandhar S, Savova G. Semeval-2014 task 7: analysis of clinical text. 2014. URL: http://alt.qcri.org/semeval2014/cdrom/pdf/SemEval2014007.pdf [accessed 2019-10-22]
26. Bethard S, Derczynski L, Savova G, Pustejovsky J, Verhagen M. Semeval-2015 task 6: clinical tempeval. 2015. URL: http://alt.qcri.org/semeval2015/cdrom/pdf/SemEval136.pdf [accessed 2019-10-22]
27. Elhadad N, Pradhan S, Gorman S, Manandhar S, Chapman W, Savova G. SemEval-2015 task 14: analysis of clinical text. 2015. URL: http://alt.qcri.org/semeval2015/cdrom/pdf/SemEval051.pdf [accessed 2019-10-22]
28. Bethard S, Savova G, Chen W, Derczynski L, Pustejovsky J, Verhagen M. SemEval-2016 task 12: clinical TempEval. 2016. URL: http://alt.qcri.org/semeval2016/task12/ [accessed 2019-10-22]
29. Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc 2011;18(5):601-606 [FREE Full text] [doi: 10.1136/amiajnl-2011-000163] [Medline: 21508414]
30. Zhang Y, Wang X, Hou Z, Li J. Clinical named entity recognition from Chinese electronic health records via machine learning methods. JMIR Med Inform 2018 Dec 17;6(4):e50 [FREE Full text] [doi: 10.2196/medinform.9965] [Medline: 30559093]
31. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. Adv Neural Info Process Sys 2013.
32. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735]
33. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K. Deep contextualized word representations. arXiv preprint 2018:180205365. [doi: 10.18653/v1/n18-1202]
34. Peters M, Ammar W, Bhagavatula C, Power R. Semi-supervised sequence tagging with bidirectional language models. arXiv preprint 2017:170500108. [doi: 10.18653/v1/p17-1161]
35. Peters M, Neumann M, Zettlemoyer L. Dissecting contextual word embeddings: architecture and representation. arXiv preprint 2018. [doi: 10.18653/v1/d18-1179]
36. Akbik A, Blythe D, Vollgraf R. Contextual string embeddings for sequence labeling. 2018. URL: https://alanakbik.github.io/papers/coling2018.pdf [accessed 2019-10-22]
37. Wu Y, Xu J, Jiang M, Zhang Y, Xu H. A study of neural word embeddings for named entity recognition in clinical text. AMIA Annu Symp Proc 2015;2015:1326-1333 [FREE Full text] [Medline: 26958273]
38. Wu Y, Jiang M, Xu J, Zhi D, Xu H. Clinical named entity recognition using deep learning models. AMIA Annu Symp Proc 2017;2017:1812-1819 [FREE Full text] [Medline: 29854252]
39. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, et al. Entity recognition from clinical texts via recurrent neural network. BMC Med Inform Decis Mak 2017 Jul 05;17(Suppl 2):67 [FREE Full text] [doi: 10.1186/s12911-017-0468-7] [Medline: 28699566]
40. Zhu H, Paschalidis I, Tahmasebi A. arXiv preprint. 2018. Clinical concept extraction with contextual word embedding. URL: https://arxiv.org/abs/1810.10566 [accessed 2019-10-22]
41. Yu M, Dredze M. Improving lexical embeddings with semantic knowledge. 2014 Presented at: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2); 2014; Baltimore. [doi: 10.3115/v1/p14-2089]
42. Weston J, Chopra S, Adams K. Semantic embeddings from hashtags. 2014 Presented at: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014; Doha. [doi: 10.3115/v1/d14-1194]
43. Wu Y, Yang X, Bian J, Guo Y, Xu H, Hogan W. Combine factual medical knowledge and distributed word representation to improve clinical named entity recognition. AMIA Annu Symp Proc 2018;2018:1110-1117 [FREE Full text] [Medline: 30815153]
44. Wang Q, Zhou Y, Ruan T, Gao D, Xia Y, He P. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. J Biomed Inform 2019 Apr;92:103133. [doi: 10.1016/j.jbi.2019.103133] [Medline: 30818005]
45. Liu Q, Jiang H, Wei S, Ling Z, Hu Y. Learning semantic word embeddings based on ordinal knowledge constraints. 2015 Presented at: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1); 2015; Beijing. [doi: 10.3115/v1/p15-1145]

XSL•FO
RenderX

46. Mencia EL, de Melo G, Nam J. Medical concept embeddings via labeled background corpora. 2016 Presented at: Proceedings of the 10th Language Resources and Evaluation Conference (LREC); 2016; Portoroz.
47. Peng S, You R, Wang H, Zhai C, Mamitsuka H, Zhu S. DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. Bioinformatics 2016 Jun 15;32(12):i70-i79 [FREE Full text] [doi: 10.1093/bioinformatics/btw294] [Medline: 27307646]
48. Celikyilmaz A, Hakkani-Tur D, Pasupat P, Sarikaya R. Enriching word embeddings using knowledge graph for semantic tagging in conversational dialog systems. URL: https://www.aaai.org/ocs/index.php/SSS/SSS15/paper/download/10333/10034 [accessed 2019-10-22]
49. Passos A, Kumar V, McCallum A. Lexicon infused phrase embeddings for named entity resolution. arXiv preprint 2014:14045367. [doi: 10.3115/v1/w14-1609]
50. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
51. Word2Vec implementation. URL: https://github.com/dav/word2vec [accessed 2019-10-22]
52. Xu J, Zhang Y, Wang J, Wu Y, Jiang M, Soysal E. UTH-CCB: the participation of the SemEval 2015 challenge—Task 14. URL: https://clamp.uth.edu/challenges-publications/UTH-CCB-%20the%20participation%20of%20the%20SemEval%202015%20challenge%E2%80%93Task%2014.pdf [accessed 2019-10-22]
53. Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 2002 Aug;35(4):222-235 [FREE Full text] [Medline: 12755517]
54. Reimers N, Gurevych I. Alternative weighting schemes for ELMo embeddings. arXiv preprint 2019:190402954.
55. Speer R, Chin J. An ensemble method to produce high-quality word embeddings. arXiv preprint 2016:160401692.
56. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint 2015:150801991.
57. Akbik A. Flair implementation. URL: https://github.com/zalandoresearch/flair/graphs/contributors2018 [accessed 2019-10-22]
58. Woolson R. Wilcoxon signed-rank test. Wiley Encyclopedia of Clinical Trials 2007:1-3. [doi: 10.1002/9780471462422.eoct979]
59. Kim Y, Jernite Y, Sontag D, Rush AM. Character-aware neural language models. URL: https://arxiv.org/abs/1508.06615 [accessed 2019-10-22]
60. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 27;521(7553):436-444. [doi: 10.1038/nature14539]
61. Devlin J, Chang M, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint 2018:181004805.

Abbreviations

biLM: bidirectional language model
biLSTM: bidirectional long short-term memory
CNN: convolutional neural network
CRF: conditional random field
HMM: hidden Markov model
i2b2: Informatics for Integrating Biology and the Bedside
LM: language model
LSTM: long short-term memory
MIMIC-III: Medical Information Mart for Intensive Care III
NER: named entity recognition
NLP: natural language processing
RNN: recurrent neural network
SSVM: structural support vector machine
SVM: support vector machine
VA: Veterans Affairs

Edited by G Eysenbach; submitted 28.05.19; peer-reviewed by F Li, B Polepalli Ramesh; comments to author 18.06.19; revised version received 16.07.19; accepted 19.10.19; published 13.11.19.

Please cite as:
Jiang M, Sanger T, Liu X
Combining Contextualized Embeddings and Prior Knowledge for Clinical Named Entity Recognition: Evaluation Study
JMIR Med Inform 2019;7(4):e14850
URL: http://medinform.jmir.org/2019/4/e14850/
doi:10.2196/14850
PMID:31719024

XSL•FO
RenderX

©Min Jiang, Todd Sanger, Xiong Liu. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 13.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Using a Large Margin Context-Aware Convolutional Neural Network to Automatically Extract Disease-Disease Association from Literature: Comparative Analytic Study

Po-Ting Lai[1], PhD; Wei-Liang Lu[2], MSc; Ting-Rung Kuo[2], MSc; Chia-Ru Chung[2], PhD; Jen-Chieh Han[2], MSc; Richard Tzong-Han Tsai[2*], PhD; Jorng-Tzong Horng[2,3*], PhD

[1]Department of Computer Science National Tsing Hua University, Hsinchu, Province of China Taiwan

[2]Department of Computer Science & Information Engineering, National Central University, Taoyuan, Province of China Taiwan

[3]Department of Bioinformatics and Medical Engineering, Asia University, Taichung, Province of China Taiwan

[*]these authors contributed equally

Corresponding Author:
Richard Tzong-Han Tsai, PhD
Department of Computer Science & Information Engineering, National Central University
No 300, Zhongda Road, Zhongli District
Taoyuan
Province of China Taiwan
Phone: 886 3 422 7151 ext 35323
Email: thtsai@csie.ncu.edu.tw

Abstract

Background: Research on disease-disease association (DDA), like comorbidity and complication, provides important insights into disease treatment and drug discovery, and a large body of the literature has been published in the field. However, using current search tools, it is not easy for researchers to retrieve information on the latest DDA findings. First, comorbidity and complication keywords pull up large numbers of PubMed studies. Second, disease is not highlighted in search results. Finally, DDA is not identified, as currently no disease-disease association extraction (DDAE) dataset or tools are available.

Objective: As there are no available DDAE datasets or tools, this study aimed to develop (1) a DDAE dataset and (2) a neural network model for extracting DDA from the literature.

Methods: In this study, we formulated DDAE as a supervised machine learning classification problem. To develop the system, we first built a DDAE dataset. We then employed two machine learning models, support vector machine and convolutional neural network, to extract DDA. Furthermore, we evaluated the effect of using the output layer as features of the support vector machine-based model. Finally, we implemented large margin context-aware convolutional neural network architecture to integrate context features and convolutional neural networks through the large margin function.

Results: Our DDAE dataset consisted of 521 PubMed abstracts. Experiment results showed that the support vector machine-based approach achieved an F1 measure of 80.32%, which is higher than the convolutional neural network-based approach (73.32%). Using the output layer of convolutional neural network as a feature for the support vector machine does not further improve the performance of support vector machine. However, our large margin context-aware-convolutional neural network achieved the highest F1 measure of 84.18% and demonstrated that combining the hinge loss function of support vector machine with a convolutional neural network into a single neural network architecture outperforms other approaches.

Conclusions: To facilitate the development of text-mining research for DDAE, we developed the first publicly available DDAE dataset consisting of disease mentions, Medical Subject Heading IDs, and relation annotations. We developed different conventional machine learning models and neural network architectures and evaluated their effects on our DDAE dataset. To further improve DDAE performance, we propose an large margin context-aware-convolutional neural network model for DDAE that outperforms other approaches.

(JMIR Med Inform 2019;7(4):e14502) doi:10.2196/14502

XSL•FO
RenderX

KEYWORDS

deep learning; disease-disease association; biological relation extraction; convolutional neural networks; biomedical natural language processing

Introduction

Background

The origin and treatment of disease is an important research field in the life sciences, covering a wide range of research topics such as comorbidity, complication, genetic disorder, drug treatment, and adverse drug reaction. As disease is involved in many areas, new scientific findings are frequently made or updated.

Disease-disease association (DDA) is an important research topic in the biomedical domain [1-5]. The influence of one disease on others is wide ranging and can manifest in any patient. Diabetes, for example, may cause macrovascular diseases [6], such as cardiovascular disease [7] and cerebrovascular disease [8]. Treating a disease without consideration of potential DDAs may result in poor treatment outcomes. Therefore, DDAs are often a prime concern for researchers and doctors involved in drug discovery and disease treatment. Figure 1 illustrates examples of DDAs in the literature (refer to Multimedia Appendix 1 for more examples, including comorbidity, complications, general associations, and risk factors). There have been several studies attempting to generate disease connectivity networks [3-5]. However, the enormous and rapidly growing disease-related literature has not been utilized.

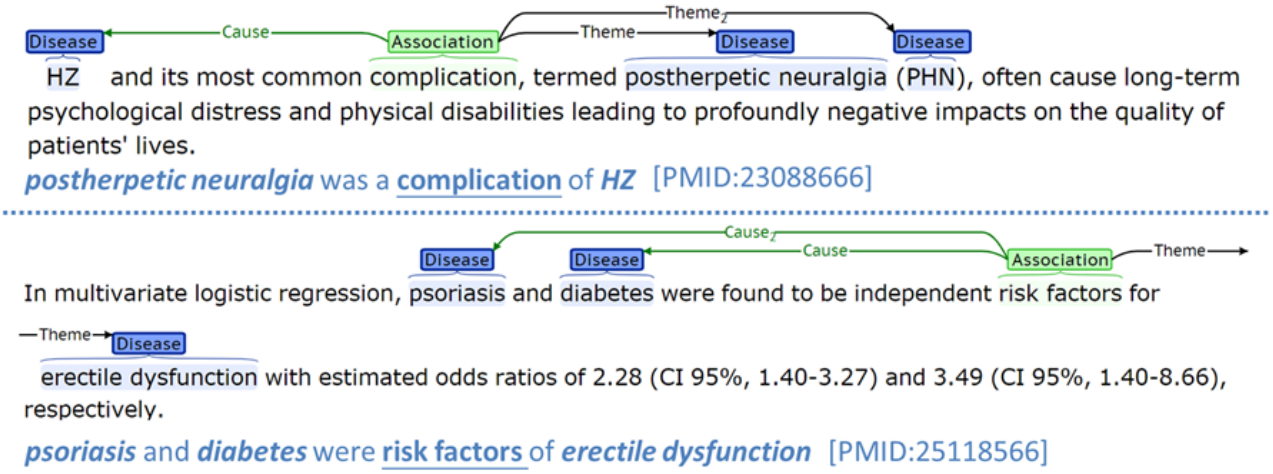
Finding DDA in the literature is a time-consuming and challenging task for researchers. First, there are huge numbers of DDA papers to sort through, and existing search engines, such as PubMed, do not mark up all relevant disease mentions in search results. Although there are text-mining tools available that could automatically identify diseases [9-11], genes [10,12,13], chemicals [14,15], and associations among them [16-22], they have not been integrated into a single interface to assist researchers in searching through the latest DDA findings. The main obstacle in creating a DDA extraction (DDAE) system is the lack of a relevant dataset. Moreover, only a few text-mining approaches [23] are suitable for extracting DDA.

In this study, we compiled a DDAE dataset consisting of 521 annotated PubMed abstracts. As it is hard for a human annotator to distinguish one DDA type from another without reading a broader context, such as a whole paragraph, we therefore annotated only 3 DDA types: positive, negative, and null associations:

1. Positive associations include comorbidity, complications, physical associations, and risk factors.
2. Negative associations are counted when the text clearly states that there is no association between 2 diseases.
3. Null associations are annotated when 2 diseases co-occur in a sentence, but no association is stated, suggested, or apparent.

In this study, we formulated DDAE as a supervised machine learning (ML) classification task in which, given a sentence containing a disease pair, the goal was to classify the pair into one of the DDA types. For classification, we employed 2 machine learning models, support vector machine (SVM) [24] and convolutional neural network (CNN) [25]. We compared different combinations of SVM and CNN to maximize performance, arriving at a novel neural network architecture, which we termed as large margin context-aware CNN (LC-CNN). LC-CNN achieved the highest F1 measure of 84.18% on our DDAE test set.

Figure 1. Disease-disease association extraction examples.



XSL•FO
RenderX

Related Work

In this section, we first review published disease annotation datasets. Then, we briefly review different methods of relation extraction in biomedical domains.

Disease Annotation Datasets

Before identifying DDAs, we have to identify diseases in the text first. Fortunately, there are many datasets for developing such disease name recognition and normalization systems. The National Center for Biotechnology Information (NCBI) disease dataset [26] is the most widely used. For instance, Leaman and Lu [9] proposed a semi-Markov model trained on an NCBI disease dataset that achieved an F1 measure of 80.7%. However, DDAs are not annotated in the NCBI dataset abstracts, limiting its usefulness for the DDAE task.

As DDAs can give insights into disease etiology and treatment, many studies focus on generating DDA networks [1-5]. For example, Sun et al [4] used disease-gene associations in the Online Mendelian Inheritance in Man [27] to predict DDAs with similar phenotypes. Bang et al [3] used disease-gene relations to define disease-disease network, and the causalities of disease pairs are confirmed through using clinical results and metabolic pathways. However, the constructed networks lack text evidence and therefore cannot be used to develop a DDAE dataset.

Xu et al [23] proposed a semisupervised iterative pattern-learning approach to learn DDA patterns from PubMed abstracts. They constructed a disease-disease risk relationship knowledge base (dRiskKB) consisting of 34,000 unique disease pairs. However, there are some limitations of dRiskKB that make it hard to use in developing DDAE systems. First, dRiskKB only provides positive DDA sentences. Owing to the lack of negative instances, it cannot be used to train ML-based classifiers. In addition, as the development of dRiskKB is based on a pattern-learning approach, it only includes DDA sentences with very simple structures and thus is not ideal for training a DDA system capable of analyzing complicated sentences.

To solve the above problems, we developed a DDAE dataset. Our dataset was different from dRiskKB in 3 aspects. First, our DDAE dataset contained positive, negative, and null DDAs. Second, it did not use patterns to annotate DDAs and therefore included DDA sentences with more complex expressions. Finally, it annotated DDAs in the entire abstract, allowing an ML-based classifier to use document-level features.

Relation Extraction

Rule-based approaches are commonly used in new domains or tasks that do not have large-scale annotated datasets. Lee et al's [28] approach is an example. They extracted protein-protein interactions (PPIs) from plain text using handcrafted dependency rules. Their approach did not require a training set, but it achieved a high precision of 97.4% on the Artificial Intelligence in Medicine (AIMed) dataset [29]. However, it was difficult for them to create rules that can extract all PPIs, and their system, therefore, achieved a low recall of 23.6%. Moreover, Nguyen et al [30] used predicate-argument structure (PAS) [31] rules to extract more general relations including PPI and drug-drug interaction. Their rules detected PPIs by examining where relation verbs and proteins are located in the spans of predicates and arguments. Their approach required less effort to design rules and was able to adapt to different relation types. Compared with Lee et al's system, it achieved a higher recall of 52.6% on the AIMed dataset but a lower precision of 30.4%.

ML-based approaches can usually achieve relatively higher performance than rule-based ones. For instance, Zhang et al [32] used hybrid feature–based and tree-based kernels implemented with SVM-LIGHT-TK [33] for PPI extraction. The feature-based kernel uses SENNA (Semantic/syntactic Extraction using a Neural Network Architecture)'s pretrained word-embedding model [34]. In the tree-based kernel configuration, the sentence dependency structure is used as input. The structure is decomposed into substructures and then transformed into one-hot encoding features for SVMs. Zhang et al's approach achieved an F score of 69.7% on the AIMed dataset, which is higher than Lee et al's 26.3% and Nguyen et al's 38.5%.

In addition to sentence-level features, document-level features are also useful in relation extraction. Peng et al [17] proposed an SVM-based approach for document-level chemical-disease relation (CDR) extraction. They used statistical features, such as whether a chemical or disease name appears in the title, to classify document-level chemical-disease pairs. By adding the features, they improved their F score from a baseline of 46.82% to 57.51% on the BioCreative V CDR dataset [35]. Our LC-CNN is partly inspired by Peng et al's [17] statistical features; our context vector adopts document-level features for sentence-level DDA classification.

Although the abovementioned feature-based approaches have made gains in many relation extraction tasks [36-38], it is difficult to find novel features to further improve performance. Several researchers are exploring deep learning approaches as a way forward. For instance, Peng and Lu [39] proposed a multichannel dependency-based CNN model (McDepCNN). McDepCNN uses 2 channels to represent an input sentence. One is the word-embedding layer, whereas the other is the head-word-embedding layer. Each embedding layer concatenates pretrained word-embedding vectors, one-hot encodings of part of speech, chunks, named entity labels, and dependency words. In PPI prediction, Peng and Lu's CNN model achieved F scores of 63.5% on AIMed and 65.3% on BioInfer.

For drug-drug interaction extraction, Lin et al [20] proposed a syntax CNN (SCNN) that integrates syntactic features, including words, predicates, and shortest dependency paths into a CNN. They trained their model with word2vec [40] and the Enju parser [31]. The Enju parser breaks the sentence into PASs, and non-PAS words or phrases are removed. The pruned sentences are then used to train the word-embedding model. Their approach achieved an F score of 68.6% on the 2013 DDIExtraction dataset.

Our LC-CNN was also inspired by Zhao et al's [20] SCNN architecture with 3 main differences. First, we replaced the log loss function with the hinge loss function. Second, SCNN uses a fully connected layer for traditional features before merging them with the CNN's output. However, LC-CNN directly

XSL•FO
RenderX

merges the CNN's output with traditional features. Finally, SCNN's traditional features only use sentence-level information, whereas LC-CNN also uses both sentence-level and document-level features.

Methods

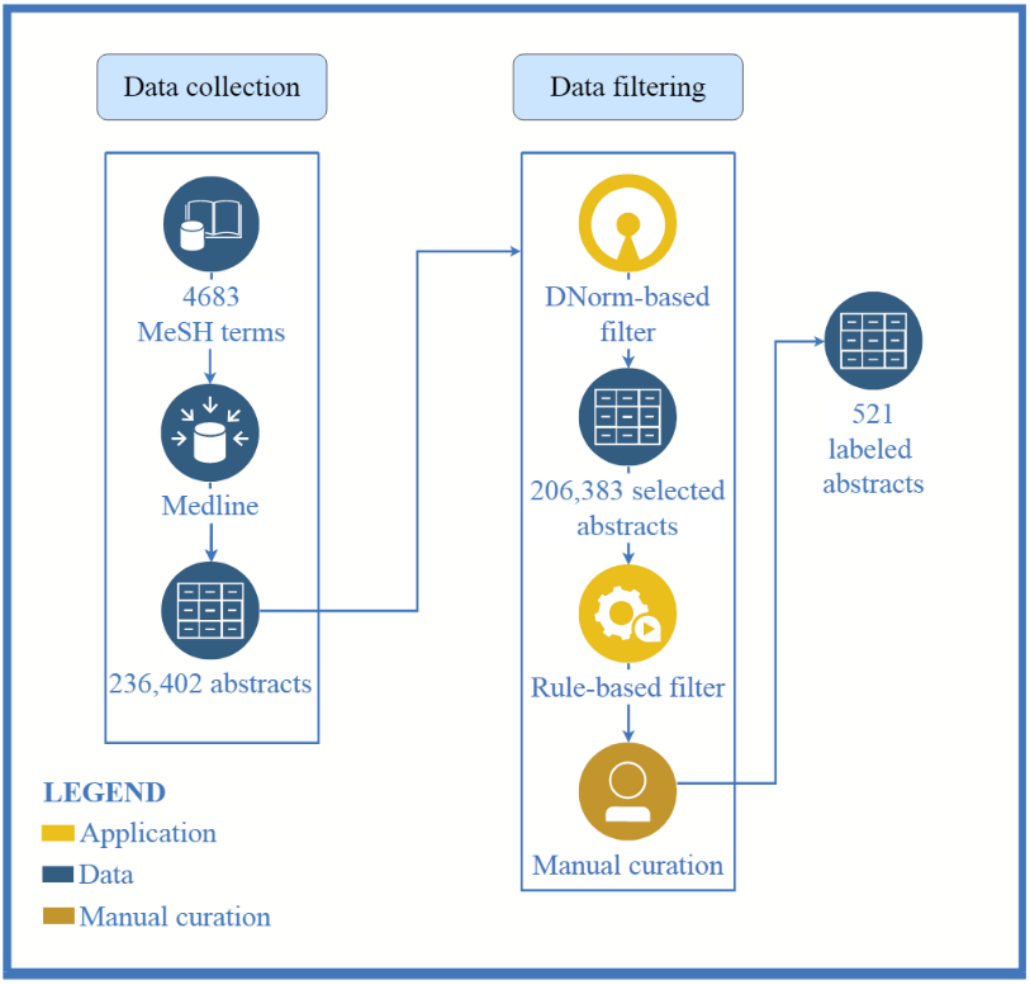
Study Process

In this section, we have first described the process of DDAE dataset construction. We then introduced our LC-CNN architecture in subsection *The Neural Network Architecture*. Further, we described each layer of LC-CNN in subsection *Composite Embedding Vector* to *Output Layer of Combined Sentence and Context Vector*. Finally, we introduced backward propagation for learning parameters of each layer.

Dataset Construction

The process of DDAE dataset construction is illustrated in Figure 2. Our DDAE dataset consisted of abstracts found in PubMed. To generate PubMed search queries related to DDA, we selected all disease nodes of the MeSH [41] tree whose tree number prefix starts with *C* and *F*, indicating diseases. We then selected any nodes related to human diseases. This produced a list of approximately 4700 disease names, which we then used to retrieve 236,000 abstracts whose titles or content contain one or more query terms.

Figure 2. Disease-disease association extraction dataset construction process. MeSH= Mesdical Subject Headings.



As some of these abstracts do not contain any DDAs, we used simple heuristic rules and a disease name recognizer/normalizer to select abstracts with a higher likelihood of containing DDAs. The process was as follows:

1. We selected only abstracts published from 2013 to 2017.
2. We used DNorm [42] to annotate disease mentions and their Medical Subject Heading (MeSH) IDs in these abstracts.
3. To ensure that the selected abstracts contain rich DDAs for training classifier, we removed abstracts that have fewer than 3 sentences that contain at least two different disease MeSH IDs.
4. To ensure the selected abstracts contain at least one DDA, we applied a DDA-adapted version of Lee et al's [28] dependency tree-based relation rules and removed any abstract not matched by any rule.
5. We randomly selected 521 abstracts from the remaining abstracts for annotation.

For the manual annotation step, we employed 2 biomedical specialists. Annotator 1 is a PhD candidate in a bioinformatics program, whereas Annotator 2 is a full-time research assistant in a hospital. Both have at least 6 years of biomedical experience. After agreeing on initial annotation guidelines (refer to Multimedia Appendix 1—Annotation Guideline), they used

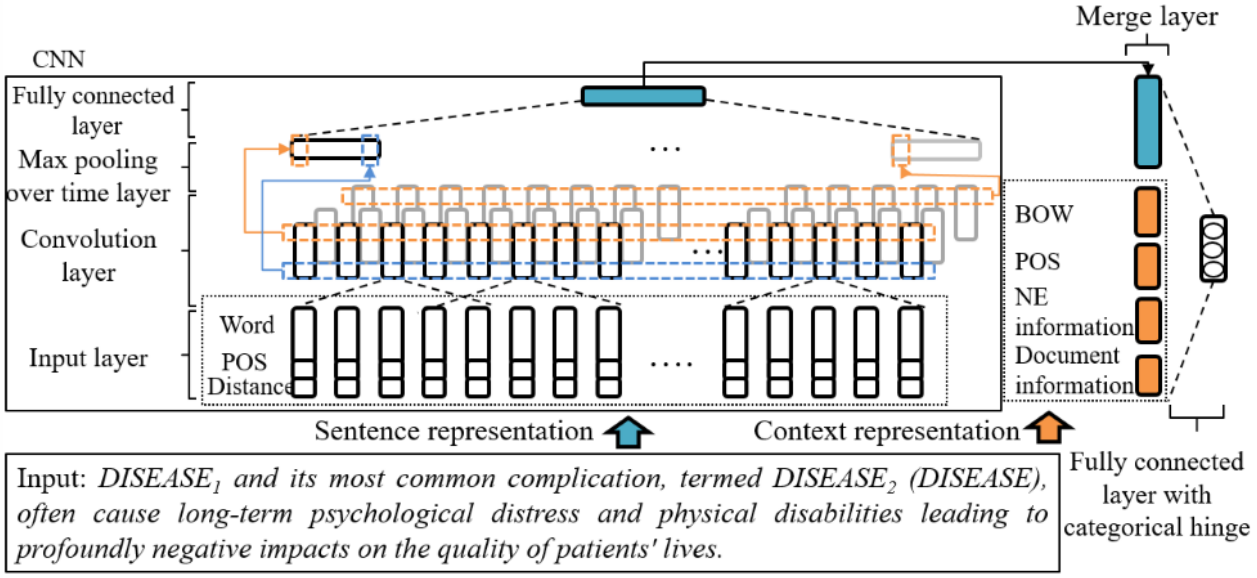
XSL•FO
RenderX

the brat rapid annotation tool [43] to annotate 10 abstracts and then compare results. In the first independent annotation processing, Cohen kappa value was 34%. Once both annotators agreed that all annotations that indicate consistency is satisfactory, they each annotated all remaining abstracts. Thus, each abstract was annotated independently twice. Inconsistent annotations were resolved afterward through discussion. The final Cohen kappa value was 76%.

The Neural Network Architecture

We formulated relation extraction as a classification problem in which, given a sentence containing a mention pair, the goal was to classify the pair into one of relation types. For classification, we propose an LC-CNN architecture as illustrated in Figure 3. The network is fed input in 2 forms: sentence representation and context representation (CR). Sentence representation is a n_{emb} x T matrix representing the sentence. n_{emb} and T are the length of composite embedding vector and the length of the sentence, respectively. The sentence representation uses only word embedding, part of speech (POS) encoding and Named Entity (NE) distance information, and parameters are learned through the next CNN and max-pool layers, which outputs an m-dimension sentence-level feature vector. The CR is a feature-rich n-dimension vector containing both syntactic and document-level features, such as whether the disease pair also appears in the title. Next, the m-dimension vector and the n-dimension vector are concatenated to form the final feature vector with (m+n) dimension. To compute the confidence of each relation type, the feature vector is fed into a fully connected layer, where we use a linear activation function with categorical hinge loss [44]. The output layer is a three-dimensional vector, with each dimension value representing the confidence of a predefined relation type.

Figure 3. Large margin context-aware convolutional neural network (LC-CNN) architecture. BOW: Bag of words; POS: Part of speech; NE: Named Entity.



Composite Embedding Vector

In a sentence, each word is represented as a composite embedding vector, as shown in Figure 3 (or in Multimedia Appendix 2). A composite embedding vector consists of 3 parts: word embedding, POS one-hot coding, and the distance between the word and disease pair. A matrix represents a sentence. The matrix contains the composite embedding vectors in the sentence, each placed in the order in its row. The sentence matrix is a matrix of size n_{emb} x T, wheren_{emb} is the dimension of the composite embedding vector and T represents the maximum length of the sentence in the dataset.

Word Embedding

The embedding of a word is a mapping of the word to a vector of real values. Generally, the word embeddings of semantically similar words are closer together in the vector space. Word embedding learned by neural networks has been demonstrated to be able to capture linguistic regularities and patterns in language models [40]. Therefore, it is commonly used in features in popular NN approaches, such as CNN [20,39] and long-short term memory (LSTM) [19]. In general, word embeddings are learned from large corpora such as Wikipedia or PubMed. For example, Pyysalo et al [45] applied word2vec to learn word embeddings from different texts, including Wikipedia, PubMed abstracts, and PubMed Central full-text papers, and developed a word-embedding lookup dictionary. Here, we employed their dictionary to generate word embeddings.

Part of Speech

The embedding of a word is a single vector and, therefore, cannot fully represent the multiple syntactic/semantic roles of a word like *good*, which can be either an adjective or a noun. The POS feature is designed to provide syntactic information (part of speech) to help the model separate the different semantic senses of a word. We used Zhao et al's [20] approach, in which similar POSs are assigned to the same group. We divided POSs into 11 groups, including adjectives, adverbs, articles, conjunctions, foreign words, interjections, nouns, prepositions, pronouns, punctuation, and verbs. If a word belongs to a POS

XSL•FO
RenderX

group, the corresponding bit value will be 1; otherwise, it will be 0.

Named Entity Distance

Zeng et al [46] proposed the use of NE distance (position features) to improve a CNN by keeping track of how close words are to the target nouns. We adopted their NE distance in this study. The NE distance feature is a two-dimensional vector (d_1, d_2). d_1 and d_2 represent the distance (number of words) between the current word and the first and second diseases of the pair.

Context Representation Layer

Contextual information, such as pair and document information, is very useful for classification and has been widely used in previous research. The purpose of using contextual representation is to introduce traditional contextual features into a neural network architecture through simple representation. We can then apply the fully connected layer to the context vector to obtain a condensed vector that combines 2 different representations.

Here are the features used in our contextual representation (refer to Multimedia Appendix 3 for more details).

Bag of Words

Word embedding has been shown to represent abstract information about words. However, word embedding can sometimes change the original meaning of a word. For example, *not* usually appears in negative relation statements. However, in the word2vec model trained on news, the 3 words most similar to *not* are *do*, *did*, and *anymore*. This violates our intuition that *don't*, *doesn't*, and *isn't* are more similar to *not* in the relation statement. As the embedded vector words of certain words may differ in the news and biomedicine domains, we use BOW features for context vector. Our BOW features include unigram, bigram, and surrounding diseases.

Part of Speech

The POS tags are commonly used for relation extraction. We used one-hot encoding to represent each word's POS tag type.

Named Entity Information

The number of diseases is useful when classifying relations. We used 3 different features to capture information, including the following:

1. The number of tokens between disease pairs.
2. The number of diseases between disease pairs.
3. The number of diseases in the sentence.

Document-Level Information

Biological papers usually follow a certain flow to describe their experimental and scientific findings. Therefore, article structure often provides valuable information about relations. We used 2 types of document-level feature, core pair and pair location. The core pair features indicate whether the current disease is a top-3 frequent disease pair in the article. The 3 most frequent pairs are treated as 3 features. The pair location feature is used to indicate the position of the sentence containing the relation in the article. If the sentence is the article title, it usually contains the subject of the article, which might be a relation investigated in the paper. Similarly, if the sentence is the last sentence of the abstract, it may summarize the main scientific discovery of the article. We used 3 binary features to represent relation pairs that appear in the title, the first sentence of the abstract, the last sentence of the abstract.

Output Layer of Combined Sentence and Context Vector

We used m_{concat} = [sr cr] to represent the concatenation of sentence representation sr and context representation cr. The size of the vector m_{concat} is $n_{concat} = n_{sr} + n_{cr}$. We then applied a fully connected layer to m_{concat} to obtain a 3D vector *out*, each value of which refers to the confidence of a predefined category.

$$out = W_{out} \times m_{concat} + Bias_{out}$$

W_{out} is a matrix with a size of n_{out} x n_{concat} and $Bias_{out}$ is a bias matrix with a size of n_{out} x 1. n_{out} is the number of predefined categories. out is the output of this fully connected layer and is defined as matrix W_{out} multiplied by matrix m_{concat}, plus bias $Bias_{out}$ Therefore, the size of out is n_{out} x 1. *out* is the final output of the prediction, and each dimension value of out refers to the score of its predefined category. out is calculated by a linear activation function, the values of out could be R × R × R.

Backward Propagation With Large Margin Loss

We used the following parameters:

1. k weight matrices, convWf each with a size of ne x f. Here, ne is the size of the input embedding vector of a word, and f is the window size of the filter.
2. k biases, convBf, each with size of ne x 1.
3. Weight matrix Wsr with a size of nsr x npool. Here, nsr is the output dimension of sentence vector and a hyperparameter.
4. Bias Biassr with a size of nsr x 1.
5. Weight matrix wout with a size of nout x nconcat. Here, nout is the number of relation types.
6. Bias BiasmaxF with a size of nout x 1.

In forward propagation, given those parameters, we calculated out with the methods mentioned in section *The Neural Network Architecture* to *Context Representation Layer*. In backward propagation, gradient descent is used to learn these parameters through minimizing the hinge loss of out. Given a sentence and its disease-disease pair, we defined a vector y as the pair's relation label vector. *y* is a 3D vector, and each dimension value of y represents the score of one relation type. According to the definition of hinge loss [44], the value is either -1 or 1. *1* means that the pair belongs to the relation type, whereas *–1* means it does not. Therefore, one value of the 3D vector must be *1*, and the others must be *–1*. For instance, the 3 vectors <1, –1, –1>, <–1, 1, –1>, and <–1, –1, 1> indicate that 3 vectors are *Positive*, *Negative*, and *Null*, respectively. We used the hinge loss function to evaluate the loss between prediction out and its truth label *y*; a larger loss indicates a larger gap between *out* and *y*. The hinge loss function is defined as follows:

$$loss(out, y) = sum_{i=1\ to\ nout}(max(1 - y_i * out_i, 0))/n_{out}$$

XSL•FO
RenderX

Here, y_i is the i-th dimension value of y. out is calculated by using forward propagation (sections *The Neural Network Architecture* to *Context Representation Layer*), and each dimension value of o refers to the prediction score of one predefined relation type. out_i is the i-th dimension value of out. out_i belongs to R. If out_i is a positive value, then the pair may be the i-th relation type. Otherwise, if out_i is a negative value, then the pair is less likely to be the i-th relation type.

In the equation, 1 is the value of the decision boundary. Ideally, $y_i * out_i$ will be larger than the decision boundary value. If y_i and out_i have the same sign, then $y_i * out_i$ will be a positive value belong to R. If $y_i * out_i$ is larger than the decision boundary value 1, then the loss(out, y) must be 0. If $y_i * out_i$ is smaller than the decision boundary value 1, then the loss(out, y) must be 1 - $y_i * out_i$ which is equal to the cost. If y_i and out_i are different signs, then $y_i * out_i$ will be a negative value ε R. Therefore, the loss(out, y) is a value greater than 1.

Given the training set

$$T=\{(x^{(i)},y^{(i)}) \mid i = 1,\ldots, N\},$$

$x^{(i)}$ is the i-th instance in the training set, $y^{(i)}$ is its label vector, and N is the number of training instances. Weight learning consists of the following optimization:

$$\text{argmin}_{convWf,\ convBf,\ Wst,\ Biassr,\ Wout,\ Biasout}\ \text{loss}(out,y)$$

Finally, mini-batch stochastic gradient descent [47] is applied to update the learned parameters in each iteration.

Results

Dataset

Currently, there are no available annotated datasets for training DDA extraction systems. To create one, we used our DDAE dataset development process, described in section *Dataset Construction*. The DDAE dataset consists of 521 annotated abstracts. After annotation, we used Cohen kappa coefficient to evaluate annotation consistency. The final kappa value is 76%, suggesting a high level of agreement.

For the experiments in this study, we divided our DDAE dataset into a training set of 400 abstracts and a test set of 121 abstracts. Before testing, we tuned the hyperparameters on one-third of abstracts randomly chosen from the training set called tuning set. Finally, our classifiers were trained on the whole training set and evaluated on the test set. A summary of the final DDAE dataset is shown in Table 1.

Table 1. Summary of disease-disease association extraction dataset.

Type	Training set, n	Test set, n	Total, n
Abstracts	400	121	521
Sentences	4820	1549	6369
Diseases	9522	2824	12,346
Total pairs	9086	2419	11,505
Positive pairs	2538	623	3161
Negative pairs	126	35	161
Null pairs	6422	1761	8183

Experiment Setup

We conducted 3 experiments to evaluate our LC-CNN. The first experiment was designed to measure the effects of different NN architectures and ML models. In the second experiment, we evaluated the effects of different approaches combining context features with NN methods. In the third experiment, we evaluated the effects of different word embeddings. The hyperparameters are listed in Multimedia Appendix 4. The performances of experiments on the tuning set can be found in Multimedia Appendix 5.

Our system is implemented on TensorFlow with Keras and runs on an Nvidia GTX 1080ti GPU. The process used in our experiments to generate the word-embedding model can be found in Multimedia Appendix 6.

Evaluation Metric

We used the F1 measure to evaluate system performance. The precision and recall are defined as given in Figure 4.

XSL•FO
RenderX

Figure 4. Precision and recall formula.

$$\text{Precision} = \frac{\text{number of correctly predicted positive and negative pairs}}{\text{number of predicted positive and negative pairs}}$$

$$\text{Recall} = \frac{\text{number of correctly predicted positive and negative pairs}}{\text{number of positive and negative pairs}}$$

$$\text{F1} - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Experiment 1—Performance Comparison With Other Models

The performance comparison between LC-CNN and different methods is listed in Table 2. It shows the performances on the tuning and test sets. The NN models (models 1 to 3) use only sentence representation. The $CR_{cross\text{-}entropy}$ and SVM methods use only CR. $CR_{cross\text{-}entropy}$ is implemented using a single hidden fully connected layer with the context vector as its input layer, and its architecture can be found in Multimedia Appendix 7. Furthermore, we also compared LC-CNN with LSTM and bidirectional LSTM (BiLSTM) models. They have been used in many relation extraction tasks, such as those seen in the studies by Hsieh et al and Zhao et al [19,48]. In our experiment, we were surprised to find that LSTM achieved the lowest F1 measure (65.02%) on the test set among all tested models. Furthermore, we also evaluated the performance of SCNN, Bidirectional Transformers for Language Understanding (BERT) [49], and BioBERT [50]. As we would like to compare the architecture of SCNN with LC-CNN, LC-CNN and SCNN use the same sentence representation, CR, and hinge loss function. The architecture of SCNN is illustrated in Multimedia Appendix 8.

As shown in Table 2, NN models trained on the entire training set (models 1 to 3) performed worse on the test set than on the tuning set. One potential reason is that the selected hyperparameters and parameters may be less likely to find unseen data, which could cause the hyperparameters and parameters of the NN models to overfit the tuning set. This problem is especially obvious in the LSTM and BiLSTM models. In contrast, $CR_{cross\text{-}entropy}$, SVM, and LC-CNN models trained on the entire training set with context information performed better on the test set than on the tuning set.

Furthermore, as shown in Table 2, CNN and $CR_{cross\text{-}entropy}$ performed similarly on the tuning set. The F1 measures of CNN and $CR_{cross\text{-}entropy}$ were 75.35% and 75.76%, respectively. CNN's recall rate was better than $CR_{cross\text{-}entropy}$'s recall rate by 2.84%, whereas $CR_{cross\text{-}entropy}$'s precision was 3.95% higher than that of CNN. This may be because the document feature provides $CR_{cross\text{-}entropy}$ with the information on the entire document, thus causing the model to generate fewer false positive cases. As CNN does not directly encode document information, it predicts more FPs. However, as CNN does not use any particular feature to separate positive, negative, and null relation pairs, it may be able to extract potential positive and negative pairs missed by $CR_{cross\text{-}entropy}$, resulting in higher recall rates. In addition, the SVM and $CR_{cross\text{-}entropy}$ use the same input features, but SVM mainly uses large margin for learning. The result shows that the SVM implemented with LibSVM [24] outperforms the $CR_{cross\text{-}entropy}$ by an F1 measure of 2.83%. Moreover, LC-CNN is able to combine the advantages of CNN and SVM to achieve the highest precision/recall/F1 measure among the tested models and outperforms SCNN, BERT, and BioBERT by F1 measures of 3.25%, 2.06%, and 1.91, respectively.

XSL•FO
RenderX

Table 2. Performances of different models. P: Precision; R: Recall; F: F1-Measure.

Input	Model	Tuning set			Test set		
		P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
SR[a]	LSTM[b]	65.53	70.15	67.76	66.13	63.95	65.02
SR	BiLSTM[c]	73.78	70.12	71.90	65.16	65.64	65.40
SR	CNN[d]	75.31	75.39	75.35	74.86	71.84	73.32
CR[e]	$CR_{cross\text{-}entropy}$	79.26	72.55	75.76	77.78	77.19	77.49
CR	SVM[f]	74.86	81.03	77.86	78.44	82.29	80.32
SR+CR	SCNN[g]	79.23	88.30	83.52	75.31	87.44	80.93
SR+CR	LC-CNN[h]	82.58	87.72	85.07	82.36	85.00	84.18
Sentence+pair	BERT	77.23	80.27	78.72	79.24	85.23	82.12
Sentence+pair	BioBERT	80.22	83.75	81.95	80.24	85.35	82.27

[a]SR: sentence representation.

[b]LSTM: long-short term memory.

[c]BiLSTM: bidirectional long-short term memory.

[d]CNN: convolutional neural network.

[e]CR: context representation.

[f]SVM: support vector machine.

[g]SCNN: syntax convolutional neural network.

[h]LC-CNN: Large margin context-aware convolutional neural network.

Experiment 2—Effect of Different Uses of Context Information

To demonstrate the advantage of integrating CNN and context information in a single LC-CNN architecture, we evaluated different ways of combining them. The performances of these combinations are shown in Table 3. There are 3 baseline models that use only either CNN or context information. Baselines 1 to 3 are $CR_{cross\text{-}entropy}$, SVM, and CNN and are used in Experiment 1. Only $CR_{cross\text{-}entropy}$ and SVM use contextual information.

SVM + CNN is an intuitive method in which the output vector of CNN is considered an additional feature vector of SVM, and its architecture is illustrated in Multimedia Appendix 9. As shown in Table 3, the F1-measure of SVM + CNN is significantly lower than that of SVM by 6.98%. One possible reason is that the CNN used in SVM + CNN is adjusted on the tuning set, so it causes the model to overfit CNN predictions, making it difficult to learn feature weights well.

We designed the LC-CNN to learn the model in a single stage. LC-CNN achieves an F1 measure of 84.18% on the test set, which is the highest score among all methods and outperform SCNN. The results showed that LC-CNN can learn CNN and context information well in a single stage.

Table 3. Performance of combined classifiers. P: Precision; R: Recall; F: F1-Measure.

Method	P (%)	R (%)	F (%)
Baseline 1 ($CR^{a}_{cross\text{-}entropy}$)	77.78	77.19	77.49
Baseline 2 (SVM[b])	78.44	82.29	80.32
Baseline 3 (CNN[c])	74.86	71.84	73.32
SCNN[d]	75.31	87.44	80.93
LC-CNN[e]	82.36	85.00	84.18
SVM+CNN (2-stage)	74.45	72.26	73.34

[a]CR: context representation.

[b]SVM: support vector machine.

[c]CNN: convolutional neural network.

[d]SCNN: syntax convolutional neural network.

[e]LC-CNN: large margin context-aware convolutional neural network.

XSL•FO
RenderX

Experiment 3—Effect of Composite Embedding Vectors on Large Margin Context-Aware Convolutional Neural Networks

In our third experiment, we evaluated the effect of different composite embedding vectors on LC-CNN (the effect of different features on LC-CNN can be found in Multimedia Appendix 10). The performance on the test set is shown in Table 4. We compared 3 different word embeddings. The word embeddings of LC-CNN$_{PubMed}$ are from Pyysalo et al [45], who learned them from Wikipedia, PubMed abstracts, and PubMed Central full texts. The word embeddings of LC-CNN$_{News}$ are learned from Google News using word2vec. In contrast, LC-CNN$_{no\ pretrain}$ does not use any pretrained word embeddings. Its word embeddings are treated as parameters and are learned through training LC-CNN$_{no\ pretrain}$ on the training set. Moreover, we also evaluated the effect of 3 different embedding features (word embedding, POS, and NE distance) by removing them individually from the LC-CNN$_{PubMed}$.

As shown in Table 4, the model with PubMed word embeddings (LC-CNN$_{PubMed}$) outperformed LC-CNN$_{News}$ and LC-CNN$_{no\ pretrain}$. In addition, our removal tests indicated that both POS and NE distance have strong impact on performance.

Table 4. The effect of different composite embedding vectors on large margin context-aware convolutional neural network performance. P: Precision; R: Recall; F: F1-Measure.

Method	P (%)	R (%)	F (%)
LC-CNN[a]$_{PubMed}$	82.36	85.00	84.18
LC-CNN$_{news}$	79.80	87.36	83.41
LC-CNN$_{no\ pretrain}$	77.83	86.58	81.97
LC-CNN$_{PubMed}$—POS[b]	80.23	84.26	82.19
LC-CNN$_{PubMed}$—distance	77.68	87.08	82.11

[a]LC-CNN: large margin context-aware convolutional neural network.

[b]POS: part of speech.

Discussion

Large Margin Context-Aware Convolutional Neural Network Error Cases Distribution

We randomly sampled approximately 60 error cases of the LC-CNN's predictions, and their distribution is illustrated in Table 5. FP and FN denote the false positive and false negative cases, respectively. As shown in Table 5, the *symptom/subclass* is a common error category in the FPs, and it contains a ratio of 28% in the sampled error cases. The *symptom/subclass* indicates that a disease is either a subclass or a symptom of another disease in the FP/FN disease pair. For example, an FP case: "Other large-artery aneurysms, including carotid, subclavian, and *iliac artery aneurysms*$_{DISEASE1}$, have also been associated with *Marfan syndrome*$_{DISEASE2}$. --- PMID:23891252" [51].

Here, the *carotid*, *subclavian*, and *iliac artery aneurysms* are 3 *Traumatic syndrome* for *Marfan syndrome*. They are the symptoms of *Marfan syndrome*. The symptom is not included in our DDA definition. Therefore, *iliac artery aneurysms*$_{DISEASE1}$ does not have a relation with the *Marfan syndrome*$_{DISEASE2}$. However, in this case, the keyword phrase *been associated with* makes LC-CNN predict it as positive relation, and thus results in an FP case.

In contrast with the FP cases, the FN cases are relatively sparse, and most of them cannot be categorized. For example, "CONCLUSION: *Cataract*$_{DISEASE1}$, uncorrected refractive error, and fundus diseases are ranked in the top 3 causes of moderate to severe *visual impairment* $_{DISEASE2}$ and blindness in adults aged 50 years or more in rural Shandong Province. --- PMID: 23714032" [52].

In the sentence, *Cataract* is one cause of *visual impairment*; however, the description also lists the other 2 diseases that cause *visual impairment*. For example, "it can be associated with any type of *vision loss*$_{DISEASE1}$ including that related to *maculardegeneration*$_{DISEASE2}$, *corneal disease*$_{DISEASE3}$, *diabetic retinopathy*$_{DISEASE4}$, and *occipital infarct*$_{DISEASE5}$. --- PMID:24339694" [53].

Here, the LC-CNN correctly identifies the relation between DISEASE1 and DISEASE2. However, it failed to identify the relations between DISEASE1 and the other diseases (DISEASE3, DISEASE4, and DISEASE5).

XSL•FO
RenderX

Table 5. The distribution of sampled large margin context-aware convolutional neural network error cases.

Type, category	Description	Ratio (%)
FP[a]		
Symptom/subclass	A disease is a symptom/subclass of another disease	28
Co-occur	2 diseases co-occur in the sentence	24
Negation	2 diseases are negative relation	8
Others	The error cannot be categorized	40
FN[b]		
Simple FN	There is an obvious relation keyword for disease pair	23
Negation	2 diseases are negative relation	16
Others	No obvious relation keyword, or the statements of DDA[c] are too complicated	61

[a]FP: False positive.

[b]FN: False negative.

[c]DDA: disease-disease association.

The Result of Using Automatic Annotated Disease Mentions

In our experiment, we used the manually annotated disease mentions, which may not reflect the actual performance of the fully automated DDAE task. Hence, we conducted an experiment, in which we used the TaggerOne [9], a state-of-the-art disease mention recognizer/normalizer, to annotate the disease mentions of the test set. Then we used the LC-CNN to extract DDAs from the TaggerOne-annotated test set. As the boundaries of some predicted mentions may be inconsistent with the gold mentions, we used an approximate matching to allow this. In the fully automatic process, the LC-CNN achieved a Precision/Recall/F1 measure of 75.28/55.03/63.57, respectively. The recall is significantly lower because it failed to recognize some diseases. However, the performance is reasonable but 7.08% lower than that of the semiautomatic process (using gold disease mentions).

Principal Findings

Our objective was to develop a DDAE dataset and a neural network–based approach to extract DDAs. In our experiments, the LC-CNN trained on our dataset achieved an F1 measure of 84.18%. We also compared LC-CNN with common NN models including CNN, Bi-LSTM, and SVM. The results showed that the LSTM and BiLSTM models achieved relatively lower F1 measures of 65.02% and 65.40%, respectively. This may be because the hyperparameters and parameters tend to overfit the training set. The CNN and SVM models achieved relatively higher F1 measures of 73.32% and 77.49%, respectively, but LC-CNN still outperformed all tested methods. In addition, the results showed that the 2-stage *SVM + CNN* model scored significantly lower in terms of F1 than SVM and LC-CNN by 6.98% and 10.84%, respectively. This suggests that simple methods may achieve better results than complex ones. Furthermore, in our experiments, the model with PubMed word embeddings ($LC\text{-}CNN_{PubMed}$) outperformed the $LC\text{-}CNN_{News}$ and $LC\text{-}CNN_{no\ pretrain}$ models, indicating that PubMed word embeddings may be more compatible with our DDAE dataset.

Conclusions

In this paper, we proposed a text-mining approach for automatically extracting DDAs from abstracts. We collected disease-related abstracts from PubMed and annotated the first publicly available DDAE dataset consisting of 521 abstracts and 3322 disease-disease pairs. Moreover, to extract DDAs, we used several different ML models, including BiLSTM, CNN, and SVM. We also evaluated the effect of combining CNN and context features. Finally, we implemented a novel neural network called LC-CNN to integrate context features and CNN through the large margin function. Our experiment results showed that LC-CNN achieved an F1 measure of 84.18%, the highest among the tested models.

Acknowledgments

The authors would like to thank the Ministry of Science and Technology, Taiwan, for financially supporting this research. This study was funded by the Ministry of Science and Technology, Taiwan, [105-2221-E-008-115-MY3] and [103-2221-E-008-044-MY3].

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO
RenderX

Annotation Guideline.
[PDF File (Adobe PDF File), 764 KB - medinform_v7i4e14502_app1.pdf]

Multimedia Appendix 2
Composite Embedding Vector.
[PDF File (Adobe PDF File), 92 KB - medinform_v7i4e14502_app2.pdf]

Multimedia Appendix 3
Context Representation Layer.
[PDF File (Adobe PDF File), 151 KB - medinform_v7i4e14502_app3.pdf]

Multimedia Appendix 4
Hyperparameters.
[PDF File (Adobe PDF File), 78 KB - medinform_v7i4e14502_app4.pdf]

Multimedia Appendix 5
Performances on Tuning Set.
[PDF File (Adobe PDF File), 81 KB - medinform_v7i4e14502_app5.pdf]

Multimedia Appendix 6
Generating Word Embedding.
[PDF File (Adobe PDF File), 133 KB - medinform_v7i4e14502_app6.pdf]

Multimedia Appendix 7
Architecture of CRcross-entropy.
[PDF File (Adobe PDF File), 217 KB - medinform_v7i4e14502_app7.pdf]

Multimedia Appendix 8
Architecture of SCNN.
[PDF File (Adobe PDF File), 196 KB - medinform_v7i4e14502_app8.pdf]

Multimedia Appendix 9
Architecture of SVM + CNN.
[PDF File (Adobe PDF File), 191 KB - medinform_v7i4e14502_app9.pdf]

Multimedia Appendix 10
Effect of Different Features on LC-CNN.
[PDF File (Adobe PDF File), 78 KB - medinform_v7i4e14502_app10.pdf]

Multimedia Appendix 11
DDAE Dataset.
[ZIP File (Zip Archive), 864 KB - medinform_v7i4e14502_app11.zip]

References

1. Žitnik M, Janjić V, Larminie C, Zupan B, Pržulj N. Discovering disease-disease associations by fusing systems-level molecular data. Sci Rep 2013 Nov 15;3:3202 [FREE Full text] [doi: 10.1038/srep03202] [Medline: 24232732]
2. Liu C, Tseng Y, Li W, Wu C, Mayzus I, Rzhetsky A, et al. DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. Nucleic Acids Res 2014 Jul;42(Web Server issue):W137-W146 [FREE Full text] [doi: 10.1093/nar/gku412] [Medline: 24895436]
3. Bang S, Kim J, Shin H. Causality modeling for directed disease network. Bioinformatics 2016 Sep 1;32(17):i437-i444. [doi: 10.1093/bioinformatics/btw439] [Medline: 27587660]
4. Sun K, Gonçalves JP, Larminie C, Przulj N. Predicting disease associations via biological network analysis. BMC Bioinformatics 2014 Sep 17;15:304 [FREE Full text] [doi: 10.1186/1471-2105-15-304] [Medline: 25228247]
5. Yang J, Wu SJ, Yang SY, Peng JW, Wang SN, Wang FY, et al. DNetDB: The human disease network database based on dysfunctional regulation mechanism. BMC Syst Biol 2016 May 21;10(1):36 [FREE Full text] [doi: 10.1186/s12918-016-0280-5] [Medline: 27209279]

XSL•FO
RenderX

6. Chawla A, Chawla R, Jaggi S. Microvasular and macrovascular complications in diabetes mellitus: distinct or continuum? Indian J Endocrinol Metab 2016;20(4):546-551 [FREE Full text] [doi: 10.4103/2230-8210.183480] [Medline: 27366724]
7. Leon BM, Maddox TM. Diabetes and cardiovascular disease: epidemiology, biological mechanisms, treatment recommendations and future research. World J Diabetes 2015 Oct 10;6(13):1246-1258 [FREE Full text] [doi: 10.4239/wjd.v6.i13.1246] [Medline: 26468341]
8. Tang B, Cao H, Wang X, Chen Q, Xu H. Evaluating word representation features in biomedical named entity recognition tasks. Biomed Res Int 2014;2014:240403-240406 [FREE Full text] [doi: 10.1155/2014/240403] [Medline: 24729964]
9. Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. Bioinformatics 2016 Sep 15;32(18):2839-2846 [FREE Full text] [doi: 10.1093/bioinformatics/btw343] [Medline: 27283952]
10. Zhu Q, Li X, Conesa A, Pereira C. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. Bioinformatics 2018 May 1;34(9):1547-1554. [doi: 10.1093/bioinformatics/btx815] [Medline: 29272325]
11. Luo ZH, Shi MW, Yang Z, Zhang HY, Chen ZX. pyMeSHSim: an integrative python package to realize biomedical named entity recognition, normalization and comparison. bioRxiv 2018:459172. [doi: 10.1101/459172]
12. Wei CH, Kao HY, Lu Z. GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. Biomed Res Int 2015;2015:918710 [FREE Full text] [doi: 10.1155/2015/918710] [Medline: 26380306]
13. Lai P, Huang M, Yang T, Hsu W, Tsai RT. Statistical principle-based approach for gene and protein related object recognition. J Cheminform 2018 Dec 17;10(1):64 [FREE Full text] [doi: 10.1186/s13321-018-0314-7] [Medline: 30560325]
14. Leaman R, Wei C, Zou C, Lu Z. Mining chemical patents with an ensemble of open systems. Database (Oxford) 2016;2016:baw065 [FREE Full text] [doi: 10.1093/database/baw065] [Medline: 27173521]
15. Tsai RT, Hsiao YC, Lai P. NERChem: adapting NERBio to chemical patents via full-token features and named entity feature with chemical sub-class composition. Database (Oxford) 2016 Oct 25;2016:baw135 [FREE Full text] [doi: 10.1093/database/baw135] [Medline: 31414701]
16. Li L, Guo R, Jiang Z, Huang D. An approach to improve kernel-based protein-protein interaction extraction by learning from large-scale network data. Methods 2015 Jul 15;83:44-50. [doi: 10.1016/j.ymeth.2015.03.026] [Medline: 25864936]
17. Peng Y, Wei CH, Lu Z. Improving chemical disease relation extraction with rich features and weakly labeled data. J Cheminform 2016;8:53 [FREE Full text] [doi: 10.1186/s13321-016-0165-z] [Medline: 28316651]
18. Ravikumar K, Rastegar-Mojarad M, Liu H. BELMiner: adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences. Database (Oxford) 2017 Jan 1;2017(1):baw156 [FREE Full text] [doi: 10.1093/database/baw156] [Medline: 28365720]
19. Hsieh Y, Chang Y, Chang N, Hsu W. Identifying Protein-protein Interactions in Biomedical Literature using Recurrent Neural Networks with Long Short-Term Memory. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2017 Presented at: IJCNLP'17; 2017; Taipei, Taiwan p. 240-245.
20. Zhao Z, Yang Z, Luo L, Lin H, Wang J. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. Bioinformatics 2016 Nov 15;32(22):3444-3453 [FREE Full text] [doi: 10.1093/bioinformatics/btw486] [Medline: 27466626]
21. Lai P, Lo YY, Huang MS, Hsiao YC, Tsai RT. BelSmile: a biomedical semantic role labeling approach for extracting biological expression language from text. Database (Oxford) 2016;2016:- [FREE Full text] [doi: 10.1093/database/baw064] [Medline: 27173520]
22. Hoyt CT, Domingo-Fernández D, Hofmann-Apitius M. BEL Commons: an environment for exploration and analysis of networks encoded in Biological Expression Language. Database 2018;2018:-. [doi: 10.1101/288274]
23. Xu R, Li L, Wang Q. dRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text. BMC Bioinformatics 2014 Apr 12;15:105 [FREE Full text] [doi: 10.1186/1471-2105-15-105] [Medline: 24725842]
24. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol 2011 Apr 1;2(3):1-27. [doi: 10.1145/1961189.1961199]
25. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification With Deep Convolutional Neural Networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. 2012 Presented at: NIPS'12; December 3-6, 2012; Lake Tahoe, Nevada p. 1097-1105. [doi: 10.1145/3065386]
26. Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inform 2014 Feb;47:1-10 [FREE Full text] [doi: 10.1016/j.jbi.2013.12.006] [Medline: 24393765]
27. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005 Jan 1;33(Database issue):D514-D517 [FREE Full text] [doi: 10.1093/nar/gki033] [Medline: 15608251]
28. Lee J, Kim D, Lee S, Lee S, Kang J. On the efficacy of per-relation basis performance evaluation for PPI extraction and a high-precision rule-based approach. BMC Med Inform Decis Mak 2013;13 Suppl 1:S7 [FREE Full text] [doi: 10.1186/1472-6947-13-S1-S7] [Medline: 23566263]
29. Bunescu R, Ge R, Kate RJ, Marcotte EM, Mooney RJ, Ramani AK, et al. Comparative experiments on learning information extractors for proteins and their interactions. Artif Intell Med 2005 Feb;33(2):139-155. [doi: 10.1016/j.artmed.2004.07.016] [Medline: 15811782]

XSL•FO
RenderX

30. Nguyen NT, Miwa M, Tsuruoka Y, Chikayama T, Tojo S. Wide-coverage relation extraction from MEDLINE using deep syntax. BMC Bioinformatics 2015 Apr 1;16:107 [FREE Full text] [doi: 10.1186/s12859-015-0538-8] [Medline: 25887686]
31. Miyao Y, Tsujii J. Feature forest models for probabilistic HPSG parsing. Comput Linguist 2008 Mar;34(1):35-80. [doi: 10.1162/coli.2008.34.1.35]
32. Zhang Y, Lin H, Yang Z, Wang J, Li Y. Hash subgraph pairwise kernel for protein-protein interaction extraction. IEEE/ACM Trans Comput Biol Bioinform 2012;9(4):1190-1202. [doi: 10.1109/TCBB.2012.50] [Medline: 22595237]
33. Moschitti A. Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. In: Proceedings of the 17th European conference on Machine Learning. 2006 Presented at: ECML'06; September 18-22, 2006; Berlin, Germany p. 318-329. [doi: 10.1007/11871842_32]
34. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. J Mach Learn Res 2011;12:2493-2537 [FREE Full text]
35. Wei C, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J, et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database (Oxford) 2016;2016 [FREE Full text] [doi: 10.1093/database/baw032] [Medline: 26994911]
36. Liu S, Tang B, Chen Q, Wang X, Fan X. Feature engineering for drug name recognition in biomedical texts: feature conjunction and feature selection. Comput Math Methods Med 2015;2015:913489 [FREE Full text] [doi: 10.1155/2015/913489] [Medline: 25861377]
37. Thomas P, Neves M, Rocktäschel T, Leser U. WBI-DDI: Drug-Drug Interaction Extraction using Majority Voting. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). 2013 Presented at: SEMEVAL'13; 2013; Atlanta, Georgia, USA p. 628-635.
38. Thuy Phan TT, Ohkawa T. Protein-protein interaction extraction with feature selection by evaluating contribution levels of groups consisting of related features. BMC Bioinformatics 2016 Jul 25;17(Suppl 7):246 [FREE Full text] [doi: 10.1186/s12859-016-1100-z] [Medline: 27454611]
39. Peng Y, Lu Z. Deep Learning for Extracting Protein-Protein Interactions From Biomedical Literature. In: Proceedings of the Biomedical Natural Language Processing Workshop (2017). 2017 Presented at: BioNLP'17; 2017; Vancouver, Canada p. 29-38. [doi: 10.18653/v1/w17-2304]
40. Mikolov T, Yih W, Zweig G. Linguistic Regularities in Continuous Space Word Representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013 Presented at: NAACL'13; 2013; Atlanta, Georgia.
41. Lipscomb CE. Medical Subject Headings (MeSH). Bull Med Libr Assoc 2000 Jul;88(3):265-266 [FREE Full text] [Medline: 10928714]
42. Leaman R, Dogan RI, Lu Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics 2013 Nov 15;29(22):2909-2917 [FREE Full text] [doi: 10.1093/bioinformatics/btt474] [Medline: 23969135]
43. Stenetorp P, Pyysalo S, Topic G, Ohta T, Ananiadou S, Tsujii J. BRAT: A Web-based Tool for NLP-Assisted Text Annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012 Presented at: EACL'12; April 23-27, 2012; Avignon, France p. 102-107.
44. Tang Y. Deep learning using linear support vector machines. arXiv preprint arXiv 2013:13060239 [FREE Full text]
45. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. editors. Distributional Semantics Resources for Biomedical Text Processing 2013.
46. Zeng D, Liu K, Lai S, Zhou G, Zhao J. Relation Classification via Convolutional Deep Neural Network. In: Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers. 2014 Presented at: COLING'14; August 23-29, 2014; Dublin, Ireland p. 2335-2344.
47. Li M, Zhang T, Chen Y, Smola A. Efficient Mini-Batch Training for Stochastic Optimization. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014 Presented at: KDD'14; August 24-27, 2014; New York, New York, USA p. 661-670. [doi: 10.1145/2623330.2623612]
48. Miwa M, Bansal M. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016 Presented at: ACL'16; August 7-12, 2016; Berlin, Germany p. 1105-1116. [doi: 10.18653/v1/p16-1105]
49. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv 2018:181004805 [FREE Full text]
50. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2019 Sep 10. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
51. Awais M, Williams DM, Deeb GM, Shea MJ. Aneurysms of medium-sized arteries in Marfan syndrome. Ann Vasc Surg 2013 Nov;27(8):1188.e5-1188.e7. [doi: 10.1016/j.avsg.2012.12.002] [Medline: 23891252]
52. Li Y, Bi HS, Wang LH, Wang T, Yang SY, Liu LP, et al. [Causes of moderate to severe visual impairment and blindness in population aged 50 years or more in rural Shandong province]. Zhonghua Yan Ke Za Zhi 2013 Feb;49(2):144-150. [Medline: 23714032]

XSL•FO
RenderX

53. Zhang J, Waisbren E, Hashemi N, Lee AG. Visual hallucinations (Charles Bonnet syndrome) associated with neurosarcoidosis. Middle East Afr J Ophthalmol 2013;20(4):369-371 [FREE Full text] [doi: 10.4103/0974-9233.119997] [Medline: 24339694]

Abbreviations

BERT: Bidirectional Transformers for Language Understanding
BiLSTM: bidirectional long-short term memory
BOW: bag of words
CDR: chemical-disease relation
CNN: convolutional neural network
CR: context representation
DDA: disease-disease association
DDAE: disease-disease association extraction
dRiskKB: disease-disease risk relationship knowledge base
LSTM: long-short term memory
McDepCNN: multichannel dependency-based convolutional neural network
MeSH: Medical Subject Headings
ML: machine learning
NCBI: National Center for Biotechnology Information
PAS: predicate-argument structure
POS: part of speech
PPI: protein-protein interaction
SCNN: syntax convolutional neural network
SVM: support vector machine

Edited by C Lovis; submitted 26.04.19; peer-reviewed by CH Wei, L Zhang, G Lim, B Polepalli Ramesh; comments to author 31.05.19; revised version received 26.07.19; accepted 11.08.19; published 26.11.19.

Please cite as:
Lai PT, Lu WL, Kuo TR, Chung CR, Han JC, Tsai RTH, Horng JT
Using a Large Margin Context-Aware Convolutional Neural Network to Automatically Extract Disease-Disease Association from Literature: Comparative Analytic Study
JMIR Med Inform 2019;7(4):e14502
URL: http://medinform.jmir.org/2019/4/e14502/
doi:10.2196/14502
PMID:31769759

©Po-Ting Lai, Wei-Liang Lu, Ting-Rung Kuo, Chia-Ru Chung, Jen-Chieh Han, Richard Tzong-Han Tsai, Jorng-Tzong Horng. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 26.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact of Automatic Query Generation and Quality Recognition Using Deep Learning to Curate Evidence From Biomedical Literature: Empirical Study

Muhammad Afzal[1,2], PhD; Maqbool Hussain[1], PhD; Khalid Mahmood Malik[2], PhD; Sungyoung Lee[3], PhD

[1]Department of Software, Sejong University, Seoul, Republic of Korea

[2]Department of Computer Science and Engineering, Oakland University, Rochester, MI, United States

[3]Department of Computer Science and Engineering, Kyung Hee University, Yongin, Republic of Korea

Corresponding Author:
Sungyoung Lee, PhD
Department of Computer Science and Engineering
Kyung Hee University
Room 313
Yongin, 446-701
Republic of Korea
Phone: 82 312012514
Fax: 82 312022520
Email: sylee@oslab.khu.ac.kr

Abstract

Background: The quality of health care is continuously improving and is expected to improve further because of the advancement of machine learning and knowledge-based techniques along with innovation and availability of wearable sensors. With these advancements, health care professionals are now becoming more interested and involved in seeking scientific research evidence from external sources for decision making relevant to medical diagnosis, treatments, and prognosis. Not much work has been done to develop methods for unobtrusive and seamless curation of data from the biomedical literature.

Objective: This study aimed to design a framework that can enable bringing quality publications intelligently to the users' desk to assist medical practitioners in answering clinical questions and fulfilling their informational needs.

Methods: The proposed framework consists of methods for efficient biomedical literature curation, including the automatic construction of a well-built question, the recognition of evidence quality by proposing extended quality recognition model (E-QRM), and the ranking and summarization of the extracted evidence.

Results: Unlike previous works, the proposed framework systematically integrates the echelons of biomedical literature curation by including methods for searching queries, content quality assessments, and ranking and summarization. Using an ensemble approach, our high-impact classifier E-QRM obtained significantly improved accuracy than the existing quality recognition model (1723/1894, 90.97% vs 1462/1894, 77.21%).

Conclusions: Our proposed methods and evaluation demonstrate the validity and rigorousness of the results, which can be used in different applications, including evidence-based medicine, precision medicine, and medical education.

(JMIR Med Inform 2019;7(4):e13430) doi:10.2196/13430

KEYWORDS

data curation; evidence-based medicine; clinical decision support systems; precision medicine; biomedical research; machine learning; deep learning

Introduction

Objective and Contributions

Personalized health care and wellness management have rapidly grown during recent years because of the increase in data influx, the development of innovative tools, and the advancement of artificial intelligence techniques. These innovations can engage patients and offer additional modalities in the treatment of chronic diseases [1]. In addition, with the advent of the next-generation sequencing and the widespread use of electronic health records (EHRs), clinicians and researchers have the

XSL•FO
RenderX

opportunity to have a wealth of data and the precise characterization of individual patient genotypes and phenotypes [2]. It is now evident that the research on internet health information–seeking behavior is on the rise [3].

Furthermore, people's interest in seeking the support of scientific research evidence is increasing daily for their level of satisfaction over medical decisions or advice, and it keeps them aware of the research about the matter. Clinicians seek for external evidences to make informed clinical decisions, particularly when internal evidences (information derived from unicenter data) are insufficient because of lack of required data. Likewise, medical researchers and students are interested in the external evidences to educate themselves on the substance of a medical problem, whereas the patients could use such evidences for their own awareness and comparative analysis of available treatments. Fortunately, an overwhelming amount of biomedical information is available in the form of scientific publications, which can be retrieved to support the process of medical decision making and for self-awareness. PubMed, which is a search engine for biomedical literature, can provide access to a set of more than 27 million articles from more than 7000 journals, including full text for about 4 million of these articles [4]. However, the current process of retrieving research publications from the external biomedical literature is a daunting task and is largely done manually, which requires not only a high level of expertise but also time and money. As the demand for evidence-based medicine (EBM) is increasing, it is important to lower the costs to identify and evaluate the best evidence. Little has been done to improve the overall efficiency of curating the quality evidences automatically from the biomedical literature until recently. One of the major challenges in this regard is to design the search query from the input information and to embed the user context in an automatic and intelligent manner to save time and cost. In addition, the low quality of the articles from where the evidence is retrieved for the decisions adds further to the challenge of an automated acquisition of evidence. Moreover, the results are summarized and ranked majorly with manual efforts.

In this paper, we contributed to the design of a comprehensive framework architecture to achieve the goal of curating biomedical literature and mining data from scientific publications to construct precise evidence to assist medical practitioners, researchers, medical students, and patients in the clinical decision-making process. The proposed framework consists of several methods for automating the process of biomedical curation. The main contributions of this paper are as follows:

- It presents the design of a comprehensive framework for biomedical literature curation. It describes proposed architecture in detail, which includes designs for methods of well-built automatic query construction, evidence quality recognition, and article summarization and ranking.
- It describes the proposed process of the construction of a well-built query. We designed a set of methods and guidelines to construct a well-structured question from the input information in a standard format for a better user understandability and content categorization.
- It presents the design of proposed extended quality recognition model (E-QRM) that identifies scientifically sound publications on the basis of content rigorousness. We developed and compared a set of machine and deep learning (DL) models with a higher level of precision as evaluation criteria.
- It offers methods for contextual ranking and summarization. We designed a cross-context interpretation model for ranking the publications based on the context captured from the input information, the user of the system, and the articles that are retrieved. In addition, we propose a conceptual model for summarization of the results based on input information.

Background and Motivation

For evidentiary support, medical professionals mostly rely on the publicly available searching services, such as PubMed [5], Google, UptoDate [6], and other search engines. These search engines are reliable, but they need to be integrated with a health care information system (HIS) in a way to make the process of evidence retrieval seamless and meaningful. In addition, an HIS is required to evaluate the retrieved evidence for quality rather than relying on a search engine's built-in evaluation mechanisms.

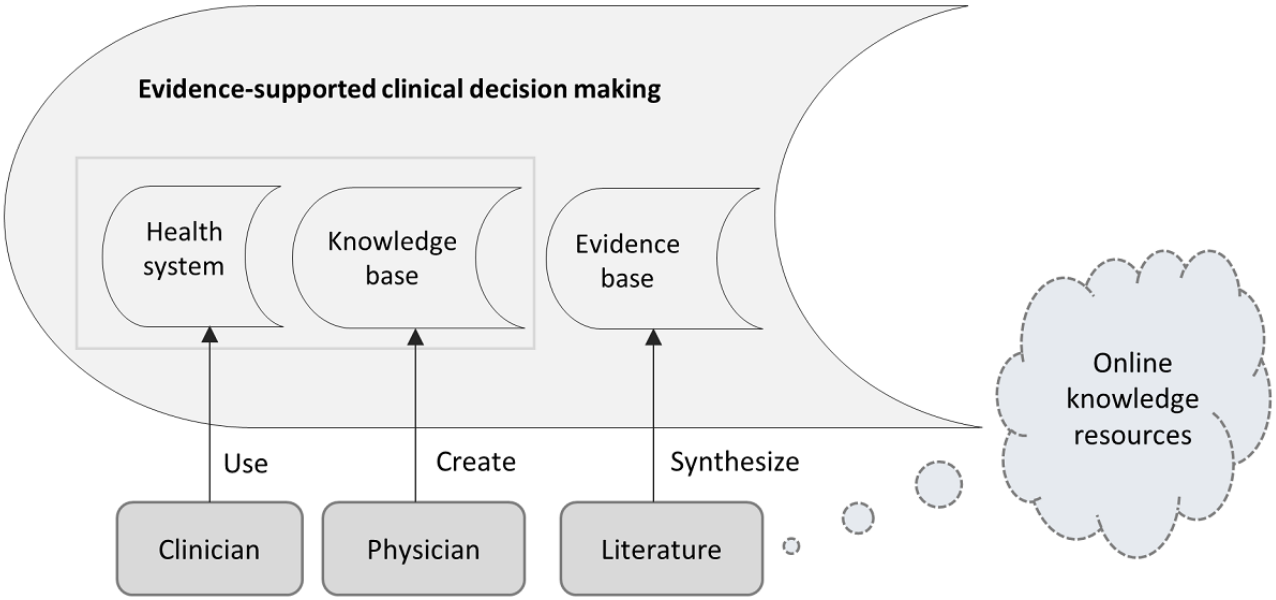
Some of today's HISs are equipped with the knowledge base (KB) of a clinical decision support system (CDSS), which provides additional support to automate the evidence retrieval from external sources in following ways: (1) it aids to automate query construction process for evidence retrieval by offering knowledge rules that consist of patient information with established logical connections and (2) it assists to enrich query for evidence with metadata such as the purpose or the query type information to improve the quality of evidence extraction. The query type information shows the purpose for which a CDSS is developed, such as a treatment plan or diagnosis recommendations.

Figure 1 shows the interaction among a health system, KB of CDSS, and the system for extraction of external evidence resources. The health system manages the patient records to be used by the clinician, and the KB of a CDSS is created with the support of expert clinicians either through directly authored rules or the machine learning (ML)–based data-driven approaches [7]. The evidence-based subsystem shows the appraised evidence synthesized from the literature through the automatic methods of acquisition and appraisal. In this study, we proposed a comprehensive framework to combine the abovementioned processes, particularly, the evidence acquisition and evidence appraisal to facilitate the clinical decision making. The proposed methodology uses the information contents from a health system and the KB of CDSS for the query construction to search and retrieve relevant research papers from the literature to support the evidence-based practice (EBP). The EBP and the CDSS have long been used in the clinical domain to enhance clinical efficacy. The EBP and the CDSS share clinical expertise as a source of data. The EBP uses the clinical expertise along with the research evidence and other factors for a clinical decision. A CDSS KB is the representation of clinical expertise of 1 or more clinical experts. The EBP is defined as "the

XSL•FO
RenderX

conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" integrated with clinical expertise and patient values to optimize outcomes and quality of life [8].

Figure 1. Clinical decision support system knowledge base linkage with evidence base synthesized from biomedical literature.



Role of User Context in Query Construction

For any evidence-based system to work efficiently, the context of the domain plays a crucial role. The context provides the features for query generation to seek the relevant information from external sources. The source and format of the data are crucial to consider automatic or semiautomatic query generation. Cimino [9] presented the idea of Infobuttons and Infobutton Manager, which attempt to determine the information needs based on the user context. Infobuttons are mainly topic-specific questions with a facility for the users to tune the query more toward the context. CDAPubMed [10] is a browser extension that aims to provide a tool to semiautomatically build complex queries. It provides additional information to the contents of the EHR to improve the biomedical literature searches. A platform called ProvCaRe [11] has the provision for search and query operations on provenance metadata to enable reproducibility of research articles. There are other approaches described in the studies by Bakal et al [12] and Sahoo et al [13] that use semantic patterns over biomedical knowledge graphs for treatment and causality predictions and semantic provenance to apprehend high-quality domain-specific information using expressive domain ontologies.

Related Work on Finding High-Quality Articles in the Literature

A decent set of approaches is available that had improvised the results of literature searching with respect to quality of studies. The PubMed Clinical Queries (CQ) [14] is one of the most prominent endeavors to retrieve scientifically sound studies from the biomedical literature. Afterward, supervised ML approaches were introduced mainly to improve the precision of the results in terms of quality checking for methodological rigorousness. Similarly, to find high-quality papers in MEDLINE, Wilczynski et al [15] developed CQ filters, which were later adapted by PubMed for use as CQ. The data collection used in the CQ filters is annotated across the following 4 dimensions: the format, the human health care, the purpose, and the scientific rigor. The experimental studies [16,17] introduced ML (supervised learning) classification models to differentiate between the methodologically rigorous and the nonrigorous articles. In an article about evidence quality prediction [18], the authors addressed the problem of automatic grading of evidence on a chosen discrete scale. The authors experimented many features, such as publication year, avenue, and type to evaluate the quality of the evidence. They found that the publication type is the most eminent feature to consider for evaluation of the evidence quality results. A DL neural network known as the Convolutional Neural Network approach [19] was very recently tried to further improve accuracy over the existing approaches of PubMed CQ and McMaster's text word search in terms of precision.

Limitation of the Existing Approaches

The existing approaches discussed mainly focus on the automation of evidence processing to overcome the central problem of time spent on searching while practicing EBM. The inclusion of the research evidence in clinical decisions varies with respect to domain context and objective. Conceptually, the evidence adaption follows the same 5As cycle as mentioned in the study by Leung [8]; however, implementation makes the scenario different. A user in a clinical setup with a CDSS implementation needs to approach the evidence differently than a user who does not have a CDSS implementation. The dataset selection, the feature engineering, and the context awareness bring uniqueness to the approach and pose challenges at the same time. The objective of this study was to circumvent the issues of efficient searching in the biomedical literature to find evidentiary articles that are qualitative and fit-to-context in the user scenario.

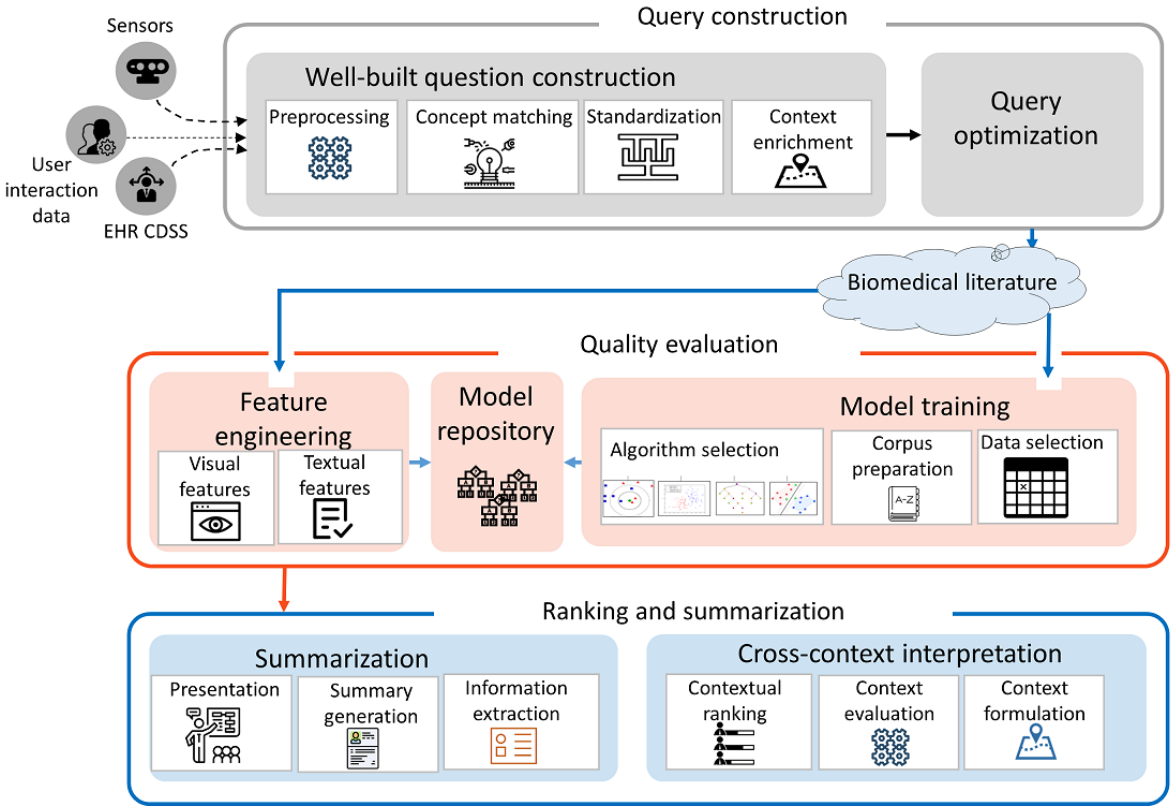
XSL•FO
RenderX

Methods

Overview

To achieve the research goal of curating and mining data from scientific publications in biomedical literature, we designed a coherent and comprehensive architecture of the framework, which is depicted in Figure 2. The architecture is divided into 3 layers to accommodate the necessary functions of connecting a health system with the scientific research. In the first layer, an optimized query is constructed in a well-built form from the input data streams. In the second layer, the quality is evaluated with data-driven approaches that include a ML or DL algorithm, which is meaningfully selected for the input set of parameters and the data requirements. Finally, in the third layer, the scientific research articles that have been evaluated for quality are summarized and ranked according to the user context to bring an article to the top, such that it is not only relevant and qualitative but also contextually viable and applicable.

Figure 2. The conceptual diagram of the proposed biomedical literature curation framework. CDSS: clinical decision support system; EHR: electronic health record.



Query Construction

The query construction is a widely studied and multiaspect topic. One aspect concerns the type of query, which could be manual, semiautomatic, or automatic. Other important aspects include the input data, the context, and the environment of the user. Finally, the query format and the structure could be either just random or well built. Here, we provide a summary of different query construction strategies and recommendations for an efficient strategy from the input clinical information. As shown in Figure 3, there are multiple paths to construct a final query. As examples, we discuss a few popular strategies in Table 1 that were and are in practice or envisioned in this study as a potential futuristic strategy.

XSL•FO
RenderX

Figure 3. Query construction strategies. EHR: electronic health record.

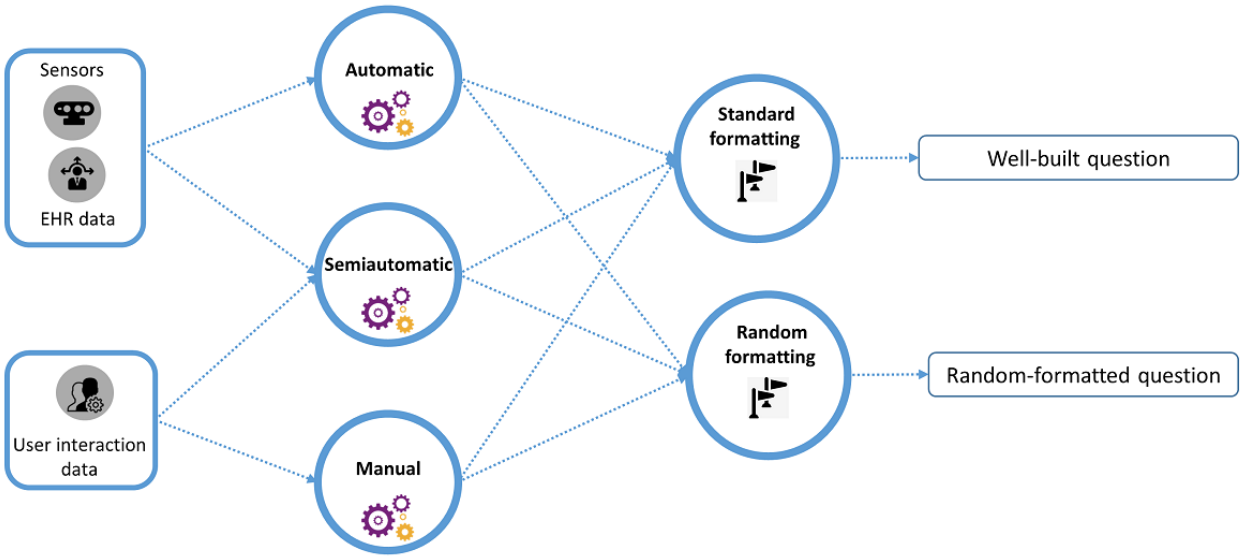


Table 1. Descriptions of different query construction strategies.

Strategy	Description	Positives	Negatives
The automatic well-built question	In this strategy, a final question is constructed without human intervention. The raw data are acquired meaningfully from the patient information stored in the patient record, and they are associated with the rules of the clinical decision support system or curated from the sensor devices. The acquired elements of the data are transformed autonomously to a well-built or standard format.	An efficient method that does not involve a human to write the query terms; it is easily understandable by the user because of the well-built format; it is comparatively straightforward to summarize the retrieved contents.	Achieving accuracy needs extra effort at the beginning; the design of the intelligent methods is required to correctly place the terms in the appropriate place of a standard or well-built structure.
The automatic random-formatted question	The input part is the same as in the first strategy, and the ingredients of the query are automatically acquired from the input sources. However, they are placed randomly without arranging in a specific format.	It is an efficient method because a human is not required to write the query terms; no effort is required to place the query terms in the required slots.	It is less understandable by the user because of the randomly placed terms; the interpretation and the summarization of the retrieved results will be a daunting task.
The semiautomatic well-built question	The acquisition of the query terms from the input source may be semiautomatic, and human involvement will be necessary to complete the missing section. In addition, placing the terms in the required slots of a standard structure will need human assistance.	Trustworthiness is higher than automatic because of the user's involvement; an edge in ranking and summarization of the retrieved results.	It is expensive in terms of time because a user will still be required to complete the query contents and the structure.
The semiautomatic random-formatted question	The input acquisition is partially automated. The arrangements of the query terms are random.	The trustworthiness is high.	The interpretation of the query terms and the summarization of results will be a problem.
The manual well-built question	In this strategy, a human is involved thoroughly to write all the contents of a query in a specified structure.	The trustworthiness and a better interpretation of the query terms, the ease in ranking, and the summarization.	It is time consuming; it is hard for naïve users to write complex queries.
The manual random-formatted question	All the contents of the query are written by humans without arranging them in a specific format.	The trustworthiness is high because all the terms are written by the humans.	It is time consuming; it is hard for naïve users; the interpretations, the ranking, and the summarization issues.

In this study, our main focus is the first strategy, which involves constructing an automatic well-built question. We have chosen to formulate the query in Patient/problem, Intervention, Comparison, and Outcome (PICO) format [20] from the input data, which included the patient structured information and the knowledge rules. PICO has a well-structured format, which differentiates different parts of a clinical question in more applicable parts that are easily understandable for the clinicians and other users. It also helps to determine the context of the question. In addition, the structure is helpful to summarize and rank the retrieved articles.

XSL•FO
RenderX

Input Acquisition and Preprocessing

The contents are preprocessed based on the nature of the input. This section summarizes the guidelines for the developers and the implementers of the representative scenarios. These scenarios represent automatic queries constructed from the following 3 types of inputs: patient record, CDSS rules, and sensory data.

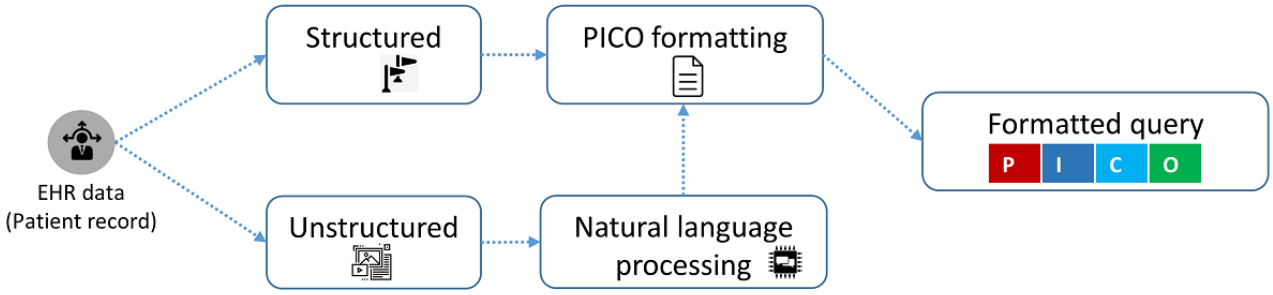
Scenario 1: An Automatic Query Construction From a Patient Record of the Electronic Health Record Data

There are 2 possibilities, which involve the data being in a structured or an unstructured format. If the data are structured with assigned labels, they are placed in the target P, I, C, and O sections accordingly. However, if the data are in an unstructured format, then an additional step of the natural language processing (NLP) is required to extract the meaningful terms from the unlabeled contents, to recognize their type and context, and then finally to place the processed terms in the target PICO format. The abstract flow of PICO construction from EHR data is depicted in Figure 4.

The following natural language preprocessing steps are applied in a pipeline: (1) the text is broken into tokens with a space delimiter; (2) stopwords of English language are removed; (3) case of the letters is changed to lower; and (4) the words are stemmed to their root words using porter stemming.

The stemmed words are mapped to the PICO format using salient term identification (STI) algorithm explained in the following section. After finding out a concept in the standard vocabulary of Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) implemented through the Unified Medical Language System (UMLS) vocabulary service application programming interface (API), the algorithm further finds out the semantic and entity types of the identified concepts. Finally, the identified concepts are mapped to PICO-corresponding slots. For example, because female belongs to population group, it is mapped to the P slot of PICO.

Figure 4. The flow of Patient/problem, Intervention, Comparison, and Outcome query construction from the patient record of the electronic health record data. EHR: electronic health record; PICO: Patient/problem, Intervention, Comparison, and Outcome.



Scenario 2: An Automatic Query Construction From the Clinical Decision Support System Knowledge Base

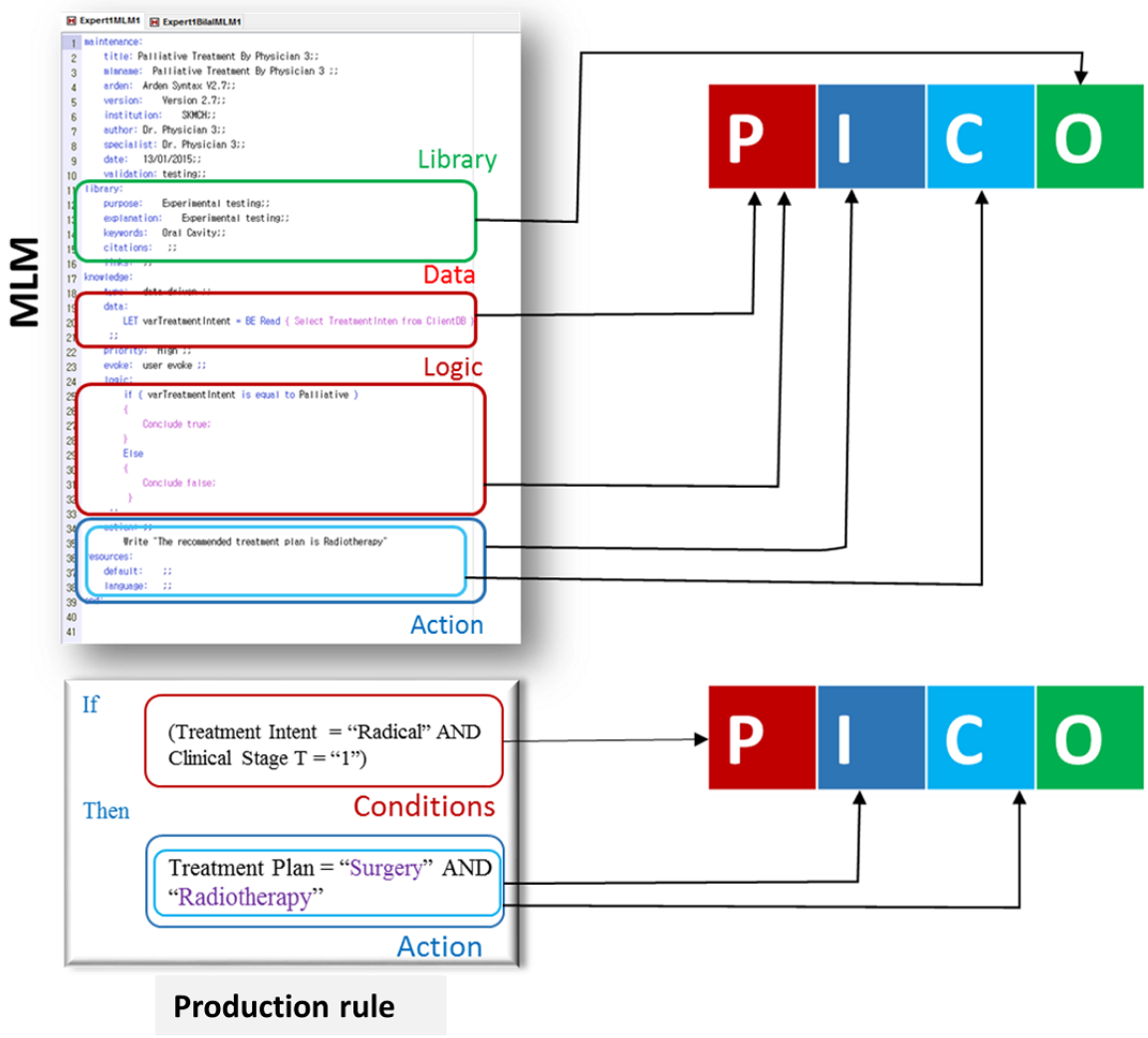
In this scenario, the contents of the query are extracted from the rules of a CDS KB. The rules are actuated against a particular decision. The extraction process relies on the representation scheme of knowledge. Of these schemes, we elaborated on the following 2 schemes for the mapping of rules to PICO: plain production rules (if-then) and the HL7 medical logic modules (MLM) [21]. We designed a general model that could be used for any of the representation schemes. According to the proposed mapping model, different parts of a rule are mapped to PICO as described in equation 1:

$$PICO=D \cap A \cap P \quad (1)$$

In equation 1, D represents the set of elements in the data part of a rule, A is the set of elements in the action part, and P shows the purpose of a rule. More specifically, D maps to P of PICO, A maps to both I and C, and P maps to O of PICO. For clarity, Figure 5 is provided to describe the mapping from rules to PICO using MLM and the plain production rules as an example. In the scenarios where there is lack of information to get outcome information from the input, some elements of PICO can be unmapped. For example, in the scenario of production rules, the O part of PICO remains unmapped.

XSL•FO
RenderX

Figure 5. An example of Patient/problem, Intervention, Comparison, and Outcome mapping from the HL7 medical logic module and the production rule. MLM: medical logic module; PICO: Patient/problem, Intervention, Comparison, and Outcome.

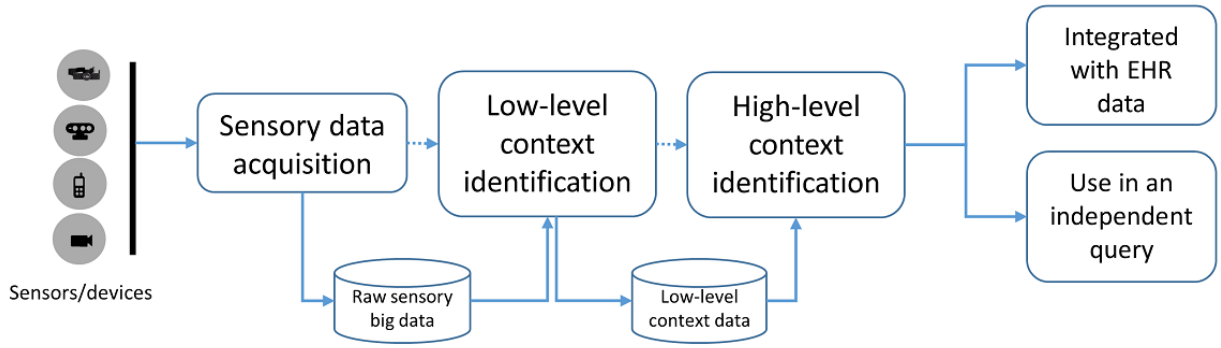


Scenario 3: Constructing a Query From the Multimodal Sensory Data

This scenario is more applicable to participatory health management, where user activity, diet, sleep, and other related information are acquired through different sensors and devices. The information from these sensors and devices are collected independently through their independent clocks with an associated time stamp. A logical clock is required to identify the data origination at the same time [22]. After synchronization, the raw data need to be labeled and persisted for other services to consume. Using the labeled data may require further processing to determine the high-level context for the appropriate usage in the query. For instance, if a user is doing a set of activities, such as walking, running, or lying down, in an adjacent frame of time, it may refer to a high-level context of exercise. In one of our preliminary work on the project of Mining Minds [23], we have developed different models for context recognition both at the lower and the higher levels on the basis of data curated from different sensors. The dataflow of the raw sensory data is briefly illustrated in Figure 6. It must be noted that the contextual information determined from the sensory data could either be used in an independent query or used as a subset information of a query constructed from the EHR data as noted in previous subsections. There could be scenarios to combine methods of the abovementioned 3 scenarios and construct a single query depending on the availability of the data and the user needs.

XSL•FO
RenderX

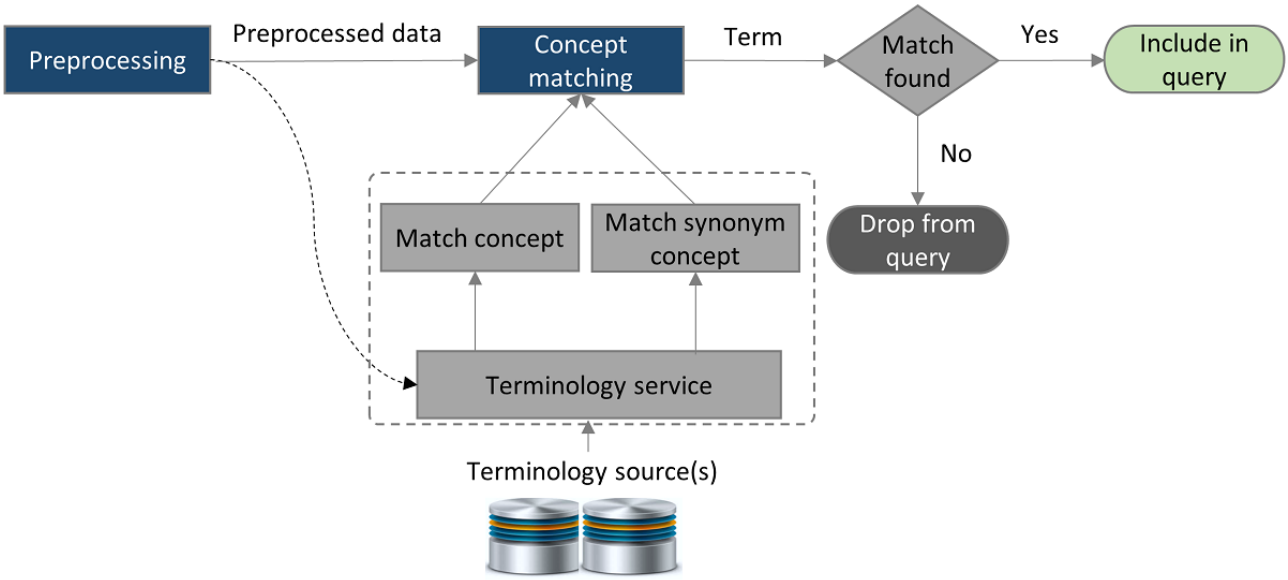
Figure 6. A dataflow diagram of the raw sensory data acquisition and the context identification. EHR: electronic health record.



Concept Matching for Term Inclusion and Exclusion

At the time of term extraction from the EHR data, it is important to include only pertinent and important terms. We developed the STI algorithm to filter out the less effective terms from the user question. The STI is a weight-based algorithm that finds an input term in a terminology source (SNOMED-CT/UMLS) and provides weight according to the matching level, such as exact match, partial match, and synonym match. The steps of the STI algorithm are described in Figure 7. According to this algorithm, if a term finds an exact match, it gets more weight (w=1.0) compared with the partial match and synonym match (w=0.5). The algorithm is formally represented in algorithm 1 in Multimedia Appendix 1.

Figure 7. The salient term identification algorithm.

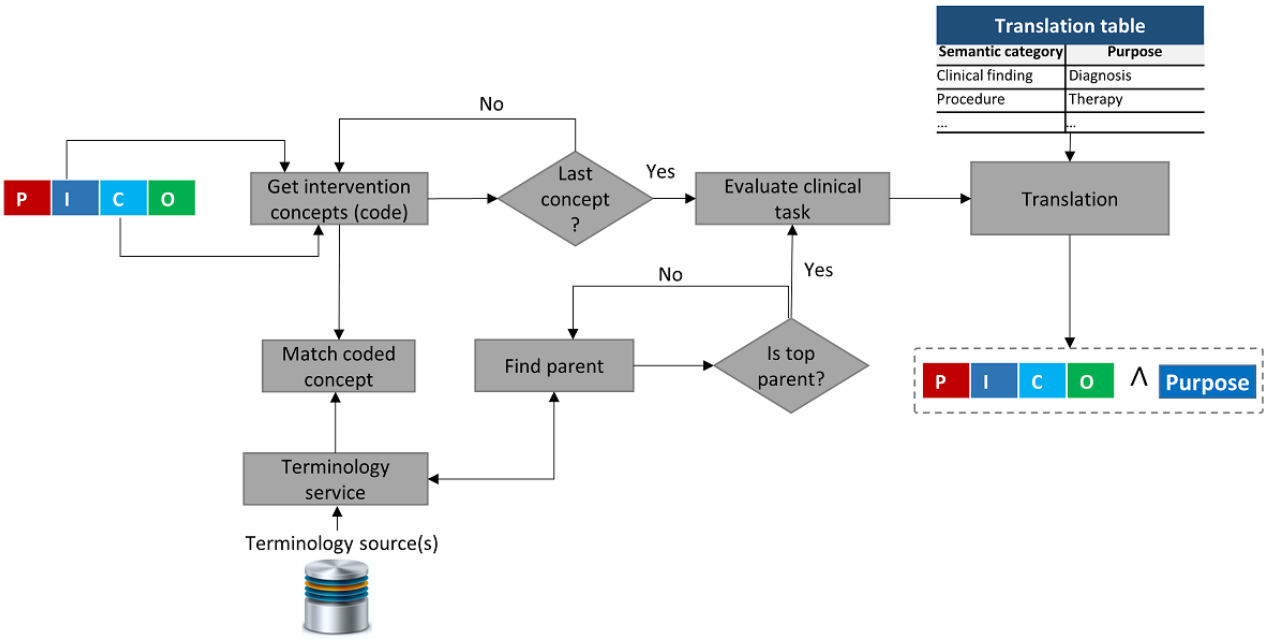


Standardization and Context Enrichment

It is not mandatory to use standard form of the terms; however, it is important to infer the overall intentions of the user from the query. In other words, the standardization helps to determine the users' interest in obtaining results that are related to the diagnosis, the treatment, or the prognosis. In addition, it helps in understanding the meaning of a particular term precisely and helps avoid confusion. To achieve standardization of the terms and determine the purpose of the query, a terminology source, preferably the UMLS, could be used. We must remember that the function of standardization is achieved at the time of executing the STI algorithm. The purpose of the query, also known as query type, is determined from identifying the semantic types and entity types of the concepts used in the I and/or C parts of the PICO. As described in Figure 8, the concepts extracted from the PICO I and/or the C parts are used to determine their parent concepts with the help of the terminology services. After getting translated from the translation table for the parental term, the inferred translated concept (diagnosis/treatment/prognosis) is attached to the PICO query. During implementation of the query to run on the PubMed, the purpose term or the query type is used as a clinical filter for an increased recall [15].

XSL•FO
RenderX

Figure 8. The steps of query type identification algorithm. PICO: Patient/problem, Intervention, Comparison, and Outcome.

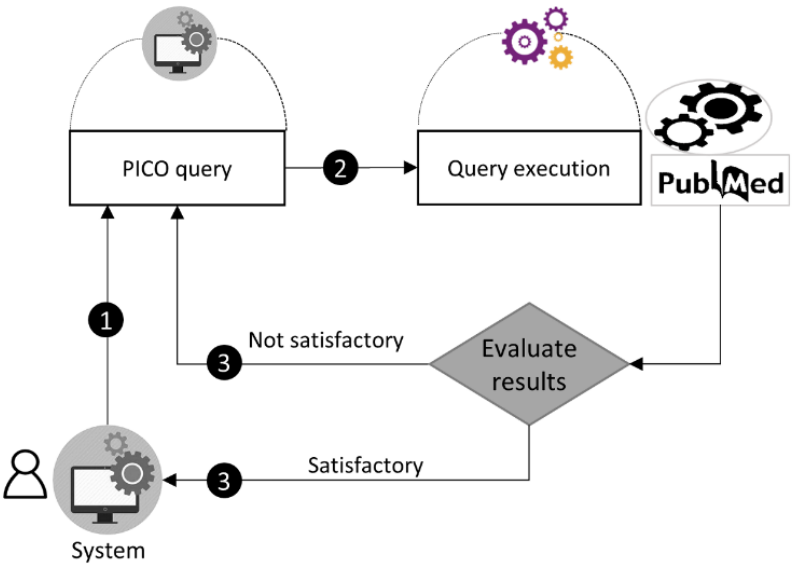


Query Optimization and Searching

A well-built structured query may be unnecessarily too lengthy to return absolutely no results or too short to return too many results. In this case, a query optimization technique is necessary to balance the resultant set. If there is a query that occasionally fails to retrieve any results, the query is optimized to exclude the least important term from a list of terms. As a general guideline, the P and I set of terms in PICO is the core, and they are considered comparatively more important than C and O. On the basis of this theory, we designed an optimization strategy, which is illustrated in Figure 9, to exclude a term from the least important parts one by one unless we retrieve an acceptable set of publications. For instance, if there are 4 terms ANDed in a query belonging to one of each PICO element and by running that query returns 0 records, we remove the C term first and check if the resultant set satisfies a threshold; if that is the case, we execute the query and continue the process further; otherwise, we remove the O term in the next cycle.

Even if the query terms consist of only P and I of PICO but are too many to retrieve any results, we can use the weights determined by the STI algorithm and remove the least effective term(s). The final optimized query was executed on biomedical literature and used the retrieved articles for further processing. For this research, we used the PubMed service to access the biomedical literature.

Figure 9. The query optimization process flow. PICO: Patient/problem, Intervention, Comparison, and Outcome.



XSL•FO
RenderX

Quality Evaluation

Articles retrieved using the PubMed built-in search strategy could possibly include quality, nonquality, or less quality articles that need to be segregated before presentation to the user. For this very reason, we designed quality parameters that need to be checked so that only quality contents come forward to be read by the users.

Guidelines for Using the Gold Standard

We learned from previous studies that the Clinical Hedges database [24], which was developed by the Hedges Group at McMaster University, can be used as a gold standard dataset. Clinical Hedges was initially employed to develop and evaluate the CQ filters [15]. It is also used for ML approaches that identify the scientifically sound PubMed clinical studies [17]. The database consists of 50,594 MEDLINE articles published in 170 clinical journals, of which 49,028 articles are unique. All the articles are manually annotated by a team of specialized experts, and they classified the articles across the following 4 dimensions: format (O=original study, R=review, GM=general and miscellaneous articles, and CR=case report), human health care interest (yes/no), scientific rigor (yes/no), and purpose (diagnosis, etiology, prognosis, treatment, economic studies, reviews, and clinical predication guides). The primary purpose of creating the database was to evaluate whether each study was scientifically sound or not using the criteria for the treatment interventions, which include clinically relevant outcomes, random allocation of study participants, and at least 80% of the follow-up of study participants.

Selection of the Search Strategy

One of the issues for common users and researchers is to choose an appropriate search strategy to satisfy their information needs. None of the state-of-the art search strategies could be considered ideal in all situations. We provide a common standpoint and recommendations to opt for a strategy based on the users' needs and rationale. We divide the set of approaches in 3 groups: (1) PubMed search strategies that include mainly CQ, (2) the ML approach, and (3) the DL approach. We also categorize the user information needs based on recall, precision, and recentness, which is the instant availability of a study. Table 2 provides the performance evidence of the existing state-of-the-art approaches, whereas Table 3 provides the recommendation of using an approach on the basis of given rationale.

Regarding the third criteria, such as the recentness, we elaborated the abovementioned approaches on a delay factor. As both the PubMed CQ and the mentioned ML approaches use a Medical Subject Headings (MeSH) filter in the search strategies, they encounter problems classifying the most recent articles. The mean delay in the MeSH indexing per journal was recorded as 162 days. The DL neural network considers only the title and the abstract features and has no dependency on the MeSH terms and does not encounter any delays to evaluate even the most recent studies.

Table 2. Average recall and precision of different search strategies.

Approach	PubMed Clinical Queries (broad)	Machine learning	Deep learning	Reference
Average recall	98.4	91.4	96.9	Reported in the study by Perez-Rey et al [10] and Del Fiol et al [19]
Average precision	22.4	86.5	34.6	Reported in the study by Perez-Rey et al [10] and Del Fiol et al [19]

Table 3. Recommendation of the search strategy in a given rationale.

Rationale	Recommendation
User top priority is high coverage (recall)	PubMed Clinical Queries
User top priority is to get precise results (precision)	Machine learning classification approach
User top priority is more recent studies	Deep learning neural network approach

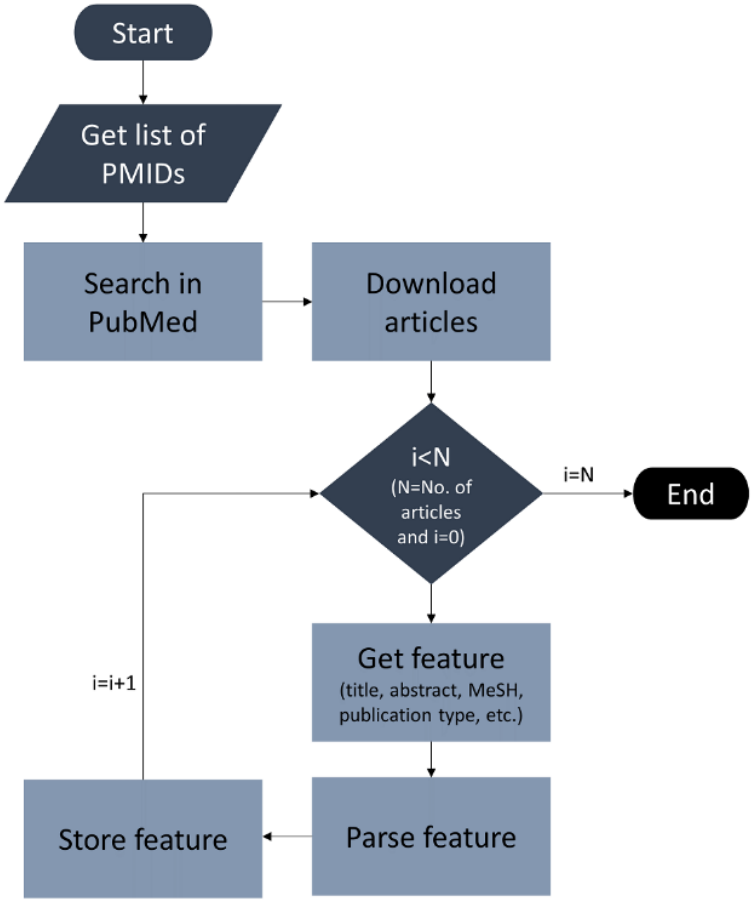
Feature Engineering and Corpus Preparation

Irrespective of the strategy, we need to engineer the features and prepare a corpus to run ML or DL classifiers. We used the Clinical Hedges dataset and acquired the PubMed identifiers, which we posted to create a custom database on PubMed through the Entrez Post service method of the Entrez Programming Utilities API, and we searched the publications using the eSearch service for the Entrez Fetch (eFetch) service by enabling the history and the environment variables to yes. The eFetch function was used to download the searched articles. The downloaded records were programmatically parsed to obtain the data for the data features, such as the title, the abstract, and the metadata features, which include the MeSH terms and the article type. The process of downloading and parsing the articles is described in Figure 10.

We engineered 2 sets of features, which included the data features and the metadata features. The data feature vector was created by tokenizing the titles and abstracts, changing the case to lower, eliminating the stop words, stemming the words using the porter stemmer, and filtering the tokens by lengths. Unlike the data features, the metadata features were created by applying only tokenization and case transformation because there was no need to remove stop words and stemming.

XSL•FO
RenderX

Figure 10. The article downloading and parsing algorithm. MeSH: Medical Subject Headings; PMIDs: PubMed Identifiers.



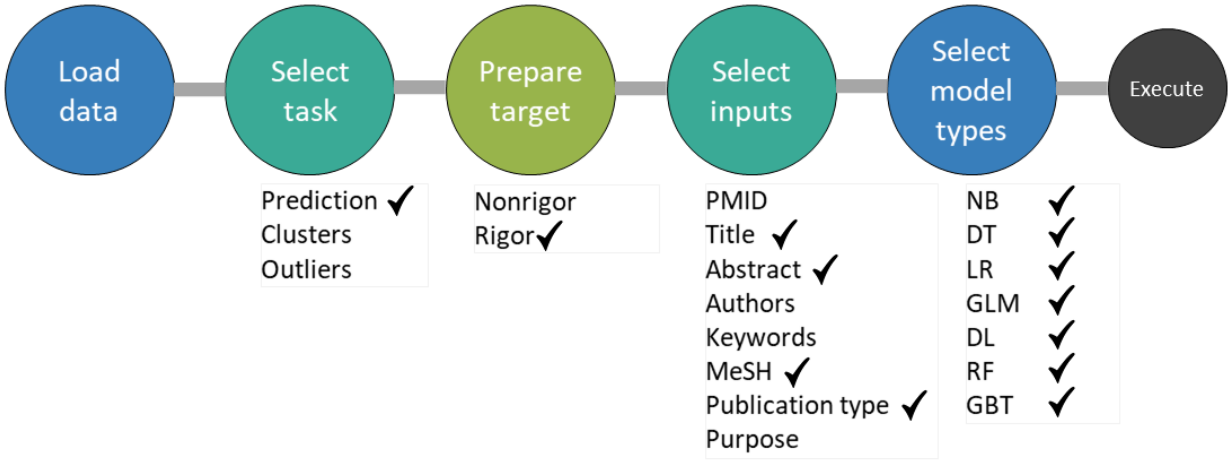
Extended Quality Recognition Model

Previously, we worked to identify the relevant and quality articles using the ML approach and developed a model called quality recognition model (QRM) [25]. The QRM was a binary classification model with the following 2 classes: rigor and nonrigor, where rigor and nonrigor represent the quality and nonquality articles, respectively. It was tested with 4 different ML algorithms, including Naïve Bayes (NB), k-nearest neighbor (kNN), decision tree (DT), and support vector machine (SVM). For this study, we extended the QRM with inclusion of DL model and ensemble technique to get a better performance. With the advent of automodel feature in the data science tool RapidMiner Studio 9.0.003 [26], it is rather more expedient to opt for an efficient model in a range of applicable models. Using the same dataset that we previously used for building QRM, the RapidMiner automodel function proposed 7 algorithms, including NB, DT, logistic regression, generalized linear model (GLM), random forest, gradient boosted trees, and DL. We keep consistency in the feature set selection, similar to the previous model of QRM. The automodel function allows us to intervene in the settings of parameters at different steps including the load of dataset, selection of the task, preparing a target, selection of the input features, and model types to execute for getting the final results. We described these steps in Figure 11 by highlighting our selection among alternatives.

We use the automodel function to select the individual learners to build our ensemble model to acquire high accuracy as compared with the performance of individual learners. We first experimented with the individual models proposed by automodel function. Later, we develop an ensemble learner over the individual learners. In the first layer of assembling, we use AdaBoost learner, whereas on the second layer, we use ensemble voting (stacking model) with a sampling type of automatic, which uses stratified sampling per default. However, if the example set does not contain a nominal label, shuffled sampling is used instead. The split was relative with a split ratio of 0.7.

XSL•FO
RenderX

Figure 11. Automodel steps and parameter settings. DL: deep learning; DT: decision tree; GBT: gradient boosted trees; GLM: generalized linear model; LR: logistic regression; MeSH: Medical Subject Headings; NB: Naïve Bayes; PMIDs: PubMed Identifiers; RF: random forest.



Ranking and Summarization

A potential problem that clinical user face is the management of the results set to identify, appraise, and synthesize the best available evidence to answer the clinical question in the best possible manner. In all this process, a lot of manual effort is required to extract the data to make a summary, and it is also subjected to error [27]. Moreover, the context of the user may change the default ranking of an article to bring it to the top or take it to the bottom. Some of the existing approaches use a grading mechanism of ranking the articles based on the strength of the contents [18,28]. The need to develop a ranking mechanism that is more patient-centered rather than only evidence-centered is needed. In addition, the model needs to consider the user's context in addition to the articles' strength and quality.

To address these issues, we conferred our previous work [25] and made possible extensions to provide guidelines for using an appropriate model in a given clinical situation. We devised a cross-context evaluation strategy that involves crossly matching 2 contexts, such as user contexts and evidence contexts. User contexts have multiple elements, such as basic information, which shows the user educational level. The background is the experience of the user, and the goal shows the short-term learning or the long-term learning. The interest represents the preferences, and the learning style is the pattern of user learning, such as textual and visual. On the other hand, an evidence context includes the article meta-features or properties, such as the publication type, the publication avenue (eg, journal and book), and the year of publication.

We devise a cross-context evaluation method that accumulates contextual parameters of both user and evidence contexts. User context parameters are represented as C1, C2, ..., Cn, whereas the evidence contextual parameters, which are properties of a publication, are represented as P1, P2, ..., Pn. These 2 sets of contexts are aggregated first vertically and then horizontally to reach to the final grading of a publication as H=high, M=medium, L=low, and U=unknown. In other words, the algorithm first finds the aggregate value of each column, such as the highest value of all the cells using the majority vote procedure, and the process is repeated for all the cells. The example described in Table 4 for a user with three contexts and a publication with two properties, the final context value is calculated for a given publication-x by learning the highest value from the aggregate contexts AggCtx-1 and AggCtx-2. The value H, M, L, and U are learned from the ranking values assigned to each of the article types described in Table 5.

Table 4. The context aggregation for article ranking.

User context	Publication-x context	
	P1	P2
C1	H[a]	M[b]
C2	L[c]	M
C3	H	H
Aggregate context	AggCtx-1=H	AggCtx-1=M

[a]H: high.

[b]M: medium.

[c]L: low.

XSL•FO
RenderX

Table 5. Values of publication types ranking and grading.

Publication types	Ranking	Grade value
Systematic reviews	1	H[a]
Meta-analysis of RCTs[b]	1	H
RCTs	3	H
Meta-analysis of CTs[c]	4	M[d]
Systematic review of CTs	5	M
CT	6	M
Cohort study/case-control study/report	7	M
Guidelines	8	L[e]
Opinion	9	L
Observational study	10	L
Any other publication type	11	L

[a]H: high.

[b]RCTs: randomized controlled trials.

[c]CT: control trials.

[d]M: medium.

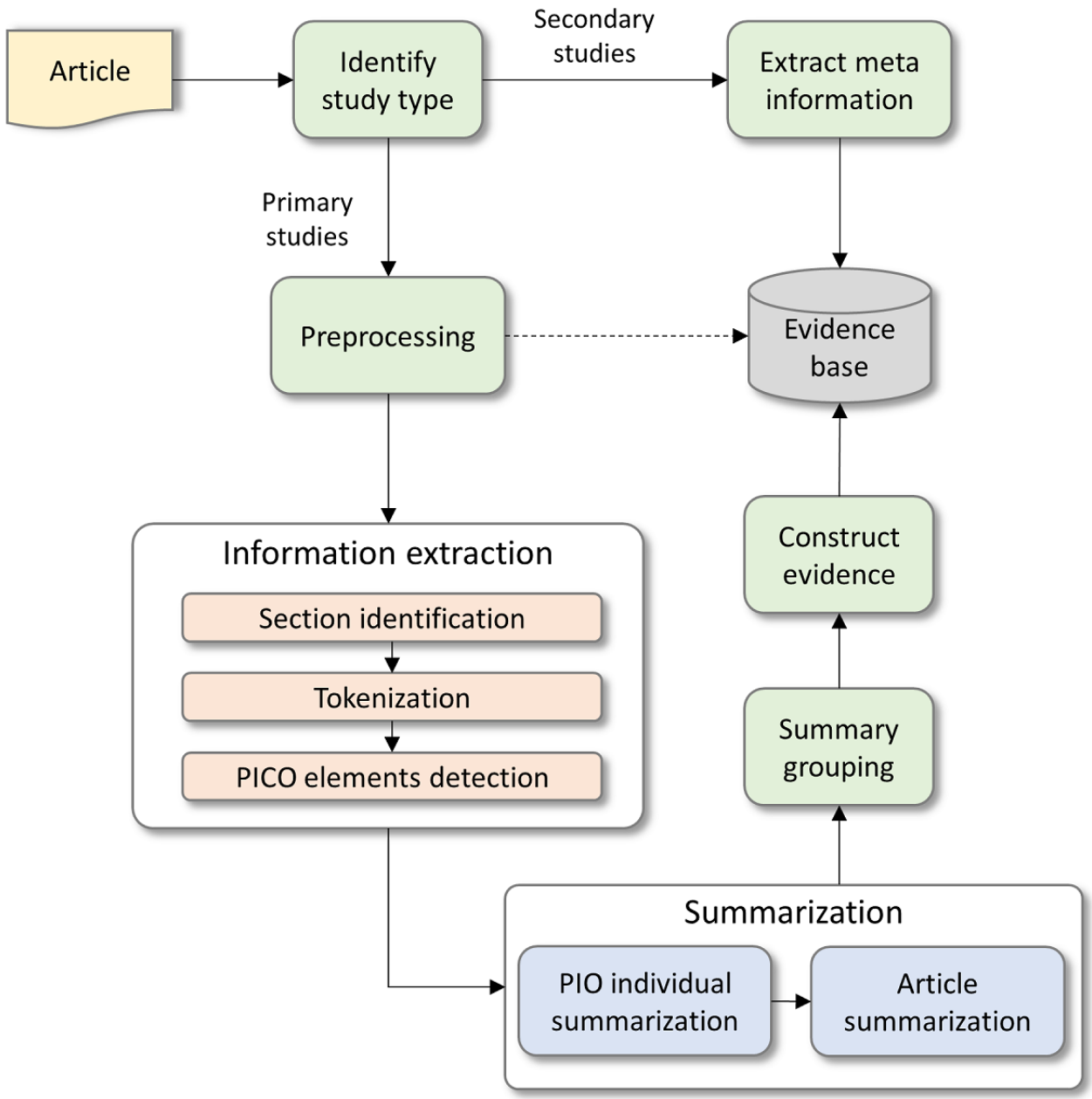
[e]L: low.

Finally, the evidence is formed from the set of the ranked articles, and it is presented in the form of a summary. As mentioned earlier, the manual summarization is a daunting task, and researchers have devised different methodologies to perform the automatic summarization of the articles. Bui et al [27] developed a computer-based ML and an NLP approach to automatically generate a summary of full-text publications by extracting the PICO values alongside the sample size and group size from the text. There are few other studies that proposed algorithms to detect PICO elements in the primary studies [29,30], which in term assisted the process of summarization.

We are convinced that the PICO-based approach assists the extractive method of automatically generating the summaries of articles. As a result, we followed the extraction of the PICO values from the text of an article. Before the extraction, we identify if the type of article is primary or secondary. The secondary articles such as systematic reviews (SRs) are formed from multiple primary studies, and they could be considered as evidence. However, the primary studies required summarization to form an evidence. We provide an abstract view of the proposed summarization system, as shown in Figure 12, to categorize articles as secondary and primary studies and construct summaries for the latter. The system works as each ranked article is taken as input and is distinguished as a primary study or secondary study. A secondary study is only processed to extract the meta information and is stored to the evidence base as an evidence. A primary study is preprocessed to convert the format from pdf to text in the first place. The text is passed to the information extraction module to identify sections and PICO elements in different section to generate individual summaries for P, I, and O elements of PICO. The element C is also an intervention, so a separate summary is not necessary; therefore, we consider it as a part of I during summarization. On the basis of individual summaries of Patient/problem, Intervention, and Outcome elements, a complete summary is constructed for the whole article. Individual articles' summaries are grouped based on their similarities to form a single evidence, which is then stored to the evidence base by including the meta information associated with each article.

XSL•FO
RenderX

Figure 12. The framework for article summarization. PICO: Patient/problem, Intervention, Comparison, and Outcome; PIO: Patient/problem, Intervention, and Outcome.



Results

Overview

As mentioned earlier, the focus of this study is to construct a PICO-compliant query from EHR data where there were 2 evident options to acquire query terms: structured data and unstructured data. One of the possible options that could be used for structured data is the knowledge rules of a CDSS, and any clinical scenario explained in simple English can be considered for the unstructured data case study. In knowledge rules, it is quite straightforward to identify the clinical concepts and map them to the PICO elements. However, unstructured data mapping to PICO requires quite a few steps to get the final PICO-based query, as discussed in the query construction section.

Case Study—Query Construction From a Clinical Scenario

Here, we present a clinical scenario and step-by-step outcomes of our proposed algorithms, mapping to PICO and finally a search query construction. The first step is to find important clinical concepts in the given clinical scenario followed by finding their semantic type and entity type. In the scenario shown in Figure 13, our proposed STI algorithm found out 8 concepts in the standard vocabulary of SNOMED-CT implemented through the UMLS vocabulary service API. For the identified concepts, we identified the semantic and entity types of the identified concepts. In the second step, the identified concepts are mapped to PICO-corresponding slots, such as female belongs to population group, so it is mapped to the P slot of PICO. Similarly, beta-blocker has the entity type of chemical and drugs, so it went to the I (intervention) slot of

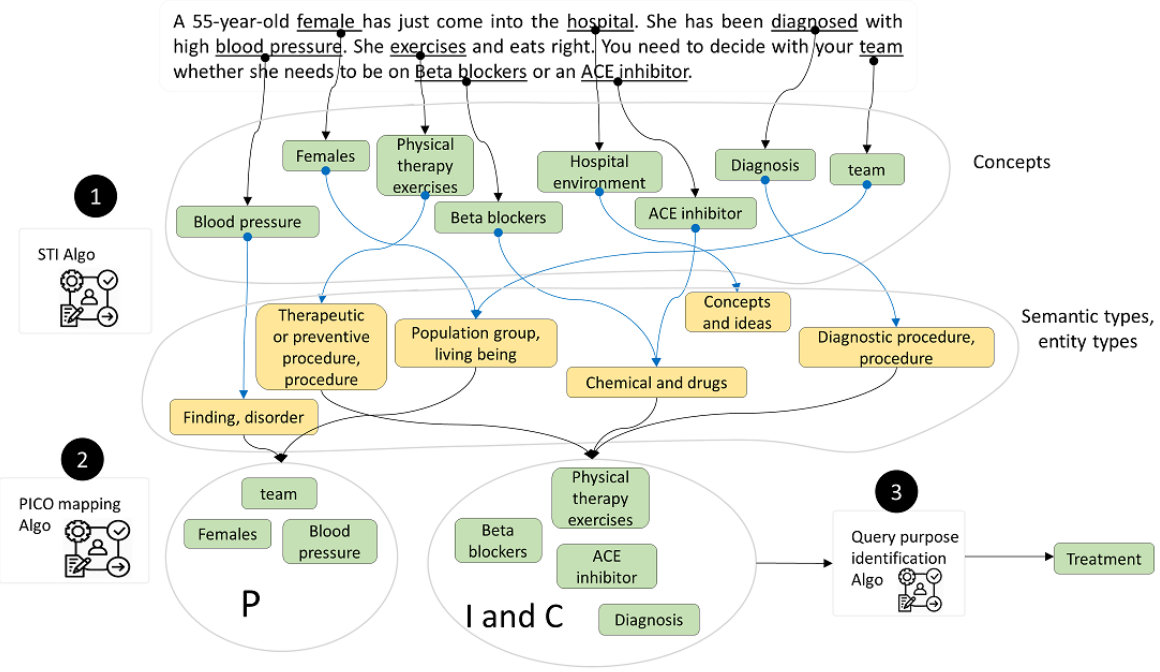
XSL•FO
RenderX

PICO. In the same way, we mapped the rest of the concepts. Finally, in the third step, we find the type of the query among diagnosis, treatment, prognosis, or etiology. Looking at the semantic/entity types of majority of the concepts in I and C, the algorithm concluded with the query type as treatment. Using the logic of searching query construction as described earlier, we will get the query as follows:

q=female blood pressure and (beta-blocker or ACE inhibitor) (2)

This query *q* is passed to our searching algorithm to search for the publications using the PubMed CQs utility.

Figure 13. Results of a clinical scenario conversion to Patient/problem, Intervention, Comparison, and Outcome with query type (purpose). ACE: angiotensin-converting-enzyme; PICO: Patient/problem, Intervention, Comparison, and Outcome; STI: salient term identification.



Quality Evaluation

The existing QRM was tested using multiple ML approaches, which included the SVM, the DT, the kNN, and the NB. The results were reported in our previous study [25], where the SVM algorithm has performed better than other algorithms. The automodel employed algorithms produced varied results for different algorithms. As shown in Table 6, the gradient boosted trees (GBT) algorithm outperformed other algorithms with accuracy of about 84.98%, followed by GLM with accuracy of about 83.80% at the individual learning stage. To minimize the instances of wrong classification, we tested ensemble method using AdaBoost on top 3 individual learners. We noticed that GBT was still on the top, with an increase of about 4% accuracy jumping from 84.98 to 88.50. The performance of GLM was slightly increased by about 1%, whereas the DL performance with AdaBoost was increased by about 9% accuracy from 75.48 to 84.57. In the final model, we use a second level of ensemble over AdaBoost (GBT) and GLM and obtained about 3% improved accuracy of 90.97%. Moreover, area under the curve (AUC) value of the E-QRM was noted as 0.989, whereas for AdaBoost (GBT), it was 0.950, followed by AdaBoost (DL) with AUC value of 0.921.

It is important to note that the experiments are performed using RapidMiner Studio version 9.3.001, which is an improved version of the descendants. All the models, particularly the automodel, may generate a different set of results even on the same data because of the changes in operators for the possible improvement from version to version.

XSL•FO
RenderX

Table 6. Extended quality recognition model performance overview using different algorithms.

Algorithm/criteria	F measure	Precision	Accuracy	Area under the curve
Naïve Bayes	0.53	0.42	0.64	0.53
GLM[a]	0.72	0.82	0.84	0.89
Logistic regression	0.46	0.39	0.62	0.63
Decision tree	0.09	0.50	0.70	0.50
DL[b]	0.57	0.59	0.75	0.78
GBT[c]	0.75	0.77	0.85	0.87
AdaBoost (GLM)	0.73	0.78	0.85	0.82
AdaBoost (DL)	0.72	0.79	0.85	0.92
AdaBoost (GBT)	0.79	0.85	0.89	0.95
Extended quality recognition model	0.85	0.83	0.91	0.98

[a]GLM: generalized linear model.

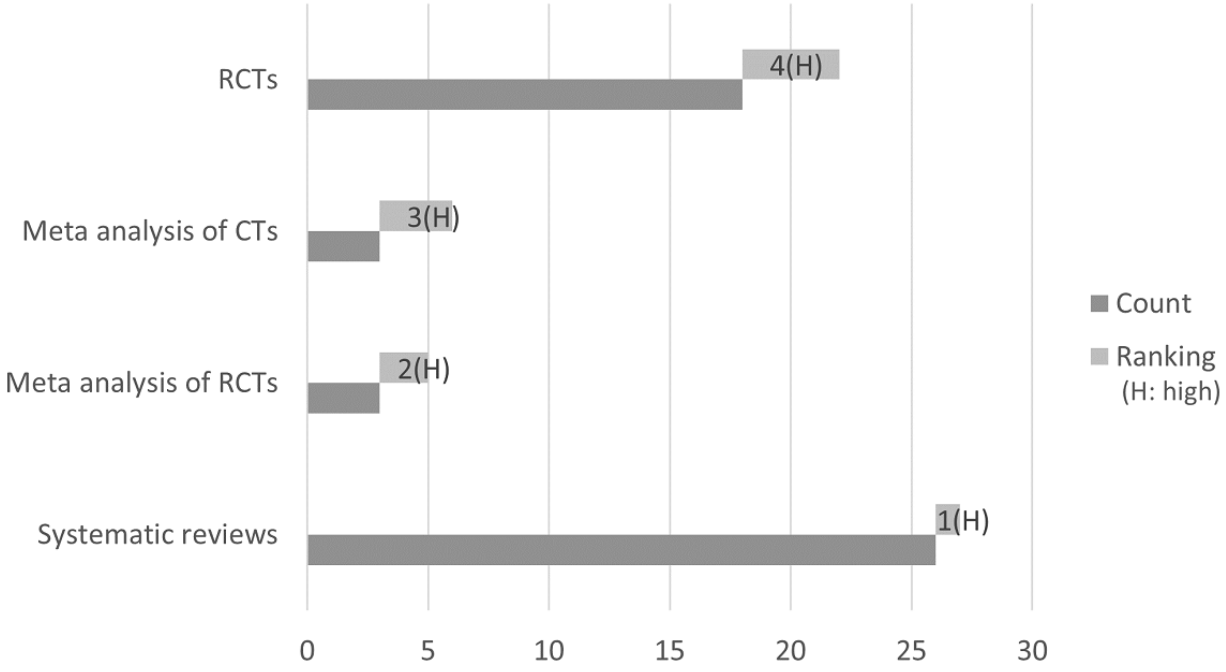
[b]DL: deep learning.

[c]GBT: gradient boosted trees.

Ranking and Summarization

Linking to our previous study results for contextual ranking, we determine the ranks for the studies retrieved against the query in equation 1. The query returned overall 5243 articles, of which 5217 were primary studies and other 26 were SRs. In the primary studies, the highest count of 5187 was assigned to randomized controlled trials, and the rest were distributed among meta-analysis (6), observational studies (5), case reports (2), and others (17). On the basis of the ranking assigned to different publication types in Table 5, the rank and grade values are produced for the selected set of top 50 out of 5243 publications, as shown in Figure 14.

Figure 14. Top 50 publication types with rank and grade values. CT: control trials; RCTs: randomized controlled trials.



Discussion

Principal Findings

The main findings of this study include the design of a comprehensive framework encompasses methods of automatic query construction using PICO, the quality assessment using data-driven approaches, and the ranking of studies using contextual aggregation matrix. Compared with the existing QRM, our high-impact ensemble classifier E-QRM obtained significantly improved accuracy (1723/1894, 90.97% vs 1462/1894, 77.21%). Moreover, the proposed work has the

XSL•FO
RenderX

significance and the implication in multiple domains, and here, we present a few of those applications.

Significance in Evidence-Based Medicine

Research evidence is one of the components of EBM. The proposed literature curation framework is best fit to locate and incorporate the best evidence from the biomedical literature in PubMed to support the evidence-based medical decisions. This work provides a comprehensive set of methods to bring automation to different levels of searching scientific publications, ranging from query construction, quality recognition, and summarization and ranking. The implementers of the EBM can extend and integrate the proposed framework with the health system to use in their daily clinical practice.

Significance in Precision Medicine

Precision medicine is a multidisciplinary approach, which involves genetic characteristics along with clinical and environmental behavior for making a precise decision. There is an opportunity to study how observational studies and clinical trials can be used in conjunction to improve health outcomes in real-world practice settings [31]. This work can greatly contribute to find relevant phenotypes and genetic information precisely from Web-based biomedical resources, including the GenBank [32], MedGen [33], ClinVar [34], and other databases. A set of studies have investigated and developed tools for the evaluation of phenotype candidates using online medical literature [35,36]. The work discussed in this study provides flexibility to apply it to find phenotypes and genome-related clinical trials and evaluate their strength.

Significance in Clinical Decision Support Systems

The CDSS decisions are more trustable if relevant evidence from external sources is timely integrated. The proposed framework could be integrated with the existing CDSS by connecting the query part with the output of the CDSS. The concept of health level seven clinical decision support hooks (HL7 CDS Hooks) [37,38] was very recently extended to include evidence information retrieved from scientific literature. Moreover, the existing CDSSs could be extended to adapt the scientific research evidence in real time.

Significance in Medical Education

Students and researchers require to educate themselves on the existing work from experts. An efficient way to access the existing research work is to implement a system that assists them in a meaningful manner. The proposed framework is capable of providing a unique opportunity to obtain the best evidence in less time and with a higher level of accuracy. Researchers need guidance on whether they have to apply a new method of intervention and at what cost. Patients do need literature to study and compare their conditions with other similar patients and find out about the outcomes of the interventions on other patients.

Limitations of the Work

The summarization research work is yet to mature; therefore, its results are not reported in this study. We have a plan to continue our investigation further to design automated methods and guidelines for the construction of summarization such as designing methods to generate a summary of different articles for the formation of a consolidated evidence. Moreover, we are also interested in investigating the strategies for discovering knowledge from the evaluated quality articles.

Conclusions

There is a great demand for consultation of external clinical evidences to be considered in a complex clinical decision-making process, particularly when internal evidences are insufficient. In addition, medical researchers, students, and patients use them for education, training, and self-awareness about their health problems, respectively. To satisfy these users' needs and desires, we proposed a comprehensive framework for automated curating of biomedical literature, which facilitates the task of bringing a quality research evidence intelligently to the users' desk to assist the users in answering clinical questions and fulfilling their informational needs. We presented a set of methods and guidelines to automate the process of curating biomedical literature at 3 levels: query construction, quality recognition, and ranking and summarization supported with sample results. This proposed automated framework is expected to improve the overall efficiency of clinical staff and researchers in terms of time and effort. However, the proposed system needs to be thoroughly tested on multiple domains before adoption. Our future work will focus on the optimization of quality evaluation and ranking and summarization. The final approval of the evidence by a human is crucial to avoid interpretations if wrong decisions are made by the system.

Acknowledgments

This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center support program (IITP-2017-0-01629) supervised by the Institute of Information & Communications Technology Planning & Evaluation (IITP). This work was supported by an IITP grant funded by the Korean government (MSIT; no 2017-0-00655).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Algorithm: salient term identification.
[DOCX File , 15 KB - medinform_v7i4e13430_app1.docx]

XSL•FO
RenderX

References

1. Milani RV, Lavie CJ. Health care 2020: reengineering health care delivery to combat chronic disease. Am J Med 2015 Apr;128(4):337-343. [doi: 10.1016/j.amjmed.2014.10.047] [Medline: 25460529]
2. Simmons M, Singhal A, Lu Z. Text mining for precision medicine: bringing structure to EHRs and biomedical literature to understand genes and health. Adv Exp Med Biol 2016;939:139-166 [FREE Full text] [doi: 10.1007/978-981-10-1503-8_7] [Medline: 27807747]
3. Li F, Li M, Guan P, Ma S, Cui L. Mapping publication trends and identifying hot spots of research on Internet health information seeking behavior: a quantitative and co-word biclustering analysis. J Med Internet Res 2015 Mar 25;17(3):e81 [FREE Full text] [doi: 10.2196/jmir.3326] [Medline: 25830358]
4. Fiorini N, Lipman D, Lu Z. Towards PubMed 2.0. Elife 2017 Oct 30;6 [FREE Full text] [doi: 10.7554/eLife.28801] [Medline: 29083299]
5. PubMed-NCBI. URL: https://www.ncbi.nlm.nih.gov/pubmed [accessed 2019-08-03]
6. UpToDate. 2016. Evidence-Based Clinical Decision Support at the Point of Care URL: http://www.uptodate.com/home [accessed 2019-08-03]
7. Hussain M, Afzal M, Ali T, Ali R, Khan WA, Jamshed A, et al. Data-driven knowledge acquisition, validation, and transformation into HL7 Arden Syntax. Artif Intell Med 2018 Nov;92:51-70. [doi: 10.1016/j.artmed.2015.09.008] [Medline: 26573247]
8. Leung G. Evidence-based practice revisited. Asia Pac J Public Health 2001;13(2):116-121. [doi: 10.1177/101053950101300210] [Medline: 12597509]
9. Cimino J. An integrated approach to computer-based decision support at the point of care. Trans Am Clin Climatol Assoc 2007;118:273-288 [FREE Full text] [Medline: 18528510]
10. Perez-Rey D, Jimenez-Castellanos A, Garcia-Remesal M, Crespo J, Maojo V. CDAPubMed: a browser extension to retrieve EHR-based biomedical literature. BMC Med Inform Decis Mak 2012 Apr 05;12:29 [FREE Full text] [doi: 10.1186/1472-6947-12-29] [Medline: 22480327]
11. Sahoo S, Valdez J, Kim M, Rueschman M, Redline S. ProvCaRe: characterizing scientific reproducibility of biomedical research studies using semantic provenance metadata. Int J Med Inform 2019 Jan;121:10-18. [doi: 10.1016/j.ijmedinf.2018.10.009] [Medline: 30545485]
12. Bakal G, Talari P, Kakani EV, Kavuluru R. Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations. J Biomed Inform 2018 Jun;82:189-199 [FREE Full text] [doi: 10.1016/j.jbi.2018.05.003] [Medline: 29763706]
13. Sahoo SS, Sheth A, Henson C. Semantic provenance for eScience: managing the deluge of scientific data. IEEE Internet Comput 2008 Jul;12(4):46-54. [doi: 10.1109/mic.2008.86]
14. Wilczynski NL, McKibbon KA, Walter SD, Garg AX, Haynes RB. MEDLINE clinical queries are robust when searching in recent publishing years. J Am Med Inform Assoc 2013;20(2):363-368 [FREE Full text] [doi: 10.1136/amiajnl-2012-001075] [Medline: 23019242]
15. Wilczynski NL, Morgan D, Haynes RB, Hedges Team. An overview of the design and methods for retrieving high-quality studies for clinical care. BMC Med Inform Decis Mak 2005 Jun 21;5:20 [FREE Full text] [doi: 10.1186/1472-6947-5-20] [Medline: 15969765]
16. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inform Assoc 2005;12(2):207-216 [FREE Full text] [doi: 10.1197/jamia.M1641] [Medline: 15561789]
17. Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards automatic recognition of scientifically rigorous clinical research evidence. J Am Med Inform Assoc 2009;16(1):25-31 [FREE Full text] [doi: 10.1197/jamia.M2996] [Medline: 18952929]
18. Sarker A, Mollá D, Paris C. Automatic evidence quality prediction to support evidence-based decision making. Artif Intell Med 2015 Jun;64(2):89-103. [doi: 10.1016/j.artmed.2015.04.001] [Medline: 25983133]
19. Del Fiol G, Michelson M, Iorio A, Cotoi C, Haynes R. A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: comparative analytic study. J Med Internet Res 2018 Jun 25;20(6):e10281 [FREE Full text] [doi: 10.2196/10281] [Medline: 29941415]
20. da Costa Santos CM, de Mattos Pimenta CA, Nobre MR. The PICO strategy for the research question construction and evidence search. Rev Lat Am Enfermagem 2007;15(3):508-511 [FREE Full text] [doi: 10.1590/s0104-11692007000300023] [Medline: 17653438]
21. HL7 International. Arden Syntax v2.8 (Health Level Seven Arden Syntax for Medical Logic Systems, Version 2.8) URL: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=268 [accessed 2019-08-04]
22. Amin M, Banos O, Khan W, Muhammad Bilal H, Gong J, Bui D, et al. On curating multimodal sensory data for health and wellness platforms. Sensors (Basel) 2016 Jun 27;16(7) [FREE Full text] [doi: 10.3390/s16070980] [Medline: 27355955]
23. Banos O, Bilal Amin M, Ali Khan W, Afzal M, Hussain M, Kang BH, et al. The Mining Minds digital health and wellness framework. Biomed Eng Online 2016 Jul 15;15 Suppl 1:76 [FREE Full text] [doi: 10.1186/s12938-016-0179-9] [Medline: 27454608]

XSL•FO
RenderX

24. Health Information Research Unit. 2015. Hedges URL: https://hiru.mcmaster.ca/hiru/HIRU_Hedges_home.aspx [accessed 2019-07-07]
25. Afzal M, Hussain M, Haynes RB, Lee S. Context-aware grading of quality evidences for evidence-based decision-making. Health Informatics J 2019 Jun;25(2):429-445. [doi: 10.1177/1460458217719560] [Medline: 28766402]
26. RapidMiner. Lightning Fast Data Science Platform for Teams URL: https://rapidminer.com/ [accessed 2019-07-07]
27. Bui DD, Del Fiol G, Hurdle JF, Jonnalagadda S. Extractive text summarization system to aid data extraction from full text in systematic review development. J Biomed Inform 2016 Dec;64:265-272 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.014] [Medline: 27989816]
28. Ebell M, Siwek J, Weiss B, Woolf S, Susman J, Ewigman B, et al. Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. J Am Board Fam Med 2004 Jan 01;17(1):59-67 [FREE Full text] [doi: 10.3122/jabfm.17.1.59] [Medline: 15014055]
29. Boudin F, Nie J, Bartlett J, Grad R, Pluye P, Dawes M. Combining classifiers for robust PICO element detection. BMC Med Inform Decis Mak 2010 May 15;10:29 [FREE Full text] [doi: 10.1186/1472-6947-10-29] [Medline: 20470429]
30. Huang K, Chiang I, Xiao F, Liao CC, Liu CC, Wong JM. PICO element detection in medical text without metadata: are first sentences enough? J Biomed Inform 2013 Oct;46(5):940-946 [FREE Full text] [doi: 10.1016/j.jbi.2013.07.009] [Medline: 23899909]
31. Chambers DA, Feero WG, Khoury MJ. Convergence of implementation science, precision medicine, and the learning health care system: a new model for biomedical research. J Am Med Assoc 2016 May 10;315(18):1941-1942 [FREE Full text] [doi: 10.1001/jama.2016.3867] [Medline: 27163980]
32. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2007 Jan;35(Database issue):D21-D25 [FREE Full text] [doi: 10.1093/nar/gkl986] [Medline: 17202161]
33. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2015 Jan;43(Database issue):D6-17 [FREE Full text] [doi: 10.1093/nar/gku1130] [Medline: 25398906]
34. Landrum M, Lee J, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 2016 Jan 04;44(D1):D862-D868 [FREE Full text] [doi: 10.1093/nar/gkv1222] [Medline: 26582918]
35. Henderson J, Bridges R, Ho J, Wallace B, Ghosh J. PheKnow-Cloud: a tool for evaluating high-throughput phenotype candidates using online medical literature. AMIA Jt Summits Transl Sci Proc 2017;2017:149-157 [FREE Full text] [Medline: 28815124]
36. Henderson J, Ke J, Ho J, Ghosh J, Wallace B. Phenotype Instance Verification and Evaluation Tool (PIVET): a scaled phenotype evidence generation framework using web-based medical literature. J Med Internet Res 2018 May 04;20(5):e164 [FREE Full text] [doi: 10.2196/jmir.9610] [Medline: 29728351]
37. GitHub. CDS Hooks URL: http://cds-hooks.org/ [accessed 2018-07-10]
38. Rasmussen L, Overby C, Connolly J, Chute C, Denny J, Freimuth R, et al. Practical considerations for implementing genomic information resources. Experiences from eMERGE and CSER. Appl Clin Inform 2016 Sep 21;7(3):870-882 [FREE Full text] [doi: 10.4338/ACI-2016-04-RA-0060] [Medline: 27652374]

Abbreviations

API: application programming interface
AUC: area under the curve
CDSS: clinical decision support system
CQ: Clinical Queries
DL: deep learning
DT: decision tree
EBM: evidence-based medicine
EBP: evidence-based practice
eFetch: Entrez Fetch
EHR: electronic health record
E-QRM: extended quality recognition model
GBT: gradient boosted trees
GLM: generalized linear model
HIS: health care information system
IITP: Institute of Information & Communications Technology Planning & Evaluation
KB: knowledge base
kNN: k-nearest neighbor
MeSH: Medical Subject Headings
ML: machine learning
MLM: medical logic modules

XSL•FO
RenderX

MSIT: Ministry of Science and ICT
NB: Naïve Bayes
NLP: natural language processing
PICO: Patient/problem, Intervention, Comparison, and Outcome
QRM: quality recognition model
SNOMED-CT: Systematized Nomenclature of Medicine-Clinical Terms
SR: systematic review
STI: salient term identification
SVM: support vector machine
UMLS: Unified Medical Language System

Edited by G Eysenbach; submitted 17.01.19; peer-reviewed by R Sadeghi, G Kolostoumpis, F Alam, L Zhang, I Tagkopoulos, G Lim; comments to author 22.03.19; revised version received 07.08.19; accepted 26.09.19; published 09.12.19.

Please cite as:
Afzal M, Hussain M, Malik KM, Lee S
Impact of Automatic Query Generation and Quality Recognition Using Deep Learning to Curate Evidence From Biomedical Literature: Empirical Study
JMIR Med Inform 2019;7(4):e13430
URL: http://medinform.jmir.org/2019/4/e13430/
doi:10.2196/13430
PMID:31815673

©Muhammad Afzal, Maqbool Hussain, Khalid Mahmood Malik, Sungyoung Lee. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 09.12.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Efficient Reuse of Natural Language Processing Models for Phenotype-Mention Identification in Free-text Electronic Medical Records: A Phenotype Embedding Approach

Honghan Wu[1,2,3], BEng, DPhil; Karen Hodgson[4], DPhil; Sue Dyson[4], MD; Katherine I Morley[4,5,6], DPhil; Zina M Ibrahim[4,7], DPhil; Ehtesham Iqbal[4], BEng; Robert Stewart[4,5], MD, DPhil; Richard JB Dobson[4,7], DPhil; Cathie Sudlow[1,3], MD, DPhil

[1]Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
[2]School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China
[3]Health Data Research UK, University of Edinburgh, Edinburgh, United Kingdom
[4]Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom
[5]South London and Maudsley NHS Foundation Trust, London, United Kingdom
[6]Centre for Epidemiology and Biostatistics, Melbourne School of Global and Population Health, The University of Melbourne, Melbourne, Australia
[7]Health Data Research UK, University College London, London, United Kingdom

Corresponding Author:
Honghan Wu, BEng, DPhil
Centre for Medical Informatics
Usher Institute
University of Edinburgh
9 Little France Road
Edinburgh, EH16 4UX
United Kingdom
Phone: 44 01316517882
Email: honghan.wu@ed.ac.uk

Abstract

Background: Much effort has been put into the use of automated approaches, such as natural language processing (NLP), to mine or extract data from free-text medical records in order to construct comprehensive patient profiles for delivering better health care. Reusing NLP models in new settings, however, remains cumbersome, as it requires validation and retraining on new data iteratively to achieve convergent results.

Objective: The aim of this work is to minimize the effort involved in reusing NLP models on free-text medical records.

Methods: We formally define and analyze the model adaptation problem in phenotype-mention identification tasks. We identify "duplicate waste" and "imbalance waste," which collectively impede efficient model reuse. We propose a phenotype embedding–based approach to minimize these sources of waste without the need for labelled data from new settings.

Results: We conduct experiments on data from a large mental health registry to reuse NLP models in four phenotype-mention identification tasks. The proposed approach can choose the best model for a new task, identifying up to 76% waste (duplicate waste), that is, phenotype mentions without the need for validation and model retraining and with very good performance (93%-97% accuracy). It can also provide guidance for validating and retraining the selected model for novel language patterns in new tasks, saving around 80% waste (imbalance waste), that is, the effort required in "blind" model-adaptation approaches.

Conclusions: Adapting pretrained NLP models for new tasks can be more efficient and effective if the language pattern landscapes of old settings and new settings can be made explicit and comparable. Our experiments show that the phenotype-mention embedding approach is an effective way to model language patterns for phenotype-mention identification tasks and that its use can guide efficient NLP model reuse.

(JMIR Med Inform 2019;7(4):e14782) doi:10.2196/14782

XSL•FO
RenderX

KEYWORDS

natural language processing; text mining; phenotype; word embedding; phenotype embedding; model adaptation; electronic health records; machine learning; clustering

Introduction

Compared to structured components of electronic health records (EHRs), free-text comprises a much deeper and larger volume of health data. For example, in a recent geriatric syndrome study [1], unstructured EHR data contributed a significant proportion of identified cases: 67.9% cases of falls, 86.6% cases of visual impairment, and 99.8% cases of lack of social support. Similarly, in a study of comorbidities using a database of anonymized EHRs of a psychiatric hospital in London (the South London and Maudsley NHS Foundation Trust [SLaM]) [2], 1899 cases of comorbid depression and type 2 diabetes were identified from unstructured EHRs, while only 19 cases could be found using structured diagnosis tables. The value of unstructured records for selecting cohorts has also been widely reported [3,4]. Extracting clinical variables or identifying phenotypes from unstructured EHR data is, therefore, essential for addressing many clinical questions and research hypotheses [5-7].

Automated approaches are essential to surface such deep data from free-text clinical notes at scale. To make natural language processing (NLP) tools accessible for clinical applications, various approaches have been proposed, including generic, user-friendly tools [8-10] and Web services or cloud-based solutions [11-13]. Among these approaches, perhaps, the most efficient way to facilitate clinical NLP projects is to adapt pretrained NLP models in new but similar settings [14], that is, to reuse existing NLP solutions to answer new questions or to work on new data sources. However, it is very often burdensome to reuse pretrained NLP models. This is mainly because NLP models essentially abstract language patterns (ie, language characteristics representable in computable form) and subsequently use them for prediction or classification tasks. These patterns are prone to change when the document set (corpus) or the text mining task (what to look up) changes. Unfortunately, when it comes to a new setting, it is uncertain which patterns have and have not changed. Therefore, in practice, random samples are drawn to validate the performance of an existing NLP model in a new setting and subsequently to plan the adaptation of the model based on the validation results.

Such "*blind*" adaptation is costly in the clinical domain because of barriers to data access and expensive clinical expertise needed for data labelling. The "*blindness*" to the similarities and differences of language pattern landscapes between the source (where the model was trained) and target (the new task) settings causes (at least) two types of potentially unnecessary, wasted effort, which may be avoidable. First, for data in the target setting with the same patterns as in the source setting, any validation or retraining efforts are unnecessary because the model has already been trained and validated on these language patterns. We call this type of wasted effort the "*duplicate waste*." The second type of *waste* occurs if the distribution of new language patterns in the target setting is unbalanced, that is, some—but not all—data instances belong to different language patterns. The model adaptation involves validating the model on these new data and further adjusting it when performance is not good enough. Without the knowledge of which data instances belong to which language patterns, data instances have to be randomly sampled for validation and adaptation. In most cases, a minimal number of instances of every pattern need to be processed, so that convergent results can be obtained. This will usually be achieved via iterative validation and adaptation process, which will inevitably cause commonly used language patterns to be over represented, resulting in the model being over validated/retrained on such data. Such unnecessary efforts on commonly used language patterns result from the pattern imbalance in the target setting, which unfortunately is the norm in almost all real-world EHR datasets. We call this "*imbalance waste*."

The ability to make language patterns *visible* and comparable will address whether an NLP model can be adapted to a new task and, importantly, provide guidance on how to solve new problems effectively and efficiently through the *smart* adaptation of existing models. In this paper, we introduce a contextualized embedding model to *visualize* such patterns and provide guidance for reusing NLP models in phenotype-mention identification tasks. Here, a phenotype mention denotes an appearance of a word or phrase (representing a medical concept) in a document, which indicates a phenotype related to a person. We note two aspects of this definition:

1. Phenotype mention ≠ Medical concept mention. When a medical concept mentioned in a document does not indicate a phenotype relating to a person (eg, cases in the last two rows of Table 1), it is not a phenotype mention.
2. Phenotype mention ≠ Phenotype. Phenotype (eg, diseases and associated traits) is a specific patient characteristic [15] and a patient-level feature, (eg, a binary value indicating whether a patient is a smoker). However, for the same phenotype, a patient might have multiple phenotype mentions. For example, "xxx is a smoker" could be mentioned in different documents or even multiple times in one document, and each of these appearances is a phenotype mention.

The focus of this work is to minimize the effort in reusing existing NLP model(s) in solving new tasks rather than proposing a novel NLP model for phenotype-mention identification. We aim to address the problem of NLP model transferability in the task of extracting mentions of phenotypes from free-text medical records. Specifically, the task is to identify the above-defined phenotype mentions and the contexts in which they were mentioned [10]. Table 1 explains and provides examples of contextualized phenotype mentions. The research question to be investigated is formally defined as mentioned in Textbox 1 and illustrated in Figure 1.

XSL•FO
RenderX

Table 1. The task of recognizing contextualized phenotype mentions is to identify mentions of phenotypes from free-text records and classify the context of each mention into five categories (listed in the second column of Table 1). The last two rows give examples of nonphenotype mentions—the two sentences are not describing incidents of a condition.

Examples	Types of phenotype mentions
49 year old man with *hepatitis c*	Positive mention[a]
With no evidence of *cancer recurrence*	Negated mention[a]
…Is concerning for local *lung cancer recurrence*	Hypothetical mention[a]
PAST MEDICAL HISTORY: (1) *Atrial Fibrillation*, (2)...	History mention[a]
Mother was A positive, *hepatitis C carrier*, and...	Mention of phenotype in another person[a]
She visited the *HIV* clinic last week.	Not a phenotype mention
The patient asked for information about *stroke*.	Not a phenotype mention

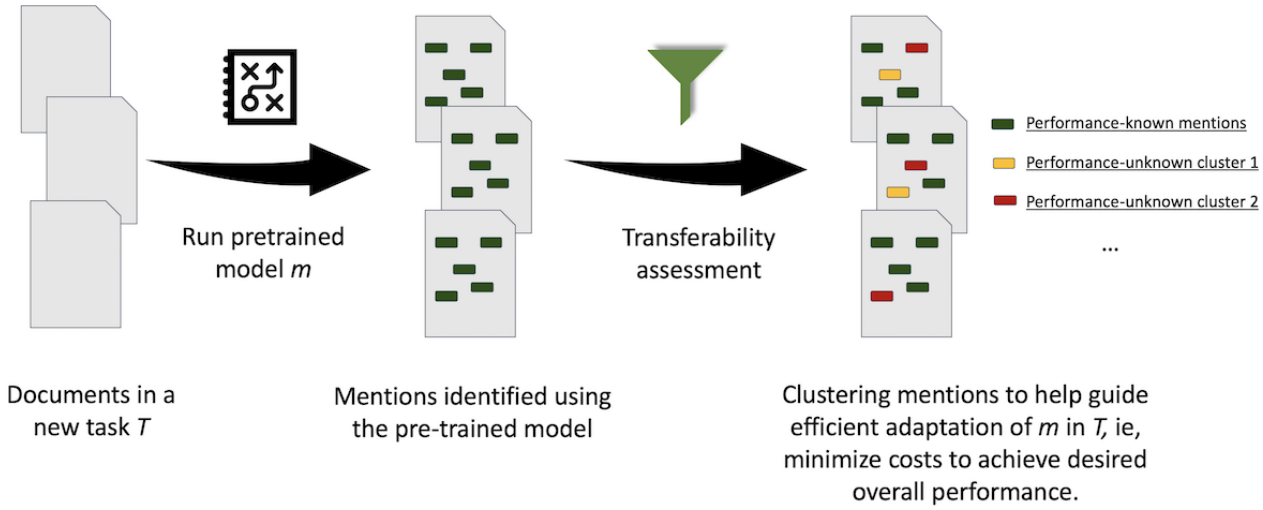
[a]Contextualized mentions.

Textbox 1. Research question.

Definition 1. Given an natural language processing model (denoted as) previously trained for some phenotype-mention identification task(s), and a new task (denoted as , where either phenotypes to be identified are new or the dataset is new, or both are new), *m* is used in to identify a set of phenotype mentions—denoted as *S*. The research question is how to partition *S* to meet the following criteria:

1. A maximum p-known subset Sknown where m's performance can be properly predicted using prior knowledge of m;
2. p-unknown subsets: {Su1, Su2…Suk}, which meet the following criteria:

Figure 1. Assess the transferability of a pretrained model in solving a new task: Discriminate between differently inaccurate mentions identified by the model in the new setting.



The identification of "*p-known*" subset (criterion 1) will help eliminate "*duplicate waste*" by avoiding unnecessary validation and adaptation on those phenotype mentions. On the other hand, separating the rest of the annotations into "*p-unknown*" subsets allows processing mentions based on their *performance-relevant* characteristics separately, which in turn helps avoid "*imbalance waste.*" The abovementioned criterion 2a ensures completeness of coverage of all performance-unknown mentions and criterion 2b ensures no overlaps between mention subsets, so that no duplicated effort will be put on the same mentions. Criterion 2c requires that the partitioning of the mentions is *performance-relevant*, meaning that model performance on a small number of samples can be generalized to the whole subset that they are drawn from. Lastly, a small (criterion 2d) enables efficient adaptation of a model.

Methods

Dataset and Adaptable Phenotype-Mention Identification Models

Recently, we developed SemEHR [10]—a semantic search toolkit aiming to use interactive information retrieval

XSL•FO
RenderX

functionalities to replace NLP building, so that clinical researchers can use a browser-based interface to access text mining results from a generic NLP model and (optionally) keep getting better results by iteratively feeding them back to the system. A SLaM instance of this system has been trained for supporting six comorbidity studies (62,719 patients and 17,479,669 clinical notes in total), where different combinations of physical conditions and mental disorders are extracted and analyzed. Multimedia Appendix 1 provides details about the user interface and model performance. These studies effectively generated 23 phenotype-mention identification models and relevant labelled data (>7000 annotated documents), which we use to study model transferability.

Foundation of the Proposed Approach

Our approach is based on the following assumption about a language pattern representation model:

- *Assumption 1*. There exists a pattern representation model, A, for identifying language patterns of phenotype mentions with the following characteristics:
  1. Each phenotype mention can be characterized by only one language pattern.
  2. Patterns are largely shared by different mentions.
  3. There is a deterministic association between NLP models' performances with such language patterns.

- *Theorem 1*. Given A, a pattern model meeting Assumption 1, m—an NLP model, T—a new task, let Pm be the pattern set Aidentifies from dataset(s) that m was trained or validated on; let PT be the pattern set A identifies from Sthe set of all mentions identified by m in T. Then, the problem defined in Definition 1 can be solved by a solution, where Pm ∩ PTis the "p-known" subset and PT – Pm ∩ PT is "p-unknown" subsets.

Proof of Theorem 1 can be found in the Multimedia Appendix 2. The rest of this section provides details of a realization of *A*using distributed representation models.

Distributed Representation for Contextualized Phenotype Mentions

In computational linguistics, statistical language models are, perhaps, the most common approach to quantify word sequences, where a distribution is used to represent the probability of a sequence of words: $P(w_1 \ldots w_n)$. Among such models, the bag-of-words (BOW) model [15] is perhaps the earliest and simplest, yet widely used and efficient in certain tasks [16]. To overcome BOW's limitations (eg, ignoring semantic similarities between words), more complex models were introduced to represent word semantics [17-19]. Probably, the most popular alternative is the distributed representation model [20], which uses a vector space to model words, so that word similarities can be represented as distances between their vectors. This concept has since been extensively followed up, extended, and shown to significantly improve NLP tasks [21-26].

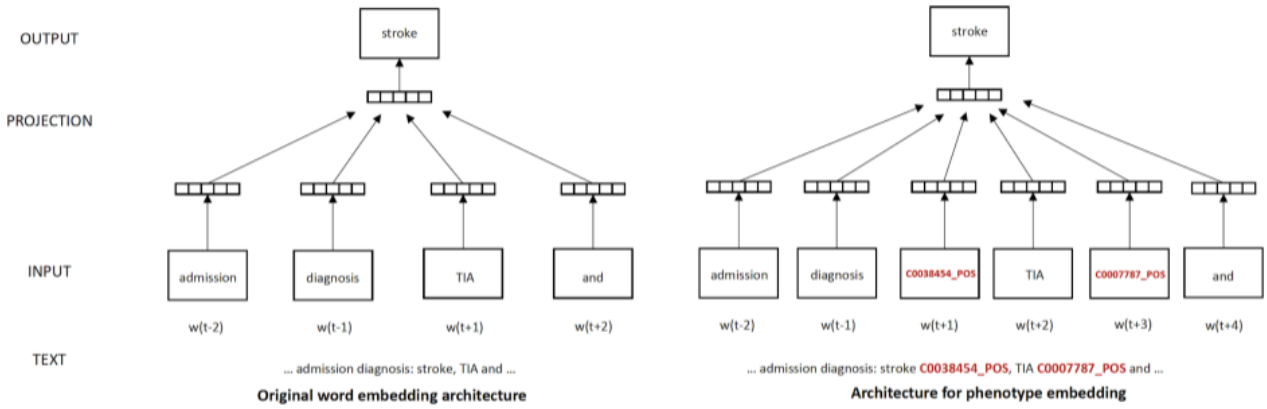
In original distributed representation models, the semantics of one word is encoded in one single vector, which makes it impossible to disambiguate different semantics or contexts that one word might be used for in a corpus. Recently, various (bidirectional) long short-term memory models were proposed to learn contextualized word vectors [27-29]. However, such linguistic contexts are not the phenotype contexts (Table 1) that we seek in this paper.

Inspired by the good properties of distributed representations for words, we propose a phenotype encoding approach that aims to model the language patterns of contextualized phenotype mentions. Compared to word semantics, phenotype semantics are represented in a larger context, at the sentence or even paragraph level (eg, *he worries about contracting HIV*; here, HIV is a hypothetical phenotype mention). The key idea of our approach is to use explicit mark-ups to represent phenotype semantics in the text, so that they can be learned through an approach similar to the word embedding learning framework.

Figure 2 illustrates our framework for extending the continuous BOW word embedding architecture to capture the semantics of contextualized phenotype mentions. Explicit *mark-ups of phenotype mentions* are added to the architecture as placeholders for phenotype semantics. A mark-up (eg, C0038454_POS) is composed of two parts: phenotype identification (eg, C0038454) and contextual description (eg, POS). The first part identifies a phenotype using a standardized vocabulary. In our implementation, the Unified Medical Language System (UMLS) [30] was chosen for its broad concept coverage and the provision of comprehensive synonyms for concepts. The first benefit of using a standardized phenotype definition is that it helps in grouping together mentions of the same phenotype using different names. For example, using UMLS concept identification of C0038454 for STROKE helps combining together mentions using *Stroke*, *Cerebrovascular Accident*, *Brain Attack*, and other 43 synonyms. The second benefit is from the concept relations represented in the vocabulary hierarchy, which helps the transferability computation that we will elaborate on later (step 3 in the next subsection). The second part of a phenotype mention mark-up is to identify the mention context. Six types of contexts are supported: POS for *positive mention*, NEG for *negated mention*, HYP for *hypothetical mention*, HIS for *history mention*, OTH for *mention of the phenotype in another person*, and NOT for *not a phenotype mention*.

XSL•FO
RenderX

Figure 2. The framework to learn contextualized phenotype embedding from labelled data that an natural language processing model m was trained or validated on. TIA: transient ischemic attack.



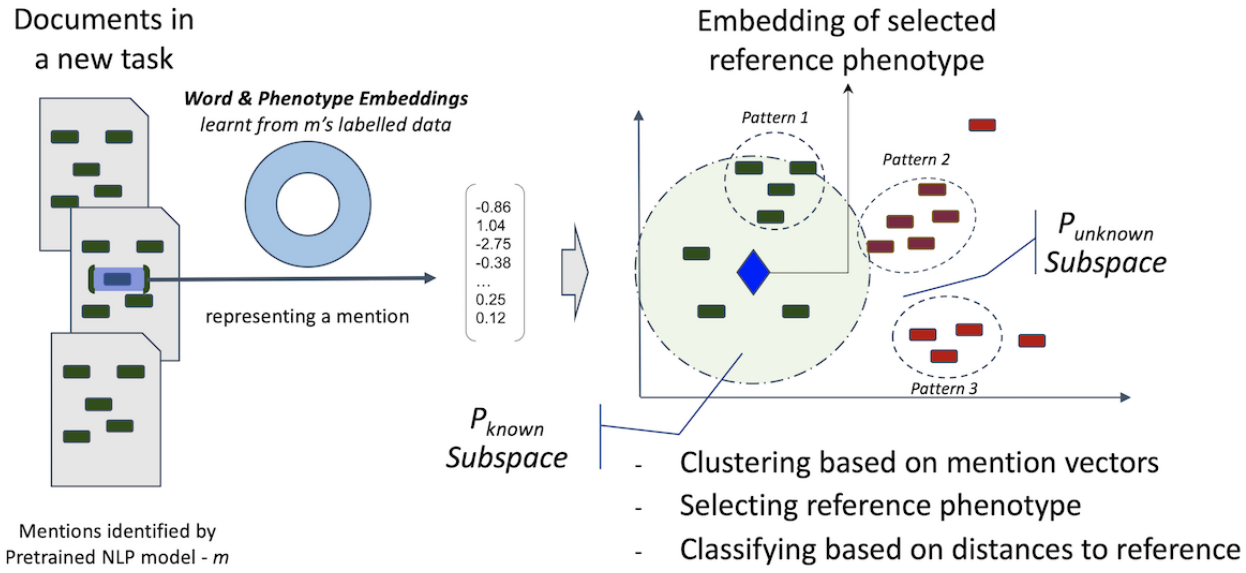
The *phenotype mention mark-ups* can be populated using labelled data that NLP models were trained or validated on. In our implementation, the mark-ups were generated from the labelled subset of SLaM EHRs.

Using Phenotype Embedding and Their Semantics for Assessing Model Transferability

The embeddings learned (including both word and contextualized phenotype vectors) are the building blocks underlying the language pattern representation model—*A*, as introduced at the beginning of this section, which is to compute P_m (the landscape of language patterns that *m* is familiar with) and P_T (the landscape of language patterns in the new task *T*) for assessing and guiding NLP model adaptation for new tasks.

Figure 3 illustrates the architecture of our approach. The double-circle shape denotes the embeddings learned from *m*'s labelled data. Essentially, the process is composed of two phases: (1) the documents from a new task (on the left of the figure) are annotated with phenotype mentions using a pretrained model *m* and (2) a classification task uses the abovementioned embeddings to assess each mention—whether it is an instance of *p-known* (something similar enough to what *m* is familiar with) or any subset of *p-unknown* (something that is new to *m*). Specifically, the process is composed of the following steps:

Figure 3. Architecture of phenotype embedding-based approach for transferring pretrained natural language processing models for identifying new phenotypes or application to new corpora. The word and phenotype embedding model is learned from the training data of the reusable models in its source domain (the task that m was trained for). No labelled data in the target domain (new setting) are required for the adaptation guidance. NLP: natural language processing.



1) Vectorize phenotype mentions in a new task: Each mention in the new task will be represented as a vector of real numbers using the learned embedding model to combine its surrounding words as context semantics. Formally, the reference is chosen as shown in Textbox 2.

XSL•FO
RenderX

Textbox 2. Vector representation of a phenotype mention.

Let *s* be a mention identified by *m* in the new task, where *s*can be represented by a function defined as follows:

(1)

Where

is the embedding model to convert a word token into a vector, t_j is the jth word in a document, *i* is the offset of the first word of *s* in the document, *l* is the number of words in *s*, and *f* is a function to combine a set of vectors into a result vector (we use *average* in our implementation).

With such representations, all mentions are effectively put in a vector space (depicted as a 2D space on the right of the figure for illustration purposes).

2) Identify clusters (language patterns) of mention vectors: In the vector space, clusters are naturally formed based on geometric distances between mention vectors. After trying different clustering algorithms and parameters, DBScan [31] was chosen on Euclidean distance in our implementation for vector clustering. Essentially, each cluster is a set of mentions considered to share the same (or similar enough) underlying language pattern, meaning that language patterns in the new task are technically the vector clusters. We chose the cluster centroid (arithmetic mean) to represent a cluster (ie, its underlying language pattern).

3) Choose a reference vector for classifying language patterns: After clusters (language patterns) are identified, the next step is to classify them as p-known or subsets of p-unknown. We choose a reference vector–based approach, classifying patterns using the distance to a selected vector. Such a reference vector is picked up (when the phenotype to be identified has been trained in m) or generated (when the phenotype is new to m) from the learned phenotype embeddings the model m has seen previously. Apparently, when the phenotype to be identified in the new task is new to m (not in the set of phenotypes it was developed for), the reference phenotype needs to be carefully selected, so that it can help produce a sensible separation between p-known and p-unknown clusters. We use the semantic similarity (distance between two concepts in the UMLS tree structure) to choose the most similar phenotype from the phenotype list m was trained for. Formally, the reference is chosen as shown in Textbox 3.

Textbox 3. Reference phenotype selection

Let c_p be the Unified Medical Language System concept for a phenotype to be identified in the new task and C_m be the set of phenotype concepts that *m* was trained for, the reference phenotype choosing function is

(2)

Where *D* is a distance function to calculate the steps between two nodes in the Unified Medical Language System concept tree.

Once the reference phenotype has been chosen, the reference vector can be selected or generated (eg, use the average) from this phenotype's contextual embeddings.

4) Classify language patterns to guide model adaptation: Once the reference vector has been selected, clusters can be classified based on the distances between their centroids (representative vectors of clusters) and the reference vector. Once a distance threshold is chosen, this distance-based classification partitions the vector space into two subspaces using the reference vector as the center: the subspace whose distance to the center is less than the threshold is called p-known subspace and the remainder is the p-unknown subspace. The union of clusters whose centroids are within the p-known subspace is p-known, meaning m's performances on them can be predicted without further validation (removing duplicate waste). Other clusters are p-unknown clusters, and m can be validated or further trained on each p-unknown cluster separately instead of blindly across all clusters. This will remove imbalance waste.

Results

Associations Between Embedding-Based Language Patterns and Model Performances

As stated in the beginning of Method section, our approach is based on three assumptions about language patterns. Therefore, it is essential to quantify to what extent the language patterns identified by our embedding-based approach meet these assumptions. The first assumption—a phenotype mention can be assigned to one and only to one language pattern—is met in our approach, since (1) (Equation 1) is a one-to-one function and (2) DBScan algorithm (the vector clustering function chosen in our implementation) is also a one-to-one function. Assumption 2 can be quantified by the percentage of mentions that can be assigned to a cluster. This percentage can be increased by increasing the epsilon (EPS) parameter (the maximum distance between two data items for them to be considered in the same neighborhood) in DBScan. However, the degree to which mentions are clustered together needs to be

XSL•FO
RenderX

balanced against the consequence of the reduced ability to identify performance-related language patterns, which is the third assumption: associations between language patterns and model performance. To quantify such an association, we propose a metric called bad guy separate power (SP), as defined in Equation 3 below (Textbox 4). The aim is to measure to what extent a clustering can assemble incorrect data items (false-positive mentions of phenotypes) together.

Textbox 4. Bad Guy Separate Power.

Let C be a set of binary data items –

(stands for true; stands for false), given a clustering result $\{C_1 \ldots C_k \mid C_1 \cup C_2 \ldots \cup C_k = C\}$, its separate power for f typed data items is defined as follows:

(3)

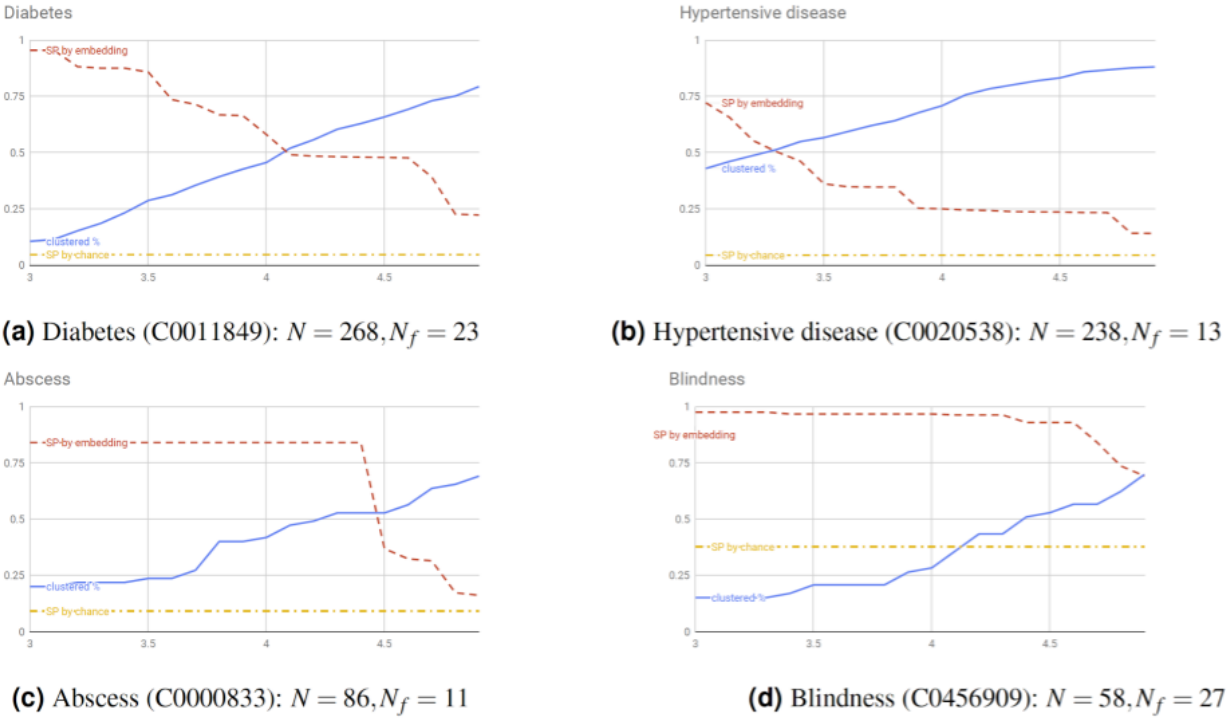
In our scenario, we would like to see clustering being able to separate easy cases (where good performance is achieved) from difficult cases (where performance is poor) for a model .

To quantify the clustering percentage, the ability to separate mentions based on model performances and the interplay between the two, we conducted experiments on selected phenotypes by continuously increasing the clustering parameter

EPS from a low level. Figure 4 shows the results. In this experiment, we label mentions into two types—correct and incorrect—using SemEHR labelled data on the SLaM corpus. Specifically, for the mention types in Table 1, incorrect mentions are those denoted “not-a-phenotype-mention” and the remainder are labelled as correct. We chose incorrect as the f in equation 3, as we evaluate the separate power on incorrect mentions. Four phenotypes were selected for this evaluation: *Diabetes* and *Hypertensive disease* were selected because they were most validated phenotypes and *Abscess* (with 13% incorrect mentions) and *Blindness* (with 47% incorrect mentions) were chosen to represent NLP models with different levels of performance. The figure shows a clear trend in all cases: As EPS increases, the clustered percentage increases, but with decreasing separate power. This confirms a trade-off between the coverage of identified language patterns and how good they are. Regarding separate power, the performance on two selected common phenotypes (Figure 4a and 4b) is generally worse than that for the other phenotypes, starting with lower power, which decreases faster as the EPS increases. The main reason is that the difficult cases (mentions with poor performance) in the two commonly encountered phenotypes are relatively rare (diabetes: 8.5%; hypertensive disease: 5.5%). In such situations, difficult cases are harder to separate because their patterns are underrepresented. However, in general, compared to random clustering, the embedding-based clustering approach brings in much better separate power in all cases. This confirms a high-level association between identified clusters and model performance. In particular, when the proportion of difficult cases reaches near 50% (Figure 4d), the approach can keep *SP* values almost constantly near 1.0 when the EPS increases. This means it can almost always group difficult cases in their own clusters.

XSL•FO
RenderX

Figure 4. Clustered percentage versus separate power on difficult cases. The x-axis is the Epsilon (EPS) parameter of the DBScan clustering algorithm---the longest distance between any two items within a cluster; the y-axis is the percentage. Two types of changing information (as functions of EPS) are plotted on each panel: clustered percentage (solid line) and SP on incorrect cases (false-positive mentions of phenotypes). The latter has two series: (1) SP by chance (dash dotted line) when clustering by randomly selecting mentions and (2) SP by clustering using phenotype embedding (dashed line). N: number of all mentions; N_f: number of false-positive mentions; SP: separate power.



(a) Diabetes (C0011849): $N = 268, N_f = 23$

(b) Hypertensive disease (C0020538): $N = 238, N_f = 13$

(c) Abscess (C0000833): $N = 86, N_f = 11$

(d) Blindness (C0456909): $N = 58, N_f = 27$

Model Adaptation Guidance Evaluation

Technically, the guidance to model adaptation is composed of two parts: avoid *duplicate waste* (skip validation/training efforts on cases the model is already familiar with) and avoid *imbalance waste* (group new language patterns together, so that validation/continuous training on each group separately can be more efficient than doing it over the whole corpus). To quantify the guidance effectiveness, the following metrics are introduced.

- Duplicate waste: This is the number of mentions whose patterns fall into what the model m is familiar with. The quantity

  is the proportion of mentions that needs no validation or retraining before reusing .
- Imbalance waste: To achieve convergence performance, an NLP model needs to be trained on a minimal number (denoted as e) of samples from each language pattern. Calling the language pattern set in a new task as C={C1…Ck}, the following equation counts the minimum number of samples needed to achieve convergent results in “blind” adaptations:

(4)

When the language patterns are identifiable, the *Imbalance waste* that can be avoided is quantified as

- Accuracy: To evaluate whether our approach can really identify familiar patterns, we quantify the accuracy of those within-threshold clusters and those within-threshold single mentions that are not clustered. Both macro-accuracy (average of all cluster accuracies) and micro-accuracy (overall accuracy) are used (detailed explanations provided elsewhere [32]).

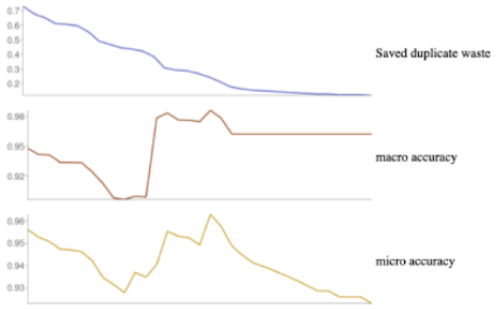
Figure 5 shows the results of our NLP model adaptation guidance on four phenotype-identification tasks. For each new phenotype-identification task, the NLP model (pre)trained for the semantically most similar (defined in Equation 2) phenotype was chosen as the reuse model. Models and labelled data for the four pairs of phenotypes were selected from six physical comorbidity studies on SLaM data. Figure 5 shows that identified mentions have a high proportion of avoidable duplicate waste in all four cases: Diabetes and heart attack start with 50%, whereas stroke and multiple sclerosis are >70%. Such avoidable duplicate waste decreases when the threshold increases. The threshold is on similarity instead of distance, meaning that new patterns need to be more similar to the reuse model’s embeddings to be counted as familiar patterns. Therefore, it is understandable that duplicate waste decreases in such scenarios. In terms of accuracy, one would expect this to increase, as only more similar patterns are left when the threshold increases. However, interestingly, in all cases, both macro- and micro-accuracies decrease slightly before increasing to reach near 1.0. This is a phenomenon worth future investigation. In general, the changes in accuracy are small (0.03-0.08), while accuracy remains high (>0.92). Given these observations, the threshold is normally set at 0.01, to optimize the avoidance of duplicate waste with minimal effect on

XSL•FO
RenderX

accuracy. Specifically, in all cases, more than half of the identified mentions (>50% for Figure 5a and 5b; >70% for Figure 5c and 5d) do not need any validation/training to obtain an accuracy of >0.95. In terms of effective adaptation on new patterns, the percentages of avoidable imbalance waste in all cases are around 80%, confirming that a much more efficient retraining on data can be achieved through language pattern-based guidance.

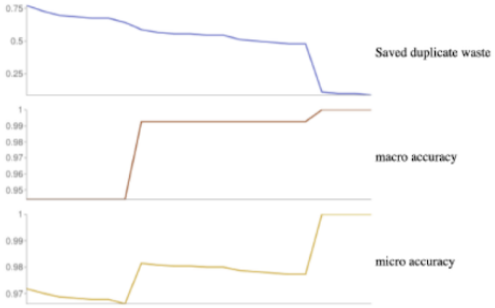
Figure 5. Identifying new phenotypes by reusing natural language processing models pretrained for semantically close phenotypes: The four pairs of phenotype-mention identification models are chosen from SemEHR models trained on SLaM data; DBScan Epsilon (EPS) value=3.8, and imbalance waste is calculated on e=3, meaning at least 3 samples are needed for training from each language pattern. The x-axis is the similarity threshold, ranging from 0.0 to 0.8; the y-axes, from top to bottom, are the proportion of duplicate waste saved over total number of mentions, macro-accuracy, and micro-accuracy, respectively.

(a) New task: *Diabetes (C0011849)*;
Reuse model: *Type 2 Diabetes(C0011860)*;
#Mentions/#not-a-mention: 268/23;
#Cluster:15;
Saved Imbalance Waste: 40 or 83%



(b) New task:*Stroke(C0038454)*;
Reuse model: *Heart Attack(C0027051)*;
#Mentions/#not-a-mention: 238/13;
#Cluster:16;
Saved Imbalance Waste: 39 or 82%

(c) New task: *Heart Attack(C0027051)*;
Reuse model: *Infarct(C0021308)*;
#Mentions/#not-a-mention: 54/11;
#Cluster:5;
Saved Imbalance Waste: 11 or 78%



(d) New task:*Multiple Sclerosis(C0026769)*;
Reuse model: *Myasthenia Gravis(C0026896)*;
#Mentions/#not-a-mention: 104/4;
#Cluster:5;
Saved Imbalance Waste: 14 or 85%

Effectiveness of Phenotype Semantics in Model Reuse

When considering NLP model reuse for a new task, if there is no existing model that has been developed for the same phenotype-mention identification task, our approach will choose a model trained for a phenotype that is most semantically similar to it (based on Equation 2). To evaluate the effectiveness of such semantic relationships in reusing NLP models, we conducted experiments on the previous four phenotypes by using phenotype models with different levels of semantic similarities. Table 2 shows the results. In all cases, reusing models trained for more similar phenotypes can identify more *duplicate waste* using the same parameter settings. The first three cases in the table can also achieve better accuracies, while *multiple sclerosis* had slightly better accuracy by reusing the *diabetes* model than the more semantically similar *myasthenia gravis*. However, the latter identified 46% more *duplicate waste*.

XSL•FO
RenderX

Table 2. Comparisons of the performance of reusing models with different semantic similarity levels. Similarity threshold: 0.01; DBScan EPS: 0.38. Reusing models trained for more (semantically) similar phenotypes achieved adaptation results with less effort (more duplicate waste identified) in all cases, and the results were more accurate in three of four cases. Performance metrics of better reusable models are highlighted as bold numbers.

Model reuse cases	Duplicate waste	Macro-accuracy	Micro-accuracy
Diabetes by Type 2 Diabetes[a]	0.502[b]	0.966[b]	0.933[b]
Diabetes by Hypercholesterolemia	0.477	0.965	0.930
Stroke by Heart Attack[a]	0.711[b]	0.948[b]	0.955[b]
Stroke by Fatigue	0.220	0.884	0.938
Heart attack by Infarct[a]	0.569[b]	0.989[b]	0.966[b]
Heart attack by Bruise	0.529	0.821	0.889
Multiple Sclerosis by Myasthenia Gravis[a]	0.761[b]	0.944	0.971
Multiple Sclerosis by Diabetes	0.522	0.993[b]	0.979[b]

[a]More similar model reuse cases.

[b]Performance metrics of better reusable models.

Ethical Approval and Informed Consent

Deidentified patient records were accessed through the Clinical Record Interactive Search at the Maudsley NIHR Biomedical Research Centre, South London, and Maudsley (SLaM) NHS Foundation Trust. This is a widely used clinical database with a robust data governance structure, which has received ethical approval for secondary analysis (Oxford REC 18/SC/0372).

Data Availability Statement

The clinical notes are not sharable in the public domain. However, interested researchers can apply for research access through https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/. The natural language processing tool, models, and code of this work are available at https://github.com/CogStack/CogStack-SemEHR.

Discussion

Principal Results

Automated extraction methods (as surveyed recently by Ford and et al [33]), many of which are freely available and open source, have been intensively investigated in mining free-text medical records [10,34-36]. To provide guidance in the efficient reuse of pretrained NLP models, we have proposed an approach that can automatically (1) identify easy cases in a new task for the reused model, on which it can achieve good performance with high confidence and (2) classify the remainder of the cases, so that the validation or retraining on them can be conducted much more efficiently, compared to adapting the model on all cases. Specifically, in four phenotype-mention identification tasks, we have shown that 50%-79% of all mentions are identifiably easy cases, for which our approach can choose the best model to reuse, achieving more than 93% accuracy. Furthermore, for those cases that need validation or retraining, our approach can provide guidance that can save 78%-85% of the validation/retraining effort. A distinct feature of this approach is that it requires no labelled data from new settings, which enables very efficient model adaptation, as shown in our evaluation: zero effort to obtain >93% accuracy among the majority (>63% in average) of the results.

Limitations

In this study, we did not evaluate the recall of adapted NLP models in new tasks. Although the models we chose can generally achieve very good recall for identifying physical conditions (96%-98%) within the SLaM records, investigating the transferability on recalls is an important aspect of NLP model adaptation.

The model reuse experiments were conducted on identifying new phenotypes on document sets that had not previously been seen by the NLP model. However, these documents were still part of the same (SLaM) EHR system. To fully test the generalizability of our approach will require evaluation of model reuse in a different EHR system, which will require a new set of access approvals as well as information governance approval for the sharing of embedding models between different hospitals.

We chose a phenotype embedding model to represent language patterns. One reason is that we have a limited number of manually annotated data items. The word embedding approach is unsupervised, and the word-level “semantics” learned from the whole corpus can help group similar words together in the vector space, so that it can help improve the phenotype-level clustering performances. However, thorough comparisons between different language pattern models are needed to reveal whether other approaches, in particular, simpler or less computing-intensive approaches can achieve similar or different performances.

In addition, implementation-wise, vector clustering is an important aspect of this approach. We have compared DBScan with k-nearest neighbors algorithm in our model, which revealed that DBScan could achieve better SP powers in most scenarios. Using a 64-bit Windows 10 server with 16 GB memory and 8 core central processing units (3.6 GHz), DBScan uses 200 MB memory and takes 0.038 seconds on about 300 data points on average of 100 executions. However, it is worth the in-depth comparisons between more clustering algorithms. In particular,

XSL•FO
RenderX

a larger dataset might be needed to compare the clustering performances on both computational aspect and SP powers.

Comparison With Prior Work

NLP model adaptation aims to adapt NLP models from a source domain (with abundant labelled data) to a target domain (with limited labelled data). This challenge has been extensively studied in the NLP community [37-41]. However, most existing approaches assume a single language model (eg, a probability distribution) from each domain. This limits the ability to identify and subsequently deal differently with data items with different language patterns. Such a limitation prevents fine-grained adaptations, such as the reuse or adaptation of one NLP model on those items for which it performs well, and the retraining of the same model or reuse of other models on those items for which the original NLP model performs poorly. In contrast, our work aimed to depict the language patterns (ie, different language models) of both source and target domains and subsequently provide actionable guidance on reusing models based on these fine-grained language patterns. Further, very few NLP model reuse studies have focused on free text in electronic medical records. To the best of our knowledge, this work is among the first to focus on model reuse for phenotype-mention identification tasks on real-world free-text electronic medical records.

Modelling language patterns have been investigated for different applications, such as the k-Signature approach [42] for identifying unique "signatures" of micro-message authors. This paper models language patterns for characterizing "landscape" of phenotype mentions. One main difference is that we do not know how many clusters (or "signatures") of language patterns exist in our scenario. Technically, we use phenotype embeddings to model such patterns and, particularly, utilize phenotype semantic similarities (based on ontology hierarchies) for reusing learned embeddings, when necessary.

Conclusions

Making fine-grained language patterns visible and comparable (in computable form) is the key to supporting "smart" NLP model adaptation. We have shown that the phenotype embedding-based approach proposed in this paper is an effective way to achieve this. However, our approach is just one way to model such fine-grained patterns. Investigating novel pattern representation models is an exciting research direction to enable automated NLP model adaptation and composition (ie, combining various models together) for efficiently mining free-text electronic medical records in new settings with maximum efficiency and minimal effort.

Acknowledgments

This research was funded by Medical Research Council/Health Data Research UK Grant (MR/S004149/1), Industrial Strategy Challenge Grant (MC_PC_18029), and the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

Conflicts of Interest

None declared.

Multimedia Appendix 1
User interface and model performances of phenotype natural language processing models.
[DOCX File , 968 KB - medinform_v7i4e14782_app1.docx]

Multimedia Appendix 2
Proof of Theorem 1.
[DOCX File , 8 KB - medinform_v7i4e14782_app2.docx]

References

1. Kharrazi H, Anzaldi LJ, Hernandez L, Davison A, Boyd CM, Leff B, et al. The Value of Unstructured Electronic Health Record Data in Geriatric Syndrome Case Identification. J Am Geriatr Soc 2018 Aug;66(8):1499-1507. [doi: 10.1111/jgs.15411] [Medline: 29972595]
2. Perera G, Broadbent M, Callard F, Chang C, Downs J, Dutta R, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. BMJ Open 2016 Mar 01;6(3):e008721 [FREE Full text] [doi: 10.1136/bmjopen-2015-008721] [Medline: 26932138]
3. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, et al. Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. PLoS Comput Biol 2011 Aug 25;7(8):e1002141. [doi: 10.1371/journal.pcbi.1002141]

XSL•FO
RenderX

4. Wang Y, Ng K, Byrd R, Hu J, Ebadollahi S, Daar Z. Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records Internet. 2015 Presented at: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2015; Milano, Italy. [doi: 10.1109/embc.2015.7318907]
5. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc 2014 Sep;21(5):801-807 [FREE Full text] [doi: 10.1136/amiajnl-2013-001915] [Medline: 24384230]
6. Margulis AV, Fortuny J, Kaye JA, Calingaert B, Reynolds M, Plana E, et al. Value of Free-text Comments for Validating Cancer Cases Using Primary-care Data in the United Kingdom. Epidemiology 2018;29(5):e41-e42. [doi: 10.1097/ede.0000000000000856]
7. Bell J, Kilic C, Prabakaran R, Wang YY, Wilson R, Broadbent M, et al. Use of electronic health records in identifying drug and alcohol misuse among psychiatric in-patients. Psychiatrist 2018 Jan 02;37(1):15-20 [FREE Full text] [doi: 10.1192/pb.bp.111.038240]
8. Jackson MSc RG, Ball M, Patel R, Hayes RD, Dobson RJB, Stewart R. TextHunter--A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research. AMIA Annu Symp Proc 2014;2014:729-738 [FREE Full text] [Medline: 25954379]
9. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010 Sep 01;17(5):507-513. [doi: 10.1136/jamia.2009.001560]
10. Wu H, Toti G, Morley K, Ibrahim Z, Folarin A, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc 2018 May 01;25(5):530-537 [FREE Full text] [doi: 10.1093/jamia/ocx160] [Medline: 29361077]
11. Christoph J, Griebel L, Leb I, Engel I, Köpcke F, Toddenroth D, et al. Secure Secondary Use of Clinical Data with Cloud-based NLP Services. Methods Inf Med 2018 Jan 22;54(03):276-282. [doi: 10.3414/me13-01-0133]
12. Tablan V, Roberts I, Cunningham H, Bontcheva K. GATECloud.net: a platform for large-scale, open-source text processing on the cloud. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 2012 Dec 10;371(1983):20120071-20120071. [doi: 10.1098/rsta.2012.0071]
13. Chard K, Russell M, Lussier YA, Mendonça EA, Silverstein JC. A cloud-based approach to medical NLP. AMIA Annu Symp Proc 2011;2011:207-216 [FREE Full text] [Medline: 22195072]
14. Carroll R, Thompson W, Eyler A, Mandelin A, Cai T, Zink R, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc 2012 Jun;19(e1):e162-e169 [FREE Full text] [doi: 10.1136/amiajnl-2011-000583] [Medline: 22374935]
15. Harris ZS. Distributional Structure. WORD 2015 Dec 04;10(2-3):146-162. [doi: 10.1080/00437956.1954.11659520]
16. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Commun ACM 1975;18(11):613-620. [doi: 10.1145/361219.361220]
17. Brown PF, Desouza PV, Mercer RL, Pietra VJD, Lai JC. Class-based n-gram models of natural language. Computational Linguistics 1992;18:479.
18. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci 1990 Sep;41(6):391-407. [doi: 10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9]
19. Blei D, Ng A, Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research 2003;3:933-1022.
20. Hinton G. Carnegie-Mellon University. 1984. Distributed representations. URL: http://www.cs.toronto.edu/~hinton/absps/pdp3.pdf [accessed 2019-11-06]
21. Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. Journal of machine learning research 2003;3:1137-1155.
22. Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. 2008 Presented at: Proceedings of the 25th international conference on Machine learning; 2008; Helsinki, Finland p. 160-167.
23. Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: A deep learning approach. 2011 Presented at: The 28th international conference on machine learning; 2011; Bellevue, Washington p. 513-520.
24. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. 2013 Presented at: Neural Information Processing Systems (NIPS); 2013; Lake Tahoe, Nevada.
25. Gouws S, Bengio Y, Corrado G. Bilbowa: Fast bilingual distributed representations without word alignments. 2015 Presented at: The 32nd International Conference on Machine Learning; 2015; Lille, France.
26. Hill F, Cho K, Korhonen A. Learning Distributed Representations of Sentences from Unlabelled Data Internet. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016 Presented at: NAACL 2016; 2016; San Diego, California.
27. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K. Deep Contextualized Word Representations Internet. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018 Presented at: NAACL 2018; 2018; New Orleans, Louisiana p. 2227-2237.

XSL•FO
RenderX

28. McCann B, Bradbury J, Xiong C, Socher R. Learned in translation: Contextualized word vectors. In: Advances in Neural Information Processing Systems. 2017 Presented at: NIPS 2017; 2017; California p. 6294-6305.
29. Peters M, Ammar W, Bhagavatula C, Power R. Semi-supervised sequence tagging with bidirectional language models Internet. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017 Presented at: ACL 2017; 2017; Vancouver, Canada.
30. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
31. Schubert E, Sander J, Ester M, Kriegel HP, Xu X. DBSCAN Revisited, Revisited. ACM Trans Database Syst 2017 Aug 24;42(3):1-21. [doi: 10.1145/3068335]
32. Van Asch V. Macro-and micro-averaged evaluation measures. 2013. URL: https://pdfs.semanticscholar.org/1d10/6a2730801b6210a67f7622e4d192bb309303.pdf [accessed 2019-11-07]
33. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016 Sep 05;23(5):1007-1015 [FREE Full text] [doi: 10.1093/jamia/ocv180] [Medline: 26911811]
34. Wu Y, Denny JC, Trent Rosenbloom S, Miller RA, Giuse DA, Wang L, et al. A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD). J Am Med Inform Assoc 2017 Apr 01;24(e1):e79-e86. [doi: 10.1093/jamia/ocw109] [Medline: 27539197]
35. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo Clinic NLP System for Patient Smoking Status Identification. Journal of the American Medical Informatics Association 2008 Jan 01;15(1):25-28. [doi: 10.1197/jamia.m2437]
36. Albright D, Lanfranchi A, Fredriksen A, Styler WF, Warner C, Hwang JD, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. J Am Med Inform Assoc 2013 Sep 01;20(5):922-930 [FREE Full text] [doi: 10.1136/amiajnl-2012-001317] [Medline: 23355458]
37. Moriokal T, Tawara N, Ogawa T, Ogawa A, Iwata T, Kobayashi T. Language Model Domain Adaptation Via Recurrent Neural Networks with Domain-Shared and Domain-Specific Representations Internet. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. 2018 Presented at: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing; 2018; Calgary, Canada p. 6084-6088.
38. Samanta S, Das S. Unsupervised domain adaptation using eigenanalysis in kernel space for categorisation tasks Internet. IET Image Processing 2015;9(11):925-930. [doi: 10.1049/iet-ipr.2014.0754]
39. Xiao M, Guo Y. Domain Adaptation for Sequence Labeling Tasks with a Probabilistic Language Adaptation Model. 2013 Presented at: International Conference on Machine Learning 2013; 2013; Atlanta, Georgia p. 293-301.
40. Xu F, Yu J, Xia R. Instance-based Domain Adaptation via Multiclustering Logistic Approximation. IEEE Intell Syst 2018 Jan;33(1):78-88. [doi: 10.1109/mis.2018.012001555]
41. Jiang J, Zhai C. Instance weighting for domain adaptation in NLP. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Association for Computational Linguistics. 2007 Presented at: ACL 2007; 2007; Prague, Czech Republic p. 264-271.
42. Schwartz R, Tsur O, Rappoport A, Koppel M. Authorship Attribution of Micro-Messages. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013 Presented at: EMNLP 2013; 2013; Seattle, Washington p. 1880-1891.

Abbreviations

BOW: bag of words
EHR: electronic health record
EPS: epsilon
LSTM: long short-term memory
NLP: natural language processing
SLaM: South London and Maudsley NHS Foundation Trust
SP: separate power

XSL•FO
RenderX

Edited by G Eysenbach; submitted 22.05.19; peer-reviewed by V Vydiswaran, B Polepalli Ramesh; comments to author 03.10.19; revised version received 08.10.19; accepted 22.10.19; published 17.12.19.

Please cite as:
Wu H, Hodgson K, Dyson S, Morley KI, Ibrahim ZM, Iqbal E, Stewart R, Dobson RJB, Sudlow C
Efficient Reuse of Natural Language Processing Models for Phenotype-Mention Identification in Free-text Electronic Medical Records: A Phenotype Embedding Approach
JMIR Med Inform 2019;7(4):e14782
URL: http://medinform.jmir.org/2019/4/e14782/
doi:10.2196/14782
PMID:31845899

©Honghan Wu, Karen Hodgson, Sue Dyson, Katherine I Morley, Zina M Ibrahim, Ehtesham Iqbal, Robert Stewart, Richard JB Dobson, Cathie Sudlow. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 17.12.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Building a Semantic Health Data Warehouse in the Context of Clinical Trials: Development and Usability Study

Romain Lelong[1,2], PhD; Lina F Soualmia[1,2,3], PhD; Julien Grosjean[1,3], PhD; Mehdi Taalba[1], MD; Stéfan J Darmoni[1,3], MD, PhD

[1]Department of Biomedical Informatics, Rouen University Hospital, Rouen, France

[2]LITIS EA 4108, TIBS, Normandy University, Rouen, France

[3]LIMICS U1142, Inserm, Sorbonne University, Paris, France

Corresponding Author:
Romain Lelong, PhD
Department of Biomedical Informatics
Rouen University Hospital
D2IM, Porte 21, Cour Leschevin
Hôpital Charles-Nicolle, 1 Rue de Germont
Rouen, 76000
France
Phone: 33 232 888 898
Email: romain.lelong@gmail.com

Abstract

Background: The huge amount of clinical, administrative, and demographic data recorded and maintained by hospitals can be consistently aggregated into health data warehouses with a uniform data model. In 2017, Rouen University Hospital (RUH) initiated the design of a semantic health data warehouse enabling both semantic description and retrieval of health information.

Objective: This study aimed to present a proof of concept of this semantic health data warehouse, based on the data of 250,000 patients from RUH, and to assess its ability to assist health professionals in prescreening eligible patients in a clinical trials context.

Methods: The semantic health data warehouse relies on 3 distinct semantic layers: (1) a terminology and ontology portal, (2) a semantic annotator, and (3) a semantic search engine and NoSQL (not only structured query language) layer to enhance data access performances. The system adopts an entity-centered vision that provides generic search capabilities able to express data requirements in terms of the whole set of interconnected conceptual entities that compose health information.

Results: We assessed the ability of the system to assist the search for 95 inclusion and exclusion criteria originating from 5 randomly chosen clinical trials from RUH. The system succeeded in fully automating 39% (29/74) of the criteria and was efficiently used as a prescreening tool for 73% (54/74) of them. Furthermore, the targeted sources of information and the search engine–related or data-related limitations that could explain the results for each criterion were also observed.

Conclusions: The entity-centered vision contrasts with the usual patient-centered vision adopted by existing systems. It enables more genericity in the information retrieval process. It also allows to fully exploit the semantic description of health information. Despite their semantic annotation, searching within clinical narratives remained the major challenge of the system. A finer annotation of the clinical texts and the addition of specific functionalities would significantly improve the results. The semantic aspect of the system combined with its generic entity-centered vision enables the processing of a large range of clinical questions. However, an important part of health information remains in clinical narratives, and we are currently investigating novel approaches (deep learning) to enhance the semantic annotation of those unstructured data.

(JMIR Med Inform 2019;7(4):e13917) doi:10.2196/13917

KEYWORDS

data warehousing; search engine; semantics; clinical trial; patient selection

XSL•FO
RenderX

Introduction

Background and Significance

Hospitals maintain important health data that can be used in various contexts: first and foremost, clinical care and then data reusability, clinical decision support systems [1], clinical research and cohort selection [2], education [3,4], and indicators. However, the exploitation of these data remains difficult for several reasons. First, the data are produced and maintained by different systems and health professionals and are consequently spread over multiple sources and even across multiple establishments. Second, the significant amount of data generated results in problematic management of data both in terms of data storage capabilities and data access performances. Health data can synthetically and legitimately be described as *big data*. For instance, according to research [5,6], in the United States, the health care system alone reached 150 exabytes (1.5×10^{20} bytes) in 2011 and will reach the yottabyte scale (10^{24} bytes) in the near future. Moreover, the health data produced are of different nature; some data are natively structured (eg, diagnosis-related group [DRG] codings and laboratory tests results), but an important part of medical information remains in unstructured free-text clinical narratives (CNs; eg, admission notes, history and physical reports, discharge summaries, radiology reports, and pathology reports) [7]. This unstructured information is particularly relevant in the context of cohort selection tasks. However, in the study by Raghavan et al [8], the authors found that not only unstructured data were essential to resolve between 59% and 77% of some clinical trials criteria but also that combining the use of structured and unstructured data enabled leverage of patient recruitment. To process unstructured data, the main approaches rely on natural language processing (NLP) methods [9,10]. The background knowledge, as represented in terminologies and ontologies (T&Os; that describe the domain), plays a crucial role in any clinical NLP task [11]. A common approach to information retrieval (IR) in clinical unstructured text outside the basic full-text search comprises partially restructuring the original texts using semantic annotators (eg, MetaMap [12]) that map words or expressions to concepts from domain knowledge databases.

Consistently aggregating all these scattered, big, complex, and diversely structured data is, in fact, the role of health data warehouses (HDWs). An HDW is defined as a grouping of data from diverse sources accessible by a single data management system [13]. This kind of data repository centralizes clinical, demographic, and administrative data within a uniform and consistent data model. Many HDWs have been proposed worldwide. From a holistic point of view, the majority of these solutions provide aggregated data mainly focusing on patient data as a result. Furthermore, they do not necessarily allow the full and independent visualization and retrieval of the different atomic entities conceptually composing the whole scope of clinical information (eg, Stanford Translational Research Integrated Database Environment or STRIDE [14] and Data Warehouse for Translational Research or DW4TR [15]). This is, nevertheless, particularly important in an IR context, as potential clinical questions and inquiries from health professionals are formulated in terms of their vision of the conceptual organization of data that derive from the actual patient management process. The Enterprise Data Trust [16] relies heavily on industrial solutions to cope with the huge amount of data. Many solutions also implement generic frameworks, such as Informatics for Integrating Biology and the Bedside (i2b2) database. This, however, implies concessions to conciliate the original conceptual representation of data with the data model required by the framework (eg, The European Hospital George Pompidou HDW [17]). Furthermore, many standardized controlled vocabularies used to semantically describe health information do not always provide access to concepts in French, and access to the data through these T&Os is not always provided for the whole set of data notably as far as the unstructured data are concerned (eg, Electronic Medical Record Search Engine or EMERSE [18] and STRIDE [14]).

In this context, in 2017, the Biomedical Informatics and Information Department of Rouen University Hospital (RUH) initiated the conception and development of a semantic HDW (SHDW). The SHDW functionally relies on 3 independent semantic layers: layer 1—the cross-terminological health T&O portal (HeTOP) [19] that provides the background knowledge necessary to semantically describe the health data; layer 2—a semantic annotator, the extracting concepts from multiple terminologies (ECMT) [20,21], that enables the annotation of unstructured data; and layer 3, the semantic search engine (SSE) [22-24] and a Web application interface semantic access to health information, ASIS, that enable access and retrieval of health data through different conceptual entities composing health information. A generic entity-attribute-value (EAV) data model and a not only structured query language (NoSQL) layer (layer 0) enable data structuring while preserving the original conceptual data model.

This study aimed to present a proof of concept (POC) of this SHDW based on the data of 250,000 patients from RUH and to assess its ability to assist health professionals in prescreening eligible patients in a clinical trial context. Since November 2018, this POC has integrated all the data of 1.8 million patients from RUH.

Related Studies

Clinical data warehousing manages health data from hospitals and is a well-addressed research field. Few generic frameworks and components exist. i2b2 [25,26] is a data mart used in >200 hospitals worldwide. Initiated within the Massachusetts General Hospital in 2004, i2b2 was developed by the Harvard Medical School and is funded by the National Institutes of Health. It enables the integration of clinical and genomic data into an EAV model known as the star schema. i2b2 enables the retrieval of patients' data using graphically built queries and querying of free-texts and coded information. Another example of a distributed solution is the Observational Medical Outcomes Partnership Common Data Model [27]. This EAV model tends to standardize data from HDW at structure and representation levels (ie, terminologies and vocabularies).

In France, a few open-source solutions exist, such as Dr Warehouse [28]—the CN-oriented data warehouse of Necker Children's Hospital, but 2 solutions really stand out from the

XSL•FO
RenderX

others: the ConSoRe system [29] used in some French Oncology Hospitals and the query engine Biomedical data warehouse of the hospital whose French acronym is eHOP [30,31] that is being deployed in 6 University Hospitals in Western France.

Owing to the specificity of the data and their private and sensitive aspect, HDWs are specific systems that are used locally in Hospital Information Systems (HISs) rather than distributed and ready-to-use solutions, and many specific HDWs have been developed worldwide in addition to the previously cited generic solutions.

The STRIDE (United States) [14] project focuses on a clinical data warehouse supporting clinical and translational research. It was initiated in 2003 at Stanford University when the functionalities of i2b2 and CAncer Biomedical Informatics Grid were not considered optimal. An Oracle database and an EAV data model derived from the Health Level 7 Reference Information Model (RIM) standard are used for data storage and representation. Several (mainly English) standardized terminologies are used to represent important biomedical concepts and their relationships (eg, Systematized Nomenclature Of MEDicine–Clinical Terms or SNOMED-CT, RxNorm, 9th revision of the International Classification of Diseases or ICD-9–Clinical Modification, and current procedural terminology). STRIDE provides hierarchical concept-based retrieval as far as structured data are concerned and provides full-text search access to more than 6 million CNs. The system is based on an n-tiered architecture and the querying of the data is distributed along several client applications whose scope targets patient cohort selection, cohort chart review, clinical data extraction, research data management, and specimen data management. The querying is done graphically using drag and drop interface-based components and returns aggregated data as a result without exposing individual patient data.

EMERSE (Michigan, United States) [18] is an electronic health record–oriented system exclusively providing full-text search capabilities into free-text clinical notes.

The Windber Research Institute (United States) developed the DW4TR [15] system to support multiple translational research projects through highly structured medical information represented in 3 dimensions (namely, clinical data, molecular data, and temporal information). Data are collected into an Oracle Relational DataBase Management System (RDBMS) with an EAV data model and are subsequently hosted in an extensible data model that organizes it into a structure of hierarchical modules inherited from especially developed ontologies. It provides 2 graphical querying interfaces designed to provide aggregated data dedicated to data analysis (eg, mean, standard deviation, counts, categorical data, and chronological view).

The Enterprise Data Trust [16] is an industrial HDW initiated in 2005 at the Mayo Clinic (50,000 employees, Rochester, Minnesota, United States). It collects patient care, education, research, and administrative data to support IR, business intelligence, and high-level decision making. The Enterprise Data Trust strongly relies on industrial technologies (eg, InfoSphere Information Server—International Business Machines; iSight and iGuard—Teleran; BusinessObjects—Systems Applications and Products; and PowerDesigner—Sybase) and enables integration and exploitation of important volumes of data (eg, more than 7 million unique patients, 64 million diagnoses, and 268 million test results). The architecture and functionalities of the Enterprise Data Trust rely on legacy technical components and long-standing governance works on data and metadata management, data modeling, and standardized vocabularies. Those initiatives provide the HDW with a reliable organization of information on patient, genomic, and research data as well as querying capabilities for cohort selection and aggregate retrieval.

In 2008, the European Hospital George Pompidou (Paris, France) initiated an HDW [17] based on the i2b2 framework. It is strongly integrated in the clinical information system (IS) of the hospital that relies on several industrial solutions (eg, OneCall—McKesson; Act management, computerized physician order entry—Medasys; and integration platform—Thales). The core HDW infrastructure relies on an Oracle database for storage (1.2 million patients and 1 million stays) and the i2b2 framework for data representation. Several client applications are connected to the system to provide technical access to the data but mainly use i2b2 client as far as researchers are concerned. The SMart Eye DATabase (SMEYEDAT) [32] is an ophthalmologic-specialized HDW developed at the University Eye Hospital in Munich in Germany. SMEYEDAT is based on a Microsoft structured query language (SQL) database updated daily from the HIS and uses a star-like patient-centered data model for data representation. The QlikView (QlikTech) [33] tool was implemented as an analytic tool to visualize and explore patient data. This interface enables patient selection using criteria and views specific to the domain.

Methods

Overall, the first prerequisite pertaining to the design of an HDW-based system is the extraction of data from the HIS. This can be achieved in 2 ways: by (1) setting up a data stream from the production environment (or a replicated database) to the HDW data storage component, or (2) using extract-transform-load (ETL) scripts. As far as the SHDW is concerned, ETL scripts are used. The following section describes the targeted sources of data of the HIS of RUH.

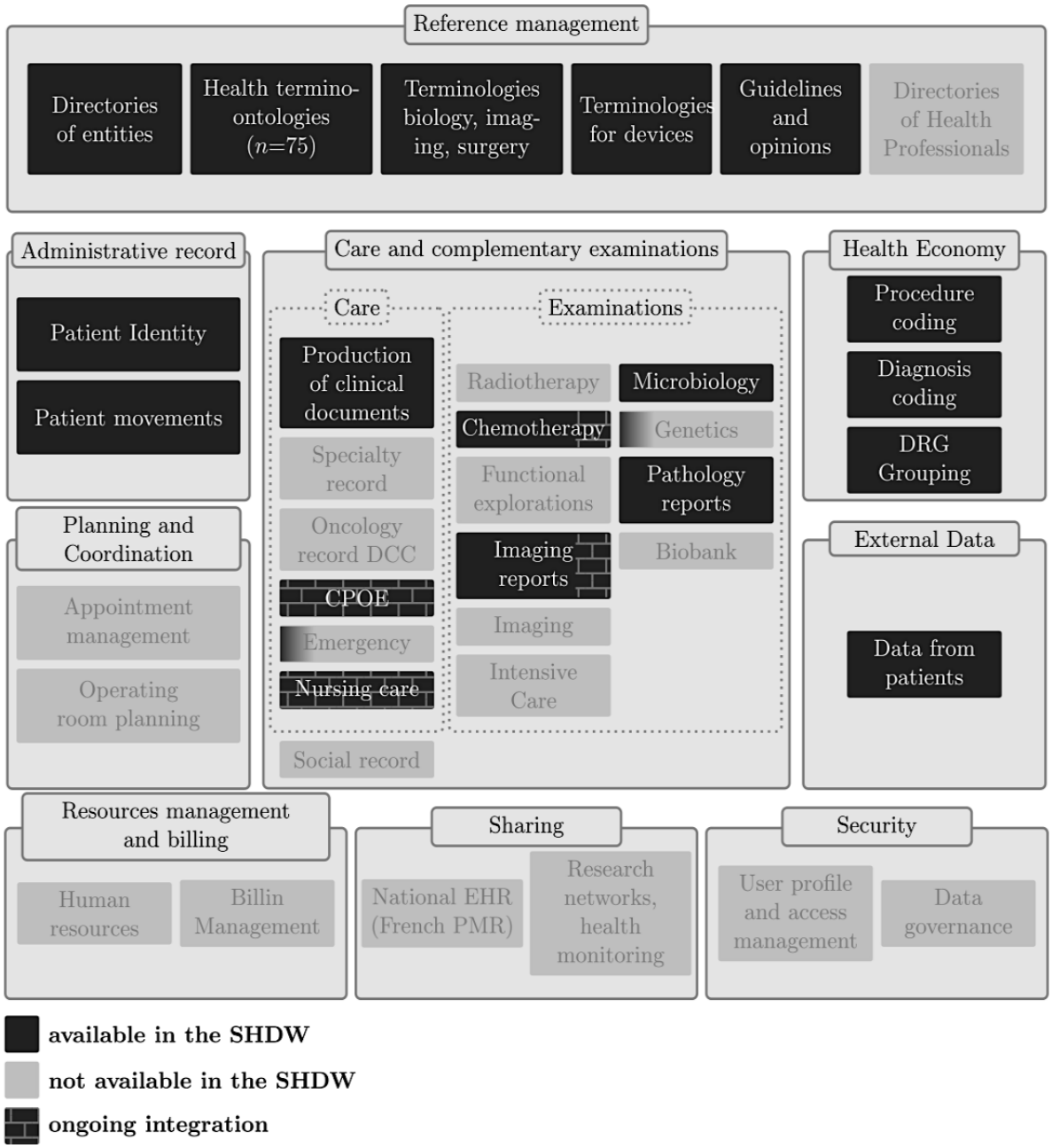
Data Sources

Since 1992, RUH has collected and maintained patient identity (eg, name, date of birth, and gender), clinical (eg, biological test results, medical procedures, visit records, letters, and discharge summaries), administrative, and less frequently, omics data [34]. The data are produced by different subsystems and applications of the IS of RUH. A subsystem called CPage Dossier Patient partially aggregates some of this important data such as laboratory results, DRGs, procedures, and clinical documents. Other data remain in other subsystems that have to be accessed separately. Overall, RUH maintains the data of 1.8 million patients that represent approximately 14.4 million visits, 11.9 million clinical documents (free-texts recorded since 2000), and 107 million single laboratory tests (eg, Sodium and Potassium being considered as 2 distinct tests; recorded since

XSL•FO
RenderX

2004). Since November 2018, the SHDW POC presented in this study includes the whole set of data. However, this study is based on a randomly chosen subset of data from 207,357 patients, 1.7 million visits, 671,442 clinical documents, and 14.2 million single laboratory tests. ETL scripts are used to incorporate data from the production environment repositories into an Oracle database. Figure 1 summarizes included data according to their specific domain (ie, reference management, administrative record, care, examinations, health economy, planning and coordination, external data, resource management and billing, sharing, and security). Data already included in the semantic health data warehouse (SHDW) are represented by a dark gray opaque background, whereas a light gray background indicates that data are neither included nor planned to be included in the short or medium term. Background partially or totally covered with bricks corresponds to data for which inclusion is in progress or is planned in the short term or medium term.

Figure 1. Functional coverage of the semantic health data warehouse in terms of data according to each domain. SHDW: semantic health data warehouse; CPOE: computerized physician order entry; DCC: French cancer communication file; PMR: personal medical record; DRG: diagnosis-related group; EHR: electronic health record.



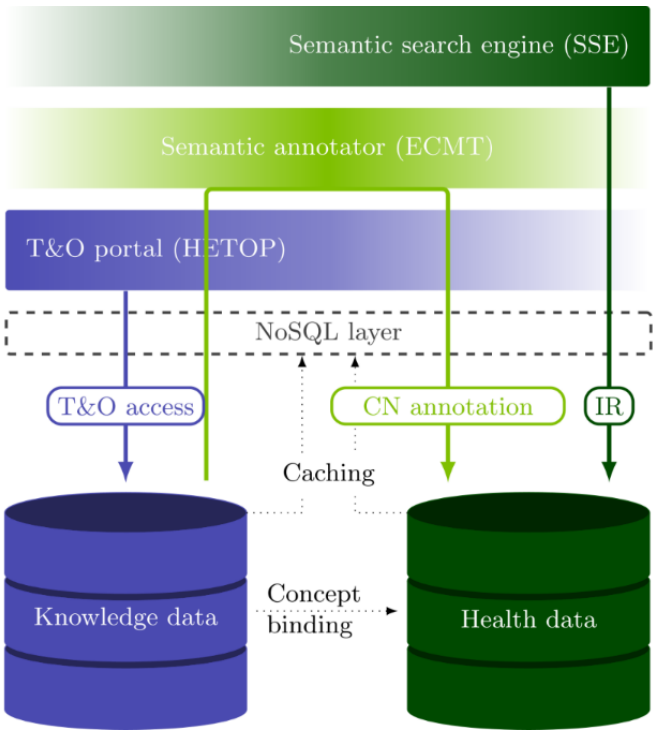
The SHDW currently focuses on clinical data and, more broadly, on health data according to a patient-centered strategy. In addition to the structured patient data, the different data pertaining to multiple admissions and events at RUH are collected (eg, diagnoses, biology, procedures, and movements). The reference-controlled vocabularies (ie, reference management domain) necessary to the understanding of those data are notably widely collected and maintained. In contrast, pure management and administrative data, such as appointment and planning data, billing data, and data governance, are not likely to be included in the short term. All those data are integrated into a modular architecture that is described in the following section.

XSL•FO
RenderX

Overall Architecture of the Semantic Health Database Warehouse

Much health information remains in CNs [7]. The 11.9 million clinical documents in French of RUH consequently play a strategic role in the context of the SHDW. Since its creation in 1995, our research team has strongly investigated French IR research domains through T&Os (and more broadly knowledge organization systems or KOSs), which has led to the development of several search tools mostly dedicated to IR from documentary and bibliographical resources [22,35]. However, the complexity of the clinical data and, more broadly, of SHDWs as a whole required the pooling of several of these acquired skills and tools. The SHDW enables the semantic retrieval of health data in French based on several T&Os and consequently relies on 2 datasets: a domain knowledge database and a health database maintaining clinical and patient data. The functionalities of the SHDW are ensured by the collaboration of 3 distinct layers, where each layer consumes data from the above layers (see Figure 2): the (1) cross-terminological HeTOP [19], (2) semantic annotator ECMT [20,21,36], and (3) SSE [22-24].

Figure 2. Functional architecture of the semantic health data warehouse that provides semantic information retrieval (IR) functionalities form clinical data. The 2 data repositories, knowledge data and health data, respectively, maintain the reference knowledge organization systems and the health data pertaining to the semantic health data warehouse. These data are accessed through a not only structured query language (NoSQL) layer by the 3 distinct components: the cross-terminological health terminology and ontology portal (HeTOP), the semantic annotator extracting concepts from multiple terminologies (ECMT), and the semantic search engine (SSE), each operating on a different range of data. CN: clinical narrative; T&O: terminology and ontology.



The HeTOP provides access to domain knowledge data. The ECMT matches words and expressions in natural language to domain knowledge concepts included in the HeTOP. In fact, ECMT enables the extraction of semantic information from unstructured data. Its functional scope consequently lies between domain knowledge and clinical data. Together, the 2 components HeTOP and ECMT serve as a base for the semantic description of the clinical information in a computer-processable form. In contrast, the SSE and the coupled Web application, are dedicated to IR tasks on health data by using this extracted semantic description.

Considering the amount of data, access to health data and domain knowledge data is made through an NoSQL layer [22] based on the Infinispan solution, an in-memory data grid (IMDG), on a server with 192 cores and 1 TB (ie, 10^{12} bytes) of random access memory (RAM) allowing vertical scaling.

Each of these layers is functionally and technically detailed below.

Semantic Representation

This section describes data and the methods for data storage and modeling and presents ECMT that enables the link between knowledge data and actual clinical data.

Domain Knowledge Data

The HeTOP provides cross-lingual access to concepts originating from 75 T&Os. A set of 2,639,620 concepts and 10,735,905 terms are available mainly in English and French. However, 32 languages are available overall. Some of the T&Os have been partially or totally translated into French (eg, SNOMED 3.5—52.3%; Medical Subject Heading descriptors—100%; National Cancer Institute Thesaurus—53.35%; Online Mendelian Inheritance in Man—79.67%; Human Phenotype Ontology—72.19%; and RadLex—22.1%). More broadly, 50% of the 2.64 million concepts accessible through the HeTOP are provided in French, and 19.1% of the 10.74 million terms have a French translation. T&Os from the HeTOP come with their original sets of

XSL•FO
RenderX

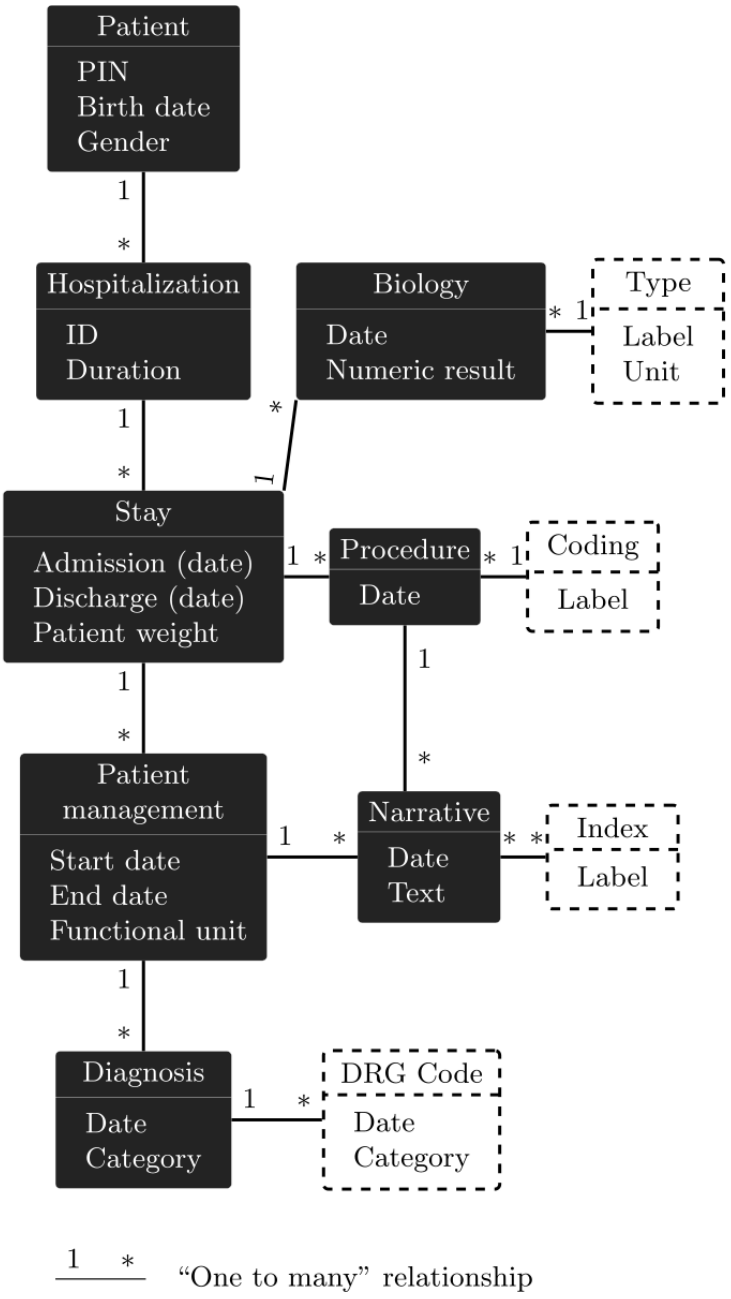
hierarchies and semantic relationships but also with additional cross-terminological exact, broader to narrower, and narrower to broader mappings performed manually or supervised by our health professionals at RUH.

As a primary use, the different concepts are bound to the different clinical entities (eg, procedure and DRG codings, and CN annotations), thus allowing a semantic description of the clinical information to be obtained. This allows both refining and broadening of the IR tasks by exploiting the underlying semantic network formed by the concepts (ie, by controlling the granularity and the depth with which this semantic network should be browsed in search processes).

Health Data Model

Health data are stored in a PostgreSQL [37] relational database. A generic and very adaptable physical EAV data model [34,38] is used to integrate the data. This model structures the information in terms of objects, attributes, and relationships and thus defines an underlying entity-association modeling of the data. It enables the preservation of any original conceptual organization of the information without altering the physical data model and consequently maintains the desired vision of the data at conceptual level. A partial and simplified representation including the main entities and a limited number of relationships and attributes of the conceptual data model used for this study is shown in Figure 3. This model is used on a daily basis to satisfy the information needs of the different health professionals of RUH.

Figure 3. Partial Conceptual Model of the semantic health data warehouse represented as a directed and attributed graph. Entities corresponding to elements from terminologies and ontologies are represented with dashed outlines. DRG: diagnosis-related group; PIN: personal identification number.



XSL•FO
RenderX

Semantic Annotator

The semantic annotator ECMT [20,21,36] matches the natural language words and expressions to the domain knowledge concepts included in the HeTOP. Technically, the ECMT relies on the bag-of-word method for concept matching but also provides pattern-matching functionalities, in particular, to deal with negation and contextual information such as numerical values in CNs. Functionally, the ECMT is used in query building processes to match user inputs to accurate sets of concepts but plays a major role in CN indexing.

Semantic Retrieval

Access to the data is allowed by a NoSQL layer before processing by the SSE.

Not Only Structured Query Language Layer

Owing to the considerable amount of health data that need to be retrieved and the well-known limitations of the RDBMSs in terms of scalability, a NoSQL layer was designed to interface access to all the data and improve data access performances. This layer is based on the IMDG Infinispan [39,40]. It is a Java NoSQL solution that uses key-value hash tables as storage structure, which allows efficient recovery of unitary data via the associated keys. Moreover, the hash tables are stored in memory and not on disk, which leverages access times.

The NoSQL layer was conceived in a generic way to mirror the EAV data model used to structure health data (ie, Java object used as values in hash tables mimics the objects and relationships of the relational databases). This generic NoSQL layer consequently preserves the conceptual data model of health and knowledge data implicitly drawn by the EAV data model. A more detailed description of this layer and an overview of performance gain compared with relational RDBMSs are presented in the study by Lelong et al [22].

Semantic Search Engine

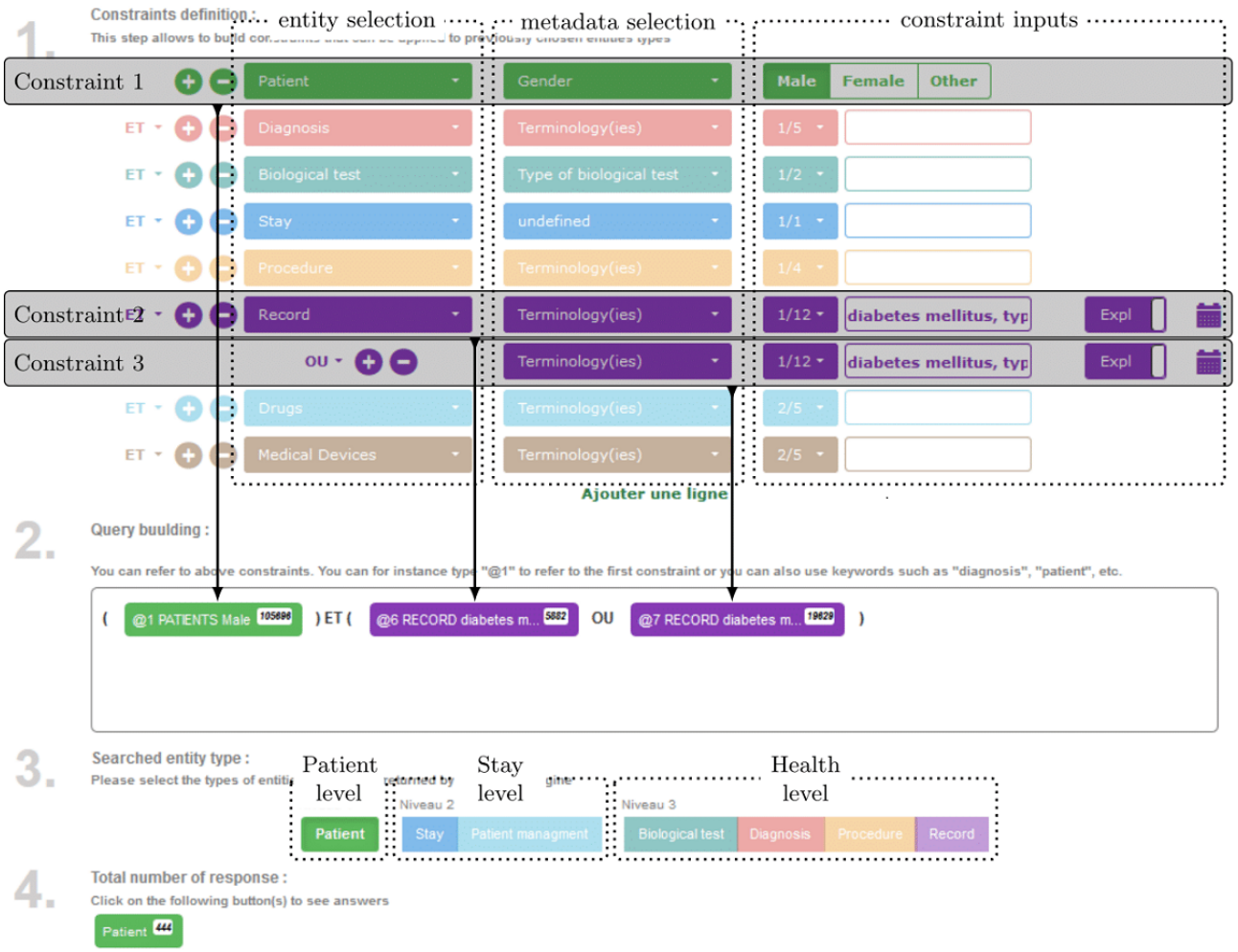
The main purpose of the SSE is to deal with the multiplicity and the diversity of conceptual entities inherent to clinical and patient data (eg, patients, stays, CNs, diagnosis, and biological tests). Overall, the entire set of data originating from the SHDW can be seen as a comprehensive oriented attributed graph that can be queried by the SSE. The SSE was designed to concentrate on semantic retrieval by allowing navigation through the semantic networks, not only included in the T&Os but also those representing clinical data conceptual entities. From a more clinically coherent point of view, the data of the SHDW can be organized in 4 levels: (1) patient level corresponding to patient identity information, (2) hospital level defining the sources of information (this level is currently not implemented as all the data originate from RUH), (3) visit level that defines much organizational and administrative information about the health care process, and (4) health level enabling group medical procedures and biological tests [24]. As an HDW can be used in various contexts (eg, health care, health research, and secondary use of health data), access and search capabilities of the full scope of those types of information must be provided. Technically, the SSE is a Boolean and entity-oriented search engine. It enables the retrieval and display of data at any of the previous clinical levels. As mentioned in section above, the NoSQL key-value store used to interface data does not provide proper querying solutions. The SSE consequently relies on a specific query language based on formal grammar. It enables the expression of queries targeting any of the different conceptual entities selected through constraints focusing on attribute values and other linked entities [24]. The SSE is used through a Web application that enables the querying of clinical data using forms and string-based queries. This application is described below.

The Semantic Access to a Health Information Web Application: ASIS

The SSE provides a powerful means to select data using textual logical queries. To bypass the complexity of the query language syntax, we designed a user-friendly Web application known as ASIS. It enables the retrieval of clinical data by means of a form that generates an SSE-processable logic-based query. The clinical data selection process is divided into 4 numbered steps clearly identifiable on the graphical interface. Step 1 comprises building a set of constraints related to any desired entity of interest as patient, diagnosis (DRG), biological tests, stays, procedures, records (CNs), drugs, and medical devices (see Figure 4). Constraints are built via (1) the choice of the entity of interest; (2) the choice of the targeted metadata of this entity as date of birth (patient), gender (patient), type of biological test (biological test), date (eg, procedures, biological tests, and stays), and coding (eg, diagnosis, records, and procedures); and finally, (3) the entry of the inputs corresponding to the chosen entity and metadata as male/female for the gender metadata of a patient constraint and the desired numeric value for the biological test constraint. To facilitate the reading of the interface, each type of entity is represented using a specific color (eg, green for patient, red for diagnoses, green-cyan for biology, and blue for stays). As the SHDW was conceived to focus on semantics, many metadata inputs concentrate on selecting T&Os and concepts by the user from fields autocompleted to facilitate the selection. For instance, constraints 2 and 3 enable retrieval of CNs indexed with the different concepts referring to type 1 and 2 diabetes (Figure 4). Step 2 comprises aggregating the constraints defined in step 1 into a Boolean query. In this form, constraints are represented as colored buttons showing their IDs, a short description of them, and the numbers of results of the subqueries corresponding to them. A click on a constraint button enables the visualization of the partial results corresponding to the constraints. The step 2 subform editable area enables the composition of the query using parentheses, Boolean operators (ie, AND, OR, and NOT), and the defined constraints that can be selected using an autocompletion feature. Nevertheless, the step 1 subform enables the predefinition of a basic Boolean query skull that is on-the-fly reported in step 2 and can be later manually modified or left untouched in step 2. Step 3 comprises choosing the desired output entity type classified according to the 3 clinical information levels: patient, visit (stay), and health levels. The choice of an entity type generates a button similar to constraint buttons in step 4.

XSL•FO
RenderX

Figure 4. The interface of the semantic access to health information, ASIS, Web application, and its 4 steps: (1) definition of constraints, (2) composition of a Boolean query from atomic constraint defined in step 1, (3) selection of the desired output entity according to its clinical coherent level, and (4) visualization of the results.



Evaluation Methodology

A total of 5 clinical trials originating from the RUH, covering a total of 95 criteria (36 inclusions/59 exclusions), were randomly selected without previous information on the content of those clinical trials. The selection was done from a pool of 57 clinical trials initiated between 2005 and 2018 and that were either still recruiting or already completed. The 5 selected clinical trials were initiated in 2012, 2014, 2015, 2016, and 2018 with various medical objectives. Of them, 3 were still in the recruitment phase. The ability of the system to automate patient prescreening was then assessed on each of those criteria, taken independently from both the originating clinical trial and the overall context of the clinical trials. For each criterion, a search strategy was designed. Each of them required the collaboration of a medical doctor (to clinically interpret the criteria and identify the different sources of information to target) and a computer engineer to master the ASIS tool querying process. The search for a single criterion can be done through multiple search directives (ie, ASIS constraints) targeting different sources of information (ie, entities). Those search directives are then aggregated into a single search strategy (ie, a global query) by combining the different search directives using Boolean operations and relational links between the entities corresponding to each search directive. The different constraints that could reduce the accuracy of each search directive were also investigated. In this study, 3 characteristics are finally considered and linked to each other to more precisely identify the different capabilities and limitations of the system: (1) the global support level of the criteria by the system, (2) the targeted source of information, and (3) the obstacle and barriers that tend to lower the effectiveness of the search.

Each of the criteria was, therefore, classified into 6 levels of global support by the system.

Fully Supported Level

This represents the criteria that can be fully automated by the system with a search strategy that retrieves all and only the resources (without false positives or false negatives) that fulfill the exact requirements of the criteria, for example, *18-year-old patients and patients with a neutrophil level below 1700/mm*^{3}.

Accurately Supported Level

This represents the criteria that are based on consistently recorded data in the IS and on reliable search. The result may, however, possibly include some irrelevant resources or miss relevant ones depending on the choices made in the elaboration process of the search strategy, mainly about the choice of

XSL•FO
RenderX

concepts to search and the exploitation of their semantic networks, for example, *patients with hepatitis B or active hepatitis C* and *patient with acute kidney failure*. In absolute terms, a relevant resource may also be ignored if the data have not been entered in a standard way in the IS.

Broadly Supported Level

This represents the criteria for which the search results in a lack of precision (ie, inclusion of irrelevant resources or absence of relevant ones). These criteria can only be reliably answered partially. This implies a broadened search of the core requirement of the criteria and a manual postfiltering of the result and/or supervision by a health professional to decide whether, or not, the retrieved resources effectively fit the criteria, for example, *patient with an evolving organic digestive and/or inflammatory pathology* and *patient with a badly regulated cardiac rhythm disturbance*.

Inaccurately Supported Level

This represents the criteria that cannot be searched precisely enough (both technically and in terms of data) to fulfill the core requirement of the criteria or systematically provide consistent results. For example, criterion *pregnant woman or breastfeeding mother* cannot be searched correctly as the information used to record pregnancy is only very rarely provided as structured data. Thus, the search for information relating to pregnancy and breastfeeding is, therefore, provided mainly in the reports and is, in addition, not systematically provided. This information is sometime provided in a roundabout way involving an effort of deduction or through hardly analyzable natural language expressions. It is, consequently, a difficult information to retrieve. Nevertheless, this information constitutes the essential part of the medical objective of the criterion. Another type of inaccuracy because of technical limitations of the system can also be observed with criterion *patient admitted for a stomach hemorrhage resulting in a favorable evolution without surgery during the hospitalization*. Although the first part of the criterion is searchable (ie, stomach hemorrhage), the second part of the criterion is not defined with metrics that can be easily interpreted in terms of a query and, more particularly, with regard to *favorable* evolution. In addition, extensive temporal functionalities would have been necessary, in particular, to exclude favorable evolution following surgery.

Nonsupported Level

This represents the criteria for which the system fails to properly select the relevant resources (ie, the medical doctor did not consider any of the first results as relevant to the criterion) and/or a search strategy is hardly feasible. For example, the criterion *patient with a regular consumption of licorice or derived substances* is not supported because the patients' diet is an information which is essentially absent from the IS. In the rare cases where the information appears in the unstructured data, this information is incomplete, unreliable, and technically difficult to identify/extract. Similarly, none of the attempts to define a research strategy to solve the criterion *abdominal pain presenting once a week during the last 3 months associated with 2 of the following criteria [...]* yielded consistent results. This criterion involves temporal considerations that are not currently within the scope of the IR system.

Not Applicable and Instruction Level

This represents criteria that either does not connect to the medical domain or that corresponds more to instructions than real requirements, for example, *patient participating in another clinical trial* and *contraception will be required during the treatment.*

A total of 6 types of source of information were identified: (P) Patient structured data as age and gender, (D) DRG data corresponding to structured diagnosis coded with the 10th revision of the ICD and related health problems, (S) stay data and other organizational structured data as medical units, (B) biological structured data, (N) CN unstructured data as full-text and/or automatic indexing including drug data, and (I) for information that is not within the scope of RUH IS.

Finally, the different obstacles or limitations that lower the effectiveness of the search were recorded for each atomic search directive and were distributed among 6 categories: (o) for search directives that are free of any obstacles, (d) for data obstacles corresponding to inconsistently provided or insufficiently accurate data from the IS, (s) for difficulties to perform an accurate search in CN or DRG data as complex information search, (t) for technical limitations of the system as chronological querying handling or search for quantitative values in CNs (partially implemented), (c) for subjective and/or generic criteria implying the interpretation or value judgment of a health professional, and (e) when it is necessary to meet the patient to complete the criteria.

The global support levels of criteria observed in this study are first detailed in section *Global Support of Criteria*. A 2-sided Wilcoxon signed-rank statistical test is used to examine the different levels of support of inclusion versus exclusion criteria. The 3 sets of scores detailed in this methodology section are then matched with each other in *Observed Sources of Information and Limitations* to objectify and identify the concrete abilities and limitation of the system.

Results

Global Support of Criteria

As a primary and holistic result, the support levels of the 36 inclusion and the 59 exclusion criteria from the 5 randomly selected clinical trials of RUH are shown in Table 1. The percentage of criteria for each of these levels was recorded.

According to the methodology used to classify criteria, 3 out of the 6 levels of support, *full*, *accurate*, and *broad* could be considered as contributing to cohort's prescreening. Taken together, the system was consequently able, at least partially, to automate the search for 15 out of 36 (15/36, 42%) of inclusion criteria versus 39 out of 59 (39/59, 66%) of exclusion criteria. Among the 5 clinical trials used in this study, the number of exclusion criteria exceeded the number of inclusion criteria by 20% on average (6 vs 16, 9 vs 13, 5 vs 4, 7 vs 13, and 9 vs 13 inclusion vs exclusion criteria, respectively).

XSL•FO
RenderX

Table 1. Number, percentage, and 95% confidence interval of the percentage of criteria for each support level and type (inclusion or exclusion).

Support level	Inclusion criteria		Exclusion criteria		Total	
	n (%)	95% CI	n (%)	95% CI	n (%)	95% CI
Full	6 (16.67)	(4.5-28.8)	5 (8.47)	(1.4-15.6)	11 (11.58)	(5.1-18.0)
Accurate	3 (8.33)	(0.0-17.4)	15 (25.42)	(14.3-36.5)	18 (18.95)	(11.1-26.8)
Broad	6 (16.67)	(4.5-28.8)	19 (32.20)	(20.3-44.1)	25 (26.32)	(17.5-35.2)
Inaccurate	4 (11.11)	(0.8-21.4)	6 (10.17)	(2.5-17.9)	10 (10.53)	(4.4-16.7)
None	3 (8.33)	(0.0-17.4)	7 (11.86)	(3.6-20.1)	10 (10.53)	(4.4-16.7)
Not applicable	14 (38.89)	(23.0-54.8)	7 (11.86)	(3.6-20.1)	21 (22.10)	(13.8-30.4)
Total	36 (100.00)	__[a]	59 (100.00)	—	95 (100.00)	—

[a]Not applicable.

A fairer and more reliable measure was also investigated. Of the 95 criteria, 21 (21/95, 22%) of this study were actually not applicable (N/A) criteria. This type of criteria is not in the scope of an HDW-based system and should consequently be set aside. Moreover, 14 of these 21 criteria (67%) were attributed to inclusion criteria. Excluding *N/A* criteria, the percentages of criteria for which the system was able to contribute increased to 68% for inclusions (15/22 criteria) and 75% for exclusions (39/52 criteria). When only considering support levels that did not imply postfiltering (ie, only *full* and *accurate*), 9 out of 22 (9/22, 41%) inclusion criteria could be answered compared with 20 out of 52 (20/52, 38%) exclusion criteria.

A 2-sided Wilcoxon signed-rank statistical test was used to compare the levels of support of inclusion versus exclusion criteria. To perform that test, a mean support score was calculated for each subset of inclusion or exclusion criteria of each clinical trial. The calculation of these means was made by assigning to each support level a score from 0 to 100. The test was not significant with a homogeneous distribution of the scores, but a trend was observed toward better support of inclusion criteria compared with exclusion criteria for distributions that weighted *full* criteria twice as much as the others. The mean support score of inclusion criteria was constantly greater than that of exclusion criteria for each clinical trial. The tests resulted in observed statistics T=15 with a *P* value equal to .06, which, even if slightly greater than the 5% significance level, suggested a better support of inclusion criteria.

Observed Sources of Information and Limitations

The results obtained in Table 1 should, nevertheless, be regarded more qualitatively than quantitatively as regard the 95% confidence intervals that show widths of 20.37% on average (15.08% when inclusion and exclusion criteria are taken together). To achieve that goal, both the targeted sources of information and the observed limitations for each support level of each criterion were investigated.

The support level of the criteria according to the combination of information sources required to search them are displayed in Table 2.

Table 2. Number of criteria of each support level according to the combination of sources necessary to search them. The sources of information are as follows: P: patient data, D: diagnoses-related group data, S: stay data, B: biological data, N: clinical narrative, and I: other information.

Support level	*P*	*S*	*B*	*D*	*N*	*I*	*S+N*	*P+N*	*D+B*	*B+N*	*D+N*	*N+I*	*P+S+D*	*S+D+N*	*D+N+I*
Full	4	0	7	0	0	0	0	0	0	0	0	0	0	0	0
Accurate	1	0	0	8	0	0	0	1	3	0	5	0	0	0	0
Broad	0	2	2	6	7	0	1	0	0	1	4	0	0	2	0
Inaccurate	0	0	0	0	5	0	0	0	0	1	1	0	1	1	1
None	0	0	0	0	2	3	0	0	0	0	3	1	0	1	0
Not applicable	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0
Total	5	2	9	14	14	24	1	1	3	2	13	1	1	4	1

Setting aside N/A criteria, 47 out of the 74 criteria (47/74, 64%) could be answered using a single search directive (ie, by exploiting a single source of information) against 27 criteria (36%) that required combined search directives. The calculation of the mean scores of level of support of these 2 groups of criteria resulted in scores between *accurate* and *broad*. Of single search directives, 23.40% versus 0% of combined search directives concerned fully supported criteria.

The different sources of information were not uniformly distributed. Patients (P) and stays (S) structured data were involved in the search for only 7 out of the 95 (7/95, 7%) and 8 out of the 95 (8/95, 8%) criteria, respectively. In contrast, the top 2 sources of information, CNs (N) and diagnoses (D), were involved in the search of 37 (37/95, 39%) and 36 (36/95, 38%) of the 95 criteria, respectively.

XSL•FO
RenderX

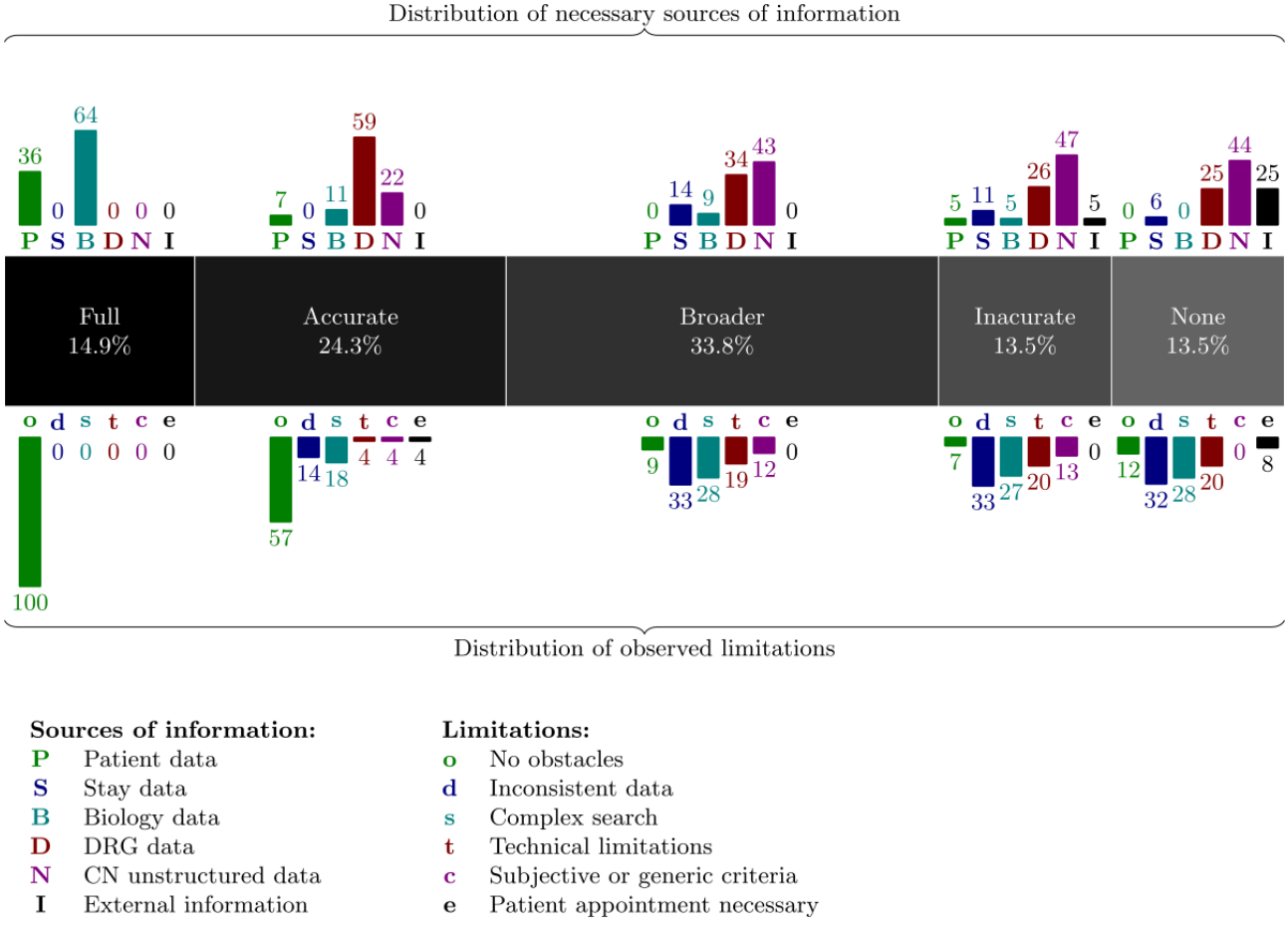
The percentages of involvement of sources of information and the percentages of involvement of observed limitations for each support level are presented in Figure 5.

Only continuously provided and fully structured data were used to answer fully supported criteria. The only sources of information used were patient structured data (P) and biological data (B). Fully supported criteria were consequently based on very precise characteristics not subject to errors or ambiguity and relying on numeric or symbolic data such as *female or male patient aged 18 to 75 years* and *patient with glycated hemoglobin ≤6.5% or ≥8%*.

Accurately supported criteria were mostly searched in DRG data (D). In practice, these criteria either rely on a single source of information (eg, *HIV-positive patient* and *type 2 diabetic subject*) or the combination of data consistently provided or properly coded as "known active hepatopathy, [...], transaminase and/or alkaline phosphatase levels twice the normal level of the laboratory" (DRGs and biological data) or *men aged 18 to 70 years or women aged 18 to 70 years in menopause* (patient data and CNs).

From a holistic point of view, we observed that DRGs (D) and CNs (N) were the 2 major sources of information used. As stated before, both were involved in the search strategy of approximately 38% of criteria (58% if taken together). The support of the criteria by the system decreased as the exploitation of CNs (N) took precedence over DRG data (D). In this study, the search within CNs (N) were performed through both full-text and semantic searching. As far as semantic searching is concerned, the 11,928,168 CNs (N) of RUH have been indexed using ECMT over 55 terminologies available in French from the HeTOP server. These CNs were first collected from the RUH IS in their original format (ie, as Microsoft Word files). The raw text was then extracted from these Word documents and stored in simple text files that were provided as input to the ECMT semantic annotator. After some performance optimizations of this annotator, the indexing of all the CNs could be completed in slightly less than 24 hours on a machine equipped with 1 TB of RAM and 144 cores. The insertion of the annotations in the database was done separately. Indexes allowing retrieval were also generated. The indexing process resulted in a total of 5,043,731,628 annotations. Some of the most redundant concepts were found clinically irrelevant, and a manual filtering process was applied based on the top 5000 most frequent medical concepts (eg, the 27 million annotations with "university hospital" were considered as irrelevant as the information was present elsewhere in the SHDW). A total of 2,087,784,055 annotations were retained after the filtering process. This set of semantic annotations served as a basis for the SSE in the semantic retrieval process.

Figure 5. The central gray band gives the percentage of criteria of each support level excluding not applicable criteria. The upper bars show, for each support level, the percentages of involvement of each source of information in the search of criteria. The lower bars show the distribution (in percentage) of the different obstacle categories identified as lowering the effectiveness of the search of criteria. CN: clinical narrative; DRG: diagnosis-related group.



XSL•FO
RenderX

The search accuracy obstacles (s) category represented 20.2% of all obstacles (32/158), and 84% (27/32) of these obstacles were attributed to CN search (N). However, the exploitation of the semantics (ie, synonyms, hierarchical, and semantic relationships) through the automatic indexing of CNs (N) by the ECMT and the ability of the SSE to combine multiple search directives (using Boolean operators) enabled a broad search support of 25 criteria out of 95 (25/95, 26.3%). Even when postfiltering was required, the system could be used effectively as a prescreening tool. For instance, the search for the criterion *patients with severe heart failure (including New York Heart Association or NYHA Classes III and IV)* was done through the search for *heart failure* in DRG data (D) and the search for *NYHA Classes III* and *IV* in CN data (N). Separately, it resulted in 11,880 diagnoses and 3311 CNs. The combination of both search directives into a single search enabled the prescreening of only 36 patients.

Data inconsistency (d) was also a major challenge, as 37 obstacles out of the 158 obstacles (37/158, 23.4%) were of that type and found across different sources of information including 10 of those (10/37, 27%) for DRGs (D) and 6 of those (6/37, 16%) for stays (S). Many data are sparsely recorded in the IS even outside CNs (eg, weight of the patient as structured data for each stay or S, diet plan in CNs, and hypersensitivity to substances in DRG data or D).

This lack of consistency of information tends to explain the focus on CNs (N) of inaccurately supported (Inaccurate level) and nonsupported (None level) criteria. In practice, these criteria suffer from the association of concurrent obstacles, often including a data consistency obstacle (d). For instance, both data inconsistency (d) and technical limitations (t) were found for the nonsupported criteria *regular consumption of alcohol exceeding 60 g per day*. Information on alcohol consumption was in fact not provided consistently in CNs (N), and technically, it would have required the following: (1) the extraction of a quantitative value from CNs and (2) the processing of this value as data (partially implemented). As another example, the criterion *patient with a creatinine clearance ≤50 ml/mn according to Cockroft formula* was inaccurately managed by searching instead for the biological tests of creatinine higher than 100 µmol/L. The criterion strongly relies on specific calculation functionalities not provided by the system and is based on sparsely provided data (eg, weight of the patient).

With regard to efficiency, the NoSQL layer used to access the data gave querying performances that were considered extremely satisfactory. On the basis of the data of 250,000 patients, each of the search directives used for this study took less than 2 seconds. As far as the POC integrating the entire patient dataset (1.8 million patients) is concerned, similar performances were observed except for some specific queries targeting and returning huge amounts of biological tests which exceeded 1 min.

Discussion

Principal Findings

To our knowledge, no formal evaluation of the criteria for clinical trial inclusion and exclusion has been performed using an SHDW. The system based on an SHDW presented in this study could be successfully used to fully automate 29 criteria out of the 74 non-N/A criteria (29/74, 39%). Moreover, with a limited postfiltering process, it could be efficiently used as a prescreening tool for 54 of those (54/74, 73%).

A trend was observed toward better support of inclusion criteria compared with exclusion criteria for distributions of scores that weighted *full* criteria twice as much as the others. However, with homogeneous distributions of scores, no conclusion could be made. A lower support of inclusion requirements tends to affect the ability of the system more to assist prescreening tasks. Manual exclusion of patients is usually a lighter task than manual inclusion (especially if the exclusion is made from a small set of patients who already meet the inclusion criteria). Furthermore, clinical trials usually rely on fewer inclusion than exclusion criteria (36 inclusion vs 59 exclusion criteria in this study), which often makes the inclusion requirements more critical prerequisites. Inclusion criteria are indeed used to target specific medical characteristics essential to clinical trials, whereas some exclusion criteria tend to be more generic. For example, one of the clinical trial used in this study included the specific inclusion criteria *type 2 diabetic subject* and *subject with a weight mass index (weight/height^2) greater than 27* and, in contrast, more generic inclusion criteria such as *patients with a severe medical or surgical history, in particular, endocrine history* or *patients treated with drugs interfering with the renin-angiotensin-aldosterone system*.

There are still many criteria (ie, 20/74 non-N/A criteria, 27%) that cannot be searched or can only be partially searched by the system. Several mishandled sources of information along with specific limitations of the system are apt to explain these results. DRG and CN data remain an important source of information for nonsupported or inaccurately supported criteria. Consistent and systematic recording of necessary information in the IS not always performed. Furthermore, this information often resides within the unstructured CNs form, which is often difficult to extract.

Technically, our system relies only on free solutions. It accesses the data through an IMDG NoSQL layer that offers very satisfactory performances with the data of 250,000 patients. Since November 2018, all the data of the 1.8 million patients from RUH have been integrated into the POC with relatively constant performance (ie, most of the queries tested in this paper are still under the 5-second threshold considered acceptable by health professionals).

Comparison to Prior Work

The general philosophy of the system relies on a generic representation of clinical information. It enables the independent search and visualization of each conceptual entity (eg, patient, biology, diagnoses, and CNs) that composes the entire health information of the SHDW. Clinical information originating

XSL•FO
RenderX

from HDWs inherits from a complex data structure. This data structure is very different from documentary and bibliographic IR context, which has been studied in the last 2 decades [22,35] and involves a limited number of entities and relationships and where data are classically more *flatly* structured with a limited depth (basically, a resource entity possibly surrounded with several other entities, such as an author or an editor entity). We believe that this entity-oriented vision of the SHDW gives added value to the IR systems dedicated to HDW, compared with existing solutions, such as i2b2, which usually adopt a patient-centered vision and provide the user with aggregated data and lists of patients as a result. Notably, the system allows the search to be conducted in an iterative manner by visualizing the search of each entity before aggregating all of them into a comprehensive and coherent search.

In addition, the underlying powerful query language used by the system makes the querying of entity-based co-occurring events more generic and more intuitive (ie, searching several events occurring in the same stay, hospitalization, and medical units) [23,24]. In contrast, that kind of functionality is usually proposed through user-friendly but predefined and specific forms (eg, STRIDE [14] and i2b2 [25,26]).

One of the fundamental aspects of SHDW is the semantic description of the health information. This was achieved with the help of many health T&Os provided by the HeTOP. The ECMT semantic annotator notably enabled to automatically annotate the 11.9 million CNs, and thus, provided a semantic access to these CNs despite the difficulties to access unstructured information contained within CNs. A bunch of semantic annotators have been proposed for English texts. Recently, Névéol et al [41] performed a literature review on NLP tools in health in languages other than English. In this study, French was the most studied language, followed by German and Chinese. Nevertheless, most of the existing semantic annotators usually extract concepts from the Unified Medical Language System (UMLS) Metathesaurus (eg, MetaMap [12]) [42] or from mainly English T&Os such as the SNOMED-CT terminology (eg, Text Analysis and Knowledge Extraction System, SNOMED-CT and RxNorm [43], and the National Center for Biomedical Ontology Annotator [44]). French is little represented in the UMLS [45]. The HeTOP includes only 17 KOSs of the 2017 edition of the UMLS. However, the UMLS only manages 11 resources providing concepts in French, and among the 978,233 concept unique identifiers of the UMLS included in the HeTOP, only 143,762 (143,762/978,233, 14.7%) concepts in French originate from the UMLS. In contrast, the HeTOP provides access in French to 428,854 of them (428,854/978,233, 43.8%; almost 3 times more than the UMLS).

The vast majority of HDWs are based on RDBMS (eg, i2b2, STRIDE, DW4TR, SMEYEDAT, and Dr Warehouse). In contrast, the system proposed in this study relies on a NoSQL solution that overcomes the limitations of RDBMS as far as the scalability of data is concerned.

Limitations

From a holistic point of view, the level of support clearly decreased as the CNs became the predominant source of information. The exploitation of unstructured data (N) is consequently considered as the major challenge for the SHDW in this study. More advanced methods of information extraction from those unstructured data, such as the extraction and exploitation of quantitative values from CNs (which is only partially implemented in our system) or the on-the-fly computation of relevant measures (eg, body mass index), could drastically improve the capabilities of the system.

Furthermore, despite the growing interest in statistical machine learning methods, rule-based NLP methods remain predominant as far as clinical information extraction is concerned mainly because of their potential of interoperability and interpretability [9,10]. Nevertheless, since 2018, our team has engaged new research on the semantic annotator ECMT to investigate the development of a hybrid approach between bag-of-words algorithm and word embeddings.

Temporal and chronological aspects are a topic of interest of many IR systems and particularly relevant to IR in clinical data [15]. For instance, DW4TR fully integrates temporal aspects to data modeling by the mean of 3 types of temporal information (eg, static, events, and intervals). Together with clinical and biological data, temporal information is 1 of the 3 dimensions of the 3-dimensional representation of health data in DW4TR. Temporal querying (ie, querying data occurring at a definite moment in time) can be achieved by the underlying search engine of the SHDW and its associated specific query language, but the ASIS Web interface still needs to be enhanced to provide specific forms able to generate the entire set of proper string-based queries. In contrast, the querying of chronologically co-occurring events (ie, searching events occurring before, after, at the same time, or within a definite time frame compared with another) is not well supported. Our department is currently discussing generic technical upgrades of the SSE that will enable us to overcome those limitations and also offer powerful functionalities beyond the scope of time handling.

One of the major drawbacks of NoSQL layer used for data access (ie, Infinispan), and more generally, of many in-memory key-value stores, is that no comprehensive query language is provided as opposed to the SQL for RDBMS. Complex querying capabilities must be fully implemented from the basic application programing interface (ie, obtaining and removing a value of a specific key) proposed by this kind of solution. In particular, in this study, neither join nor reverted index functionality is natively fully provided by Infinispan and requires, respectively, the maintenance of custom maps and the use of Lucene [46] tools to enable the search from concrete values (ie, text and numerical and data values).

An optimized version of the SHDW is, nevertheless, currently in progress.

Conclusions

An HDW is defined as a grouping of data from diverse sources accessible by a single data management system [13] that centralizes clinical, demographic, and administrative data within a uniform and consistent data model. In this study, a POC of an SHDW based on the data of 250,000 patients from RUH is presented along with a graphical interface semantic access to health information. The system provides semantic IR capabilities

XSL•FO
RenderX

and relies on 3 distinct semantic layers. The system was evaluated for its ability to support prescreening of eligible patients in 5 randomly selected clinical trials from RUH. The system showed encouraging results in accurately automating the search of the criteria and good results when used as a prescreening tool. However, this study underlines some limitations of the system especially in relation to information extraction from unstructured CNs, which is still an essential source of information. Since November 2018, all the data of 1.8 million patients from RUH have been included in the POC, and an optimized version is in progress since July 2019.

Acknowledgments

The authors are grateful to Nikki Sabourin-Gibbs, RUH, for help in editing the paper.

Conflicts of Interest

None declared.

References

1. O'Connor PJ, Sperl-Hillen JM, Rush WA, Johnson PE, Amundson GH, Asche SE, et al. Impact of electronic health record clinical decision support on diabetes care: a randomized trial. Ann Fam Med 2011;9(1):12-21 [FREE Full text] [doi: 10.1370/afm.1196] [Medline: 21242556]
2. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014;21(2):221-230 [FREE Full text] [doi: 10.1136/amiajnl-2013-001935] [Medline: 24201027]
3. Krasowski M, Schriever A, Mathur G, Blau J, Stauffer S, Ford B. Use of a data warehouse at an academic medical center for clinical pathology quality improvement, education, and research. J Pathol Inform 2015;6:45 [FREE Full text] [doi: 10.4103/2153-3539.161615] [Medline: 26284156]
4. VanLangen K, Wellman G. Trends in electronic health record usage among US colleges of pharmacy. Curr Pharm Teach Learn 2018 May;10(5):566-570. [doi: 10.1016/j.cptl.2018.01.010] [Medline: 29986815]
5. Cottle M, Hoover W, Kanwal S, Kohn M, Strome T, Treister NW. Transforming Health Care Through Big Data. 2013. Transforming Health Care Through Big Data URL: http://c4fd63cb482ce6861463-bc6183f1c18e748a49b87a25911a0555.r93.cf2.rackcdn.com/iHT2_BigData_2013.pdf [accessed 2019-10-02]
6. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014;2:3 [FREE Full text] [doi: 10.1186/2047-2501-2-3] [Medline: 25825667]
7. Petro J. KevinMD. 2011 Sep 1. Natural Language Processing in Electronic Health Records URL: https://www.kevinmd.com/blog/2011/09/natural-language-processing-electronic-health-records.html, [accessed 2019-10-02]
8. Raghavan P, Chen JL, Fosler-Lussier E, Lai AM. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? AMIA Jt Summits Transl Sci Proc 2014;2014:218-223 [FREE Full text] [Medline: 25717416]
9. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
10. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform 2017 Sep;73:14-29 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.012] [Medline: 28729030]
11. Doan S, Conway M, Phuong TM, Ohno-Machado L. Natural language processing in biomedicine: a unified system architecture overview. Methods Mol Biol 2014;1168:275-294. [doi: 10.1007/978-1-4939-0847-9_16] [Medline: 24870142]
12. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
13. International Organization for Standardization. 2010. Health Informatics — Deployment of a Clinical Data Warehouse URL: https://www.iso.org/standard/45582.html [accessed 2019-10-02]
14. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE--An integrated standards-based translational research informatics platform. AMIA Annu Symp Proc 2009 Nov 14;2009:391-395 [FREE Full text] [Medline: 20351886]
15. Hu H, Correll M, Kvecher L, Osmond M, Clark J, Bekhash A, et al. DW4TR: A data warehouse for translational research. J Biomed Inform 2011 Dec;44(6):1004-1019 [FREE Full text] [doi: 10.1016/j.jbi.2011.08.003] [Medline: 21872681]
16. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc 2010;17(2):131-135 [FREE Full text] [doi: 10.1136/jamia.2009.002691] [Medline: 20190054]
17. Zapletal E, Rodon N, Grabar N, Degoulet P. Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case. Stud Health Technol Inform 2010;160(Pt 1):193-197. [doi: 10.3233/978-1-60750-588-4-193] [Medline: 20841676]

XSL•FO
RenderX

18. Hanauer DA, Mei Q, Law J, Khanna R, Zheng K. Supporting information retrieval from electronic health records: a report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). J Biomed Inform 2015 Jun;55:290-300 [FREE Full text] [doi: 10.1016/j.jbi.2015.05.003] [Medline: 25979153]
19. Grosjean J, Merabti T, Dahamna B, Kergourlay I, Thirion B, Soualmia LF, et al. Health multi-terminology portal: a semantic added-value for patient safety. Stud Health Technol Inform 2011;166:129-138. [doi: 10.3233/978-1-60750-740-6-129] [Medline: 21685618]
20. Soualmia LF, Cabot C, Dahamna B, Darmoni SJ. SIBM at CLEF e-Health Evaluation Lab 2015. In: Proceedings of the 2015 Conference and Labs of the Evaluation forum. 2015 Presented at: CLEF'15; September 8-11, 2015; Toulouse, France URL: http://ceur-ws.org/Vol-1391/125-CR.pdf
21. Cabot C, Soualmia LF, Dahamna B, Darmoni SJ. SIBM at CLEF eHealth Evaluation Lab 2016: Extracting Concepts in French Medical Texts with ECMT and CIMIND. In: Proceedings of the 2016 Conference and Labs of the Evaluation forum. 2016 Presented at: CLEF'16; September 5-8, 2016; Évora, Portugal URL: http://ceur-ws.org/Vol-1609/16090047.pdf
22. Lelong R, Lina F S, Sakji S, Dahamna B, Darmoni S. NoSQL technology in order to support Semantic Health Search Engine. : April 2018 Presented at: Medical Informatics Europe, MIE 2018; Goteborg, Sweden URL: https://hal.archives-ouvertes.fr/hal-02103574
23. Lelong R, Soualmia L, Dahamna B, Griffon N, Darmoni SJ. Querying EHRs with a semantic and entity-oriented query language. Stud Health Technol Inform 2017;235:121-125. [doi: 10.3233/978-1-61499-753-5-121] [Medline: 28423767]
24. Lelong R, Cabot C, Soualmia LF, Darmoni SJ. Semantic Search Engine to Query into Electronic Health Records with a Multiple-Layer Query Language. In: Proceedings of the Association for Computing Machinery's (ACM) Special Interest Group on Information Retrieval (SIGIR). 2016 Presented at: SIGIR'16; 2016; Pisa, Italy URL: http://medir2016.imag.fr/data/MEDIR_2016_paper_8.pdf
25. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc 2006:1040 [FREE Full text] [Medline: 17238659]
26. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130 [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
27. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [doi: 10.3233/978-1-61499-564-7-574] [Medline: 26262116]
28. Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, et al. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. J Biomed Inform 2018 Apr;80:52-63 [FREE Full text] [doi: 10.1016/j.jbi.2018.02.019] [Medline: 29501921]
29. Heudel P, Livartowski A, Arveux P, Willm E, Jamain C. [The ConSoRe project supports the implementation of big data in oncology]. Bull Cancer 2016 Nov;103(11):949-950. [doi: 10.1016/j.bulcan.2016.10.001] [Medline: 27816168]
30. Cuggia M, Garcelon N, Campillo-Gimenez B, Bernicot T, Laurent JF, Garin E, et al. Roogle: an information retrieval engine for clinical data warehouse. Stud Health Technol Inform 2011;169:584-588. [doi: 10.3233/978-1-60750-806-9-584] [Medline: 21893816]
31. Delamarre D, Bouzille G, Dalleau K, Courtel D, Cuggia M. Semantic integration of medication data into the EHOP Clinical Data Warehouse. Stud Health Technol Inform 2015;210:702-706. [doi: 10.3233/978-1-61499-512-8-702] [Medline: 25991243]
32. Kortüm KU, Müller M, Kern C, Babenko A, Mayer WJ, Kampik A, et al. Using electronic health records to build an ophthalmologic data warehouse and visualize patients' data. Am J Ophthalmol 2017 Jun;178:84-93. [doi: 10.1016/j.ajo.2017.03.026] [Medline: 28365240]
33. Qlik: Data Analytics for Modern Business Intelligence. QlikView URL: https://www.qlik.com/us/products/qlikview [accessed 2019-02-15] [WebCite Cache ID 76Cws45GA]
34. Cabot C, Lelong R, Grosjean J, Soualmia LF, Darmoni SJ. Retrieving clinical and omic data from electronic health records. Stud Health Technol Inform 2016;221:115. [Medline: 27071889]
35. Darmoni SJ, Thirion B, Leroyt JP, Douyère M, Lacoste B, Godard C, et al. A search tool based on 'encapsulated' MeSH thesaurus to retrieve quality health resources on the internet. Med Inform Internet Med 2001;26(3):165-178. [doi: 10.1080/14639230110064488] [Medline: 11706927]
36. Cabot C, Soualmia LF, Grosjean J, Griffon N, Darmoni SJ. Evaluation of the terminology coverage in the french corpus LISSa. Stud Health Technol Inform 2017;235:126-130. [doi: 10.3233/978-1-61499-753-5-126] [Medline: 28423768]
37. PostgreSQL. URL: https://www.postgresql.org/ [accessed 2019-02-15] [WebCite Cache ID 76CxVKdFk]
38. Grosjean J. CISMeF - CHU de Rouen. 2014. Modélisation, réalisation et évaluation d'un portail Multi-terminologique, Multi-discipline, Multi-lingue (3M) dans le cadre de la Plateforme d'Indexation Régionale (PlaIR) URL: http://www.chu-rouen.fr/tibs/wp-content/uploads/th%C3%A8se_Julien_Grosjean_final.pdf [accessed 2019-10-02]
39. Infinispan Data Grid Platform. URL: http://infinispan.org/ [accessed 2019-02-15] [WebCite Cache ID 76Cw0We8b]
40. Marchioni F, Surtani M. Infinispan Data Grid Platform. Birmingham, United Kingdom: Packt Publishing; 2012.

XSL•FO
RenderX

41. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than English: opportunities and challenges. J Biomed Semantics 2018 Mar 30;9(1):12 [FREE Full text] [doi: 10.1186/s13326-018-0179-8] [Medline: 29602312]
42. Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Yearb Med Inform 2018;02(01):41-51. [doi: 10.1055/s-0038-1637976]
43. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
44. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 2009 Jul;37(Web Server issue):W170-W173 [FREE Full text] [doi: 10.1093/nar/gkp440] [Medline: 19483092]
45. Névéol A, Grosjean J, Darmoni SJ, Zweigenbaum P. Language Resources for French in the Biomedical Domain. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation. 2014 Presented at: LREC'14; May 26-31, 2014; Reykjavik, Iceland p. 2146-2151 URL: https://www.aclweb.org/anthology/L14-1487/
46. Appache Lucene. URL: https://lucene.apache.org/ [accessed 2019-10-02]

Abbreviations

CN: clinical narrative
DRG: diagnosis-related group
DW4TR: Data Warehouse for Translational Research
EAV: entity-attribute-value
ECMT: extracting concepts from multiple terminologies
EMERSE: Electronic Medical Record Search Engine
ETL: extract-transform-load
HDW: health data warehouse
HeTOP: health T&O portal
HIS: Hospital Information System
i2b2: Informatics for Integrating Biology and the Bedside
ICD: International Classification of Diseases
IMDG: in-memory data grid
IR: information retrieval
IS: information system
KOS: knowledge organization system
NLP: natural language processing
NoSQL: not only structured query language
NYHA: New York Heart Association
POC: proof of concept
RAM: random access memory
RDBMS: Relational DataBase Management System
RIM: Reference Information Model
RUH: Rouen University Hospital
SHDW: semantic health data warehouse
SMEYEDAT: SMart Eye DATabase
SNOMED-CT: Systematized Nomenclature Of MEDicine–Clinical Terms
SQL: structured query language
SSE: semantic search engine
STRIDE: Stanford Translational Research Integrated Database Environment
T&O: terminology and ontology
UMLS: Unified Medical Language System

XSL•FO
RenderX

Edited by G Eysenbach; submitted 05.03.19; peer-reviewed by C Liang, B Schreiweis; comments to author 07.06.19; revised version received 02.08.19; accepted 19.08.19; published 20.12.19.

Please cite as:
Lelong R, Soualmia LF, Grosjean J, Taalba M, Darmoni SJ
Building a Semantic Health Data Warehouse in the Context of Clinical Trials: Development and Usability Study
JMIR Med Inform 2019;7(4):e13917
URL: http://medinform.jmir.org/2019/4/e13917/
doi:10.2196/13917
PMID:31859675

©Romain Lelong, Lina F Soualmia, Julien Grosjean, Mehdi Taalba, Stéfan J Darmoni. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 20.12.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Characterizing Artificial Intelligence Applications in Cancer Research: A Latent Dirichlet Allocation Analysis

Bach Xuan Tran[1,2], PhD; Carl A Latkin[2], PhD; Noha Sharafeldin[3,4], PhD, MBChB; Katherina Nguyen[5], BA; Giang Thu Vu[6], MSc; Wilson W S Tam[7], PhD; Ngai-Man Cheung[8,9], PhD; Huong Lan Thi Nguyen[10], MSc; Cyrus S H Ho[11], MBBS; Roger C M Ho[12,13,14], MBBS

[1]Institute for Preventive Medicine and Public Health, Hanoi Medical University, Hanoi, Vietnam
[2]Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States
[3]Division of Hematology & Oncology, Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, United States
[4]Institute for Cancer Outcomes and Survivorship, School of Medicine, University of Alabama at Birmingham, Birmingham, AL, United States
[5]Department of Science, Technology, and Society, Stanford University, Palo Alto, CA, United States
[6]Center of Excellence in Evidence-based Medicine, Nguyen Tat Thanh University, Ho Chi Minh, Vietnam
[7]Alice Lee Centre for Nursing Studies, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[8]Center of Excellence in Artificial Intelligence in Medicine, Nguyen Tat Thanh University, Ho Chi Minh, Vietnam
[9]Information Systems Technology and Design, Singapore University of Technology and Design, Singapore, Singapore
[10]Institute for Global Health Innovations, Duy Tan University, Da Nang, Vietnam
[11]Department of Psychological Medicine, National University Hospital, Singapore, Singapore
[12]Center of Excellence in Behavior Medicine, Nguyen Tat Thanh University, Ho Chi Minh, Vietnam
[13]Institute for Health Innovation and Technology, National University of Singapore, Singapore, Singapore
[14]Department of Psychological Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Corresponding Author:
Bach Xuan Tran, PhD
Institute for Preventive Medicine and Public Health
Hanoi Medical University
No 1 Ton That Tung Street
Hanoi, 100000
Vietnam
Phone: 84 982228662
Email: bach.ipmph@gmail.com

Abstract

Background: Artificial intelligence (AI)–based therapeutics, devices, and systems are vital innovations in cancer control; particularly, they allow for diagnosis, screening, precise estimation of survival, informing therapy selection, and scaling up treatment services in a timely manner.

Objective: The aim of this study was to analyze the global trends, patterns, and development of interdisciplinary landscapes in AI and cancer research.

Methods: An exploratory factor analysis was conducted to identify research domains emerging from abstract contents. The Jaccard similarity index was utilized to identify the most frequently co-occurring terms. Latent Dirichlet Allocation was used for classifying papers into corresponding topics.

Results: From 1991 to 2018, the number of studies examining the application of AI in cancer care has grown to 3555 papers covering therapeutics, capacities, and factors associated with outcomes. Topics with the highest volume of publications include (1) machine learning, (2) comparative effectiveness evaluation of AI-assisted medical therapies, and (3) AI-based prediction. Noticeably, this classification has revealed topics examining the incremental effectiveness of AI applications, the quality of life, and functioning of patients receiving these innovations. The growing research productivity and expansion of multidisciplinary approaches are largely driven by machine learning, artificial neural networks, and AI in various clinical practices.

XSL•FO
RenderX

Conclusions: The research landscapes show that the development of AI in cancer care is focused on not only improving prediction in cancer screening and AI-assisted therapeutics but also on improving other corresponding areas such as precision and personalized medicine and patient-reported outcomes.

(JMIR Med Inform 2019;7(4):e14401) doi:10.2196/14401

KEYWORDS

scientometrics; cancer; artificial intelligence; global; mapping

Introduction

Background

Every year, over 200 million healthy life years are lost because of cancer, making it one of the highest health care burden causing disability and mortality among men and women [1]. Fortunately, many types of cancers can be prevented or effectively treated if patients are diagnosed in a timely manner and offered optimal therapies. In many parts of the world, however, programs for cancer control and prevention are facing multiple barriers because of limited health service infrastructure, availability of treatment options, and health worker capacities.

Artificial intelligence (AI) is considered a disruptive innovation in health and medicine. Over the past six decades, AI has been widely applied to many areas of medical research and clinical practice. The number of published papers on AI and its impacts has been rapidly growing within the research community over the past decade. A bibliometric study has shown that the number of studies on AI applications in medicine has tripled in the past 3 years, with the highest interest in cancer research [2]. Various techniques, such as robotics, machine learning, and artificial neural networks, have been applied to the study of cancer, showing promising improvements in clinical prediction, treatment, and diagnosis. For instance, machine learning techniques in the application of proteomics and genomics could increase precision in estimating survival and inform the selection of therapies [3]. In large populations, the development and application of AI also holds potential in screening for cancer and scaling up treatment services in a timely manner.

Literature Review

Many approaches and products have been developed to support cancer treatment and for prevention at health facilities and within communities. However, the synthesis of resulting evidence from these efforts is necessary to inform decision making. Some authors have conducted systematic reviews of the performance and effectiveness of AI techniques and products in specific cancers [3-10]. Overall, these reviews found that almost all AI-assisted interventions led to greater effectiveness than conventional approaches. However, insights from these efforts have raised some important points for further exploration. Lisboa et al reviewed predictive models using artificial neural networks and suggested the need for rigorous evaluation of results [4]. In addition, Spelt et al emphasized the importance of justifying the complex structure of datasets and individual factors in these models [5]. Ray et al reviewed the wearable systems for cancer detection and found that cloud computing and long-range communication paradigms are still lacking, and that AI and machine learning should be applied to current products [8]. Other authors affirmed the greater performance of image-based AI applications to breast cancer diagnosis, but few studies have been supported by a high level of evidence. Conducting further clinical research and health technology assessment is recommended.

Objectives

With the rapid development of technologies, AI-based therapeutics, devices, and systems will be vital innovations in cancer control. To accelerate research and development, it is critical to understand current approaches in the applications of AI in cancer care, multiple disciplines involved, and the trends and establishment of the research landscapes. To our knowledge, none of the previous studies have systematically quantified the development of AI in the bibliographic literature of cancer studies. This study analyzes the global trends, patterns, and development of interdisciplinary landscapes in AI and cancer studies.

Methods

Search Strategy

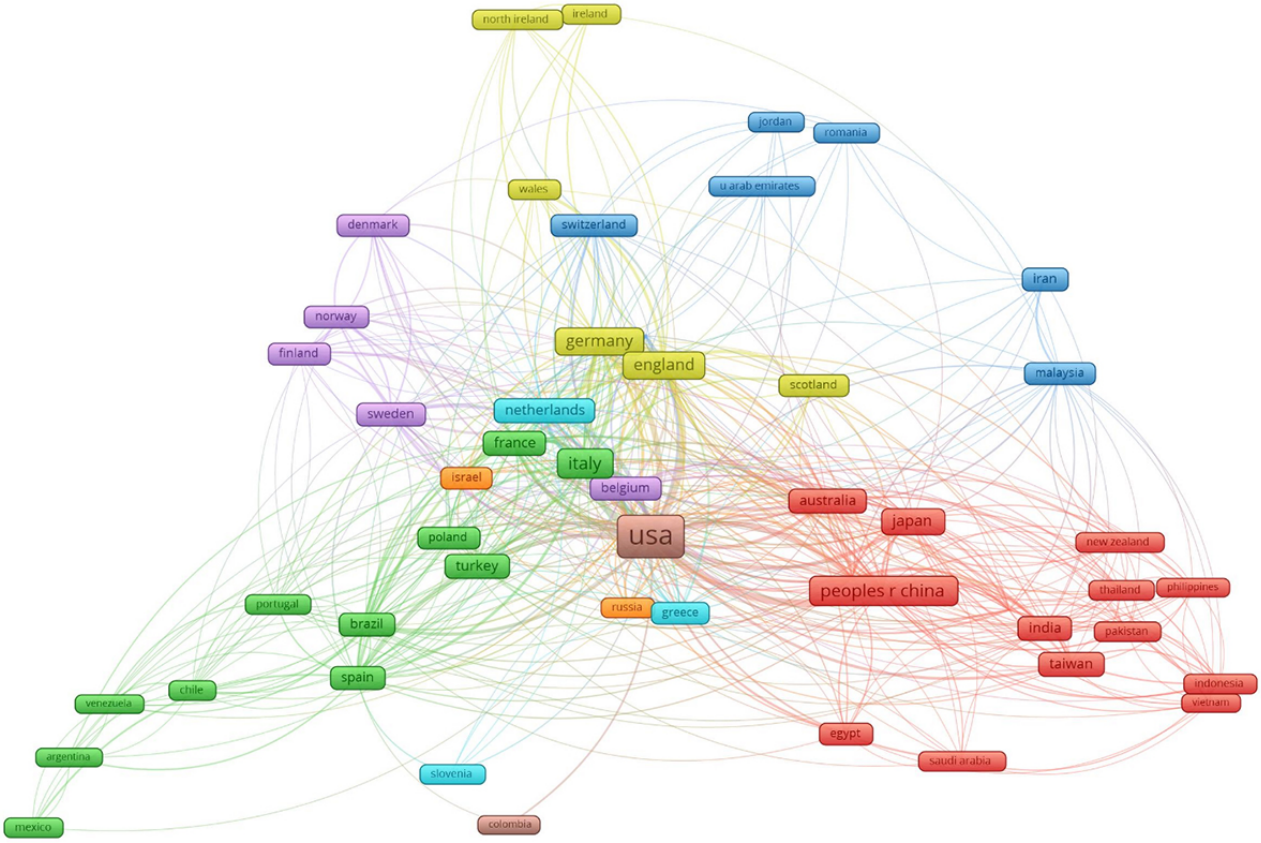
We searched and retrieved all papers related to AI in cancer care on the Web of Science (WOS) that is a Web-based database covering the largest proportion of peer-reviewed literature in this field. The full search strategy has been presented elsewhere [2]. In short, we used a set of predefined search terms related to *artificial intelligence* and *health and medicine* to search the WOS for publications (inclusion step) and then excluded those that did not satisfy our eligibility criteria of publication year from 1991 to 2018 and publications other than articles and reviews (exclusion step). In this analysis, we selected all the documents of retrieved data on AI applications related to *cancer care*.

Data Extraction

We downloaded all data from the WOS database in .txt format, including all information such as author names, paper title, journals, keywords, affiliations of institutions, the prevalence of citation, categories, and abstracts. All of these data were converted to an Excel file (Microsoft Excel, Microsoft Corporation) for checking the data error. A process of standardization was carried out by 2 researchers to bring together the different names of an author. Then, we filtered all downloaded data using the following criteria: (1) not original articles and reviews, (2) not about cancer and AI, and (3) not in English. Any conflict was solved by discussion (Figure 1). The combined dataset was transferred into Stata (version 14.0, STATA Corporation) for further analysis.

XSL•FO
RenderX

Figure 1. The global networking of 53 countries having at least five coauthorships classified in 8 clusters.



Data Analysis

Data were resolved based on basic indicators of publication (number of authors, publication years, and main categories), keywords (most common keywords and co-occurrence keywords), citations, usages, and abstracts. After downloading and extracting the data, we applied the descriptive statistical analysis using Stata to calculate country citations and intercountry collaboration. A network graph illustrating the network of countries by sharing the co-authorships was created, along with the author keyword co-occurrence network and countries network. VOSviewer (version 1.6.8, Center for Science and Technology, Leiden University) was used to establish a co-occurrence network and a countries network. The principles of underlying algorithms used by the software for clustering have been documented elsewhere [11-14] For content analysis of the abstracts, we applied the exploratory factor analysis to identify research domains emerging from all content of the abstracts, loadings of 0.4 [15]. The Jaccard similarity index was utilized to identify research topics or terms most frequently co-occurring with each other [16]. Latent Dirichlet Allocation (LDA) was used for classifying papers into corresponding topics [17-21]. The summary of analytical techniques for each data type is presented in Table 1.

Table 1. Summary of data analytical techniques.

Type of data	Unit of analysis	Analytical methods	Presentations of results
Authors, keywords, countries	Words	Frequency of co-occurrence	Map of authors keywords clusters
Abstracts	Words	Exploratory factors analyses	Top 50 constructed research domains; clustering map of the landscapes constructed by these domains
Abstracts	Papers	Latent Dirichlet Allocation	10 classifications of research topics
WOS[a] classification of research areas	WOS research areas	Frequency of co-occurrence	Dendrogram of research disciplines (WOS classification)

[a]WOS: Web of Science.

Results

The Number of Published Items and Publication Trend

There has been a rapid increase in the number of studies applying AI to cancer research from 1991 to 2018. In particular, the research productivity of the past 10 years has accounted for over 90.66% (3223/3555) of the total papers. Rates of citation and usage are also growing fast. The mean usage (downloads) in the past 6 months of papers published in the past 1 to 2 years was twice that of those published in the past 3 to 4 years (Table 2).

XSL•FO
RenderX

In Table 3, we examine the study settings mentioned in the abstracts of publications. The bibliography included country settings 749 times, and in those, the United States was mentioned 46.5% of the times. Over 90% of the total settings were in developed countries. Noticeably, 2 countries with large populations, China and India, accounted for 3.3% and 4.4%, respectively.

Table 2. General characteristics of publications.

Year published	Total number of papers	Total citations	Mean citation rate per year[a]	Total usage in the last 6 months[b]	Total usage in the last 5 years[b]	Mean use rate for the last 6 months[c]	Mean use rate for the last 5 years[d]
2018	661	809	1.22	2489	3858	3.77	1.17
2017	503	3206	3.19	994	4663	1.98	1.85
2016	435	3680	2.82	431	5529	0.99	2.54
2015	349	4524	3.24	304	3713	0.87	2.13
2014	284	4131	2.91	140	2914	0.49	2.05
2013	268	5167	3.21	118	2893	0.44	2.16
2012	202	4642	3.28	66	1511	0.33	1.50
2011	173	4706	3.40	64	1150	0.37	1.33
2010	146	5474	4.17	51	881	0.35	1.21
2009	114	3550	3.11	55	729	0.48	1.28
2008	88	3671	3.79	36	478	0.41	1.09
2007	68	2480	3.04	24	388	0.35	1.14
2006	58	2324	3.08	18	238	0.31	0.82
2005	45	1885	2.99	14	219	0.31	0.97
2004	26	1582	4.06	6	134	0.23	1.03
2003	39	3115	4.99	22	399	0.56	2.05
2002	17	3208	11.10	32	297	1.88	3.49
2001	15	964	3.57	2	75	0.13	1.00
2000	18	2040	5.96	11	192	0.61	2.13
1999	13	1043	4.01	5	51	0.38	0.78
1998	12	548	2.17	4	31	0.33	0.52
1997	9	420	2.12	5	28	0.56	0.62
1996	2	52	1.13	0	4	0.00	0.40
1995	2	297	6.19	5	28	2.50	2.80
1994	4	172	1.72	0	9	0.00	0.45
1993	0	0	0	0	0	0.00	0.00
1992	3	105	1.30	2	6	0.67	0.40
1991	1	2	0.07	0	2	0.00	0.40

[a]Mean citation rate per year=total citations/(total citations×[2018−that year]).

[b]Total usage: total downloads.

[c]Mean use rate for the last 6 months=total usage in the last 6 months/total number of papers.

[d]Mean use rate for the last 5 years=total usage in the last 5 years/(total number of papers×5).

XSL•FO
RenderX

Table 3. Number of papers by countries as study settings (N=749).

Rank	Country settings	Frequency, n (%)
1	United States	348 (46.5)
2	Ireland	47 (6.3)
3	Taiwan	44 (5.9)
4	Japan	41 (5.5)
5	United Kingdom	37 (4.9)
6	India	33 (4.4)
7	China	25 (3.3)
8	Australia	22 (2.9)
9	Italy	14 (1.9)
10	Mali	12 (1.6)
11	Sweden	11 (1.5)
12	Wallis and Futuna	10 (1.3)
13	Germany	9 (1.2)
14	Netherlands	9 (1.2)
15	Poland	9 (1.2)
16	France	7 (0.9)
17	Spain	7 (0.9)
18	Denmark	6 (0.8)
19	Hong Kong	6 (0.8)
20	Canada	5 (0.7)
21	Finland	5 (0.7)
22	Iran	5 (0.7)
23	Singapore	4 (0.5)
24	Belgium	3 (0.4)
25	Brazil	3 (0.4)
26	Egypt	3 (0.4)
27	Israel	3 (0.4)
28	Malaysia	3 (0.4)
29	Turkey	3 (0.4)
30	New Zealand	2 (0.3)
31	Norway	2 (0.3)
32	Antarctica	1 (0.1)
33	Austria	1 (0.1)
34	Georgia	1 (0.1)
35	Greece	1 (0.1)
36	Iceland	1 (0.1)
37	Indonesia	1 (0.1)
38	Jersey	1 (0.1)
39	Jordan	1 (0.1)
40	Pakistan	1 (0.1)
41	Saint Pierre	1 (0.1)
42	Saudi Arabia	1 (0.1)

XSL•FO
RenderX

Figure 1 presents the global network among 53 countries having at least five co-authorships with other countries. The range of nodes represents the contribution of each country to the total number of publications, and the thickness of lines indicates the proportion of the volume of collaborations. These countries were classified into 8 clusters depending on their level of international collaborations.

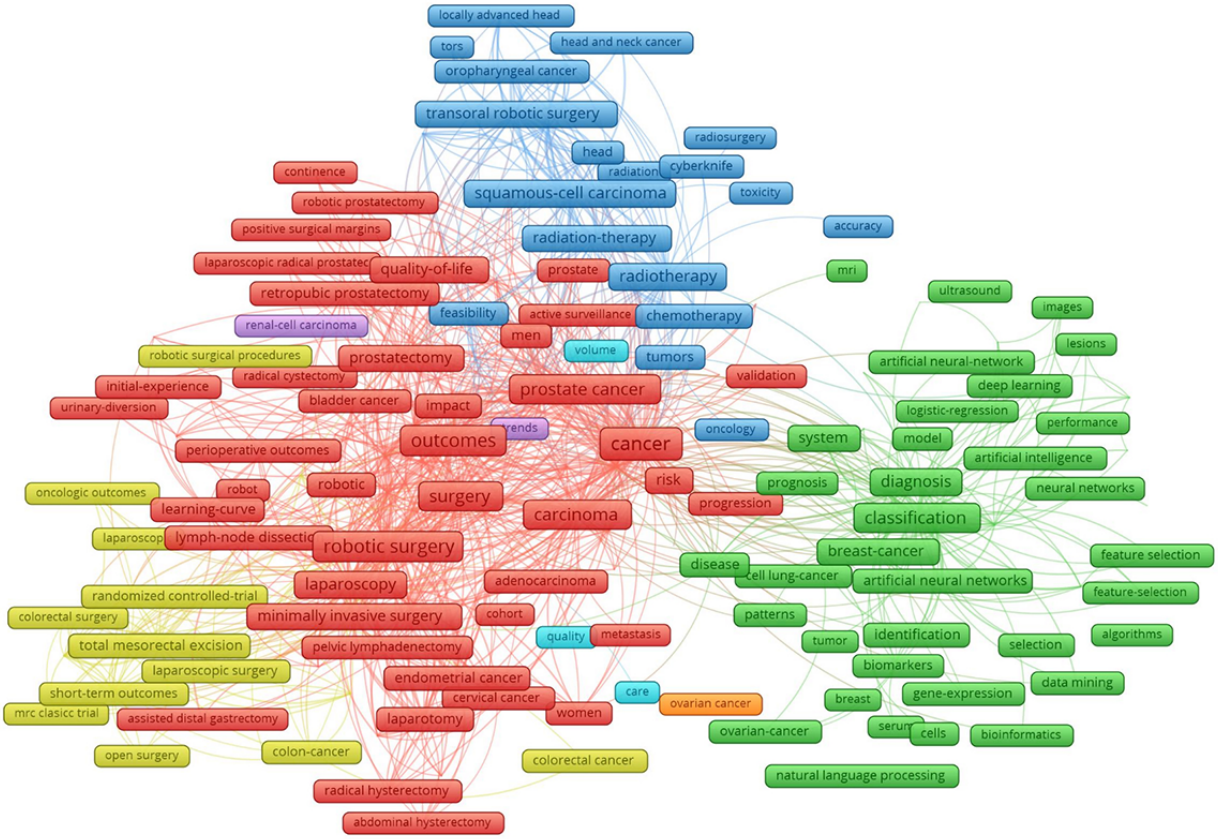
Analyses of keywords and abstract contents provide us with a better understanding of the scopes of studies and development of the research landscapes. Figure 2 describes the co-occurrence of keywords with the most frequent groups of terms. There were 8 major clusters emerging from 180 most frequent keywords with a co-occurrence of 30 times and higher. Some major clusters included the following: Cluster 1 (red) refers to surgery and treatment outcomes; Cluster 2 (green) focuses on the applications of AI techniques in some specific cancers; Cluster 3 (yellow) describes the therapies for colorectal cancers; and Cluster 4 (blue) illustrates applications of chemotherapy and radiotherapy. The colors of the nodes indicate principal components of the data structure; the node size was scaled to the keyword occurrences; and the thickness of the lines is based on the strength of the association between 2 keywords.

As for the content analysis of abstracts, the top 50 emerging research domains are listed in Table 4. AI techniques have been applied to various aspects of cancer research, including therapies (radiotherapy, chemotherapy, and surgery), capacities (prediction, screening, and treatment), and factors associated with outcomes (physical, social, and economic).

Figure 3 illustrates the classification of the co-occurrence of research domains into principal components. Primarily, we have the following major landscapes: (1) robotic surgery (blue), (2) AI techniques for detection and prediction (gray), (3) chemotherapy (jade), and (4) radiotherapy (yellow).

In Table 5, we present the research topics that were constructed using LDA. The labels of the topics were manually annotated by scrutinizing the most frequent words and titles for each topic. Topics with the highest volume of publications included (1) machine learning, (2) comparative effectiveness evaluation of AI-assisted medical therapies, and (3) AI-based prediction. Noticeably, this classification has revealed topics examining the incremental effectiveness of AI applications (Topic 2) and, more interestingly, the quality of life outcomes and functioning of patients receiving these innovations. The changes in research productivity over time are illustrated in Figure 4, which shows the rapid growth of Topics 1, 2, 3, and 4, especially in recent years.

Figure 2. Co-occurrence of the most frequent author's keywords.



XSL•FO
RenderX

Table 4. Top 50 research domains that emerged from the exploratory factor analysis of the content of all abstracts.

Number	Name	Keywords	Eigen value	Cases, n (%)
1	Classification; feature selection	Classification; feature; proposed; features; selection; breast; performance; algorithm; diagnosis; classifier; paper; accuracy; machine	4.71	6104 (58.09)
2	Disease-free survival	Survival; free; recurrence; follow; disease; local	2.94	3413 (52.69)
3	Medical; processing	Medical; processing; information; system; systems	1.85	2218 (42.31)
4	Blood loss; hospital stay	Loss; blood; stay; operative; complications; length; hospital; min; postoperative; conversion; perioperative; complication; safe; intraoperative; feasible	17.16	5050 (41.13)
5	Prostate; assisted radical	Prostatectomy; prostate; radical; RARP[a]; localized; men; assisted; radical prostatectomy; robot; Gleason; RALP[b]; PSA[c]	3.97	4050 (41.04)
6	Gy; radiation dose	Gy; dose; SBRT[d]; radiation; radiotherapy; therapy; local; body; treated	5.18	2495 (37.16)
7	Predict; prediction	Predict; prediction; predictive; models; predicting; prognostic; variables; validation	2.34	2562 (36.71)
8	Machine learning	Learning; machine; accuracy	2.27	2398 (36.68)
9	Cohort; risk	Cohort; risk; outcome; retrospective	1.44	1759 (36.54)
10	PSA; Gleason	PSA; Gleason; specific; biopsy; serum; prostate	1.38	1970 (34.04)
11	Early stage	Early; cervical; stage; hysterectomy	1.48	1700 (33.47)
12	Evaluate	Evaluate; evaluated; according	1.37	1185 (28.16)
13	Training set	Training; set; test; sets; validation	1.60	1573 (27.74)
14	Adjuvant chemotherapy	Chemotherapy; adjuvant; therapy; advanced	1.51	1323 (27.34)
15	Tumor	Tumor; tumors; size	1.46	1253 (27.12)
16	Morbidity and mortality	Mortality; morbidity; rate	1.36	1194 (26.69)
17	Staging for endometrial; hysterectomy	Endometrial; hysterectomy; laparotomy; staging; lymphadenectomy; pelvic; cervical; laparoscopy; women	3.09	1759 (25.26)
18	Sensitivity and specificity	Specificity; sensitivity; serum; detection; diagnostic	2.17	1405 (23.74)
19	Plans; planning	Plans; planning; target; mm; volume; dose; average	1.53	1267 (23.57)
20	Cystectomy; bladder	Cystectomy; bladder; RARC[e]; urinary; radical	2.83	1235 (22.53)
21	Case	Cases; case	1.26	913 (22.50)
22	Artificial neural	Neural; artificial; network; ANN[f]; networks	3.32	2060 (22.17)
23	Image	Images; image; imaging; deep; CT[g]; MRI[h]	2.59	1363 (21.77)
24	Quality of life	Life; quality; health; sexual	2.09	1148 (21.55)
25	Lymph node	Lymph; node; dissection; nodes; pelvic; lymphadenectomy	2.47	1909 (21.24)
26	Safe and feasible	Safe; feasible; procedure	1.25	1020 (20.39)
27	Decision support	Support; SVM[i]; decision; classifier	1.45	1062 (19.86)
28	Rectal resection	Rectal; colorectal; resection; conversion	1.57	926 (19.16)
29	Oncological and functional; sexual function	Functional; function; sexual; oncological	1.31	903 (19.10)
30	Gene expression	Gene; expression; genes; molecular; protein; samples; mutations	3.24	1369 (18.90)
31	Purpose	Purpose; materials	1.51	815 (18.71)
32	Pathology; reports	Pathology; reports; processing; report	1.41	824 (18.54)
33	Women diagnosed	Diagnosed; screening; women	1.29	728 (17.64)
34	Transoral; TORS[j]	Transoral; tors; oropharyngeal; neck; head; HPV[k]; carcinoma	3.48	1183 (16.48)
35	Margin; PT[l]	Margin; PT; margins; pathologic; Gleason; RALP	1.80	884 (16.15)
36	Cost	Cost; costs; care	1.93	712 (16.06)

XSL•FO
RenderX

Number	Name	Keywords	Eigen value	Cases, n (%)
37	Surgeon experience	Experience; surgeons; surgeon	1.70	690 (15.22)
38	Small cell lung	Lung; small	1.29	628 (14.91)
39	Body mass	Mass; index; body	1.89	823 (14.91)
40	Multi drug	Multiple; multi; drug	1.56	588 (14.68)
41	Operating curve	Curve; operating; AUC[m]	1.68	649 (12.52)
42	Benign and malignant	Malignant; benign; lesions	1.59	579 (11.14)
43	Normal tissue	Tissue; normal	1.40	460 (11.11)
44	Imaging (MRI)	MRI; imaging	1.39	435 (10.63)
45	Metastases	Metastases; metastasis; liver	1.39	416 (9.51)
46	Renal	Renal; partial; sparing	1.74	376 (8.27)
47	HPV-negative	Negative; HPV	1.33	312 (8.07)
48	Trials	Trials	1.34	239 (6.72)
49	Biomarkers	Biomarkers	1.25	171 (4.81)
50	Gastrectomy for gastric	Gastric; gastrectomy	1.77	205 (3.63)

[a]RARP: robotic-assisted radical prostatectomy.

[b]RALP: robot assisted laparoscopic prostatectomy.

[c]PSA: prostate specific antigen.

[d]SBRT: stereotactic body radiation therapy.

[e]RARC: remittance advice remark code.

[f]ANN: artificial neural network.

[g]CT: computed tomography.

[h]MRI: magnetic resonance imaging.

[i]SVM: support vector machine.

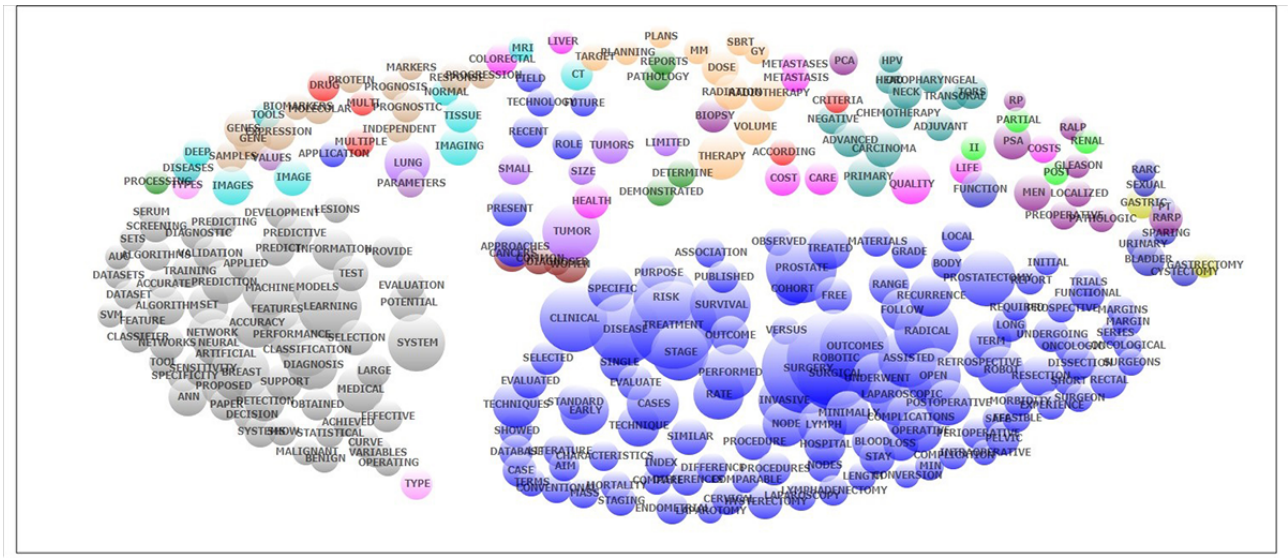
[j]TORS: transoral robotic surgery.

[k]HPV: human papilloma virus.

[l]PT: prothrombin time.

[m]AUC: area under the curve.

Figure 3. Co-occurrence of most frequent topics emerged from exploratory factor analysis of abstracts contents.



XSL•FO
RenderX

Table 5. Ten research topics classified by Latent Dirichlet Allocation.

Topics	Research areas	Frequency (N=3555), n (%)
Topic 1	Machine learning	824 (23.18)
Topic 2	Comparative effectiveness evaluation of AI[a]-assisted medical therapies	513 (14.43)
Topic 3	AI-based prediction	456 (12.83)
Topic 4	Multidisciplinary care, precision, and personalized medicine	371 (10.44)
Topic 5	Quality of life outcomes, physical and mental health, and functioning	312 (8.78)
Topic 6	Enhanced radiotherapy	270 (7.59)
Topic 7	Robotic surgery	229 (6.44)
Topic 8	AI-assisted imaging and signals	215 (6.05)
Topic 9	Data mining and natural language processing	183 (5.15)
Topic 10	AI and robotic-assisted cancer diagnosis and therapies	182 (5.12)

[a]AI: artificial intelligence.

Figure 4. Changes in the applications of artificial intelligence to cancer research during 1991-2018.

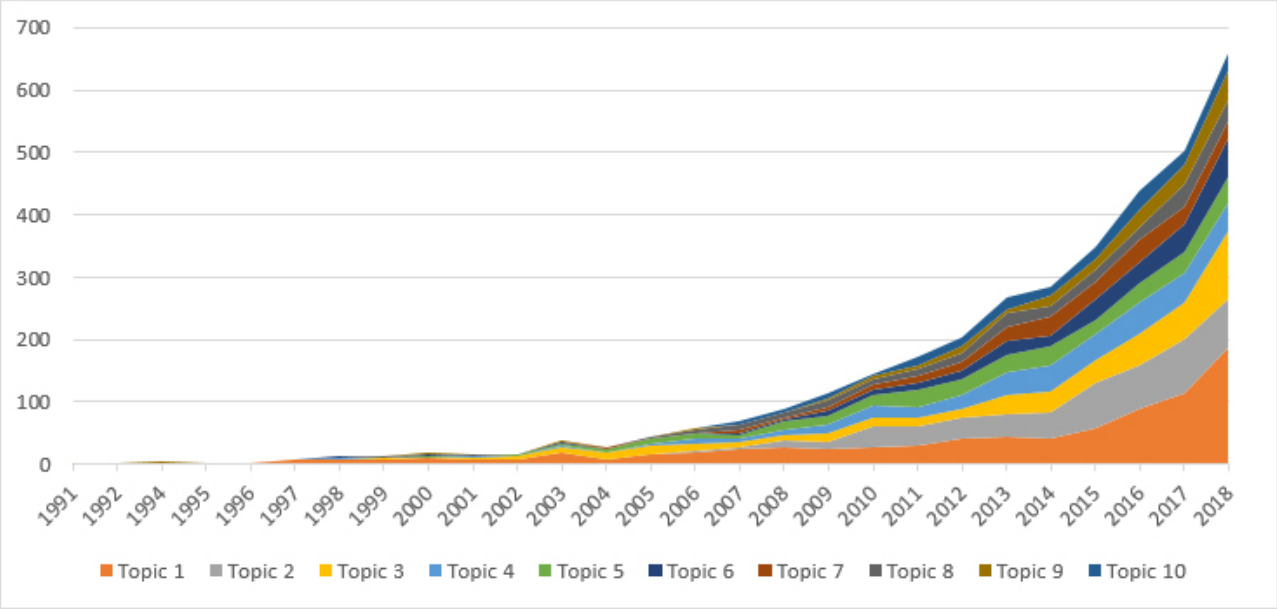
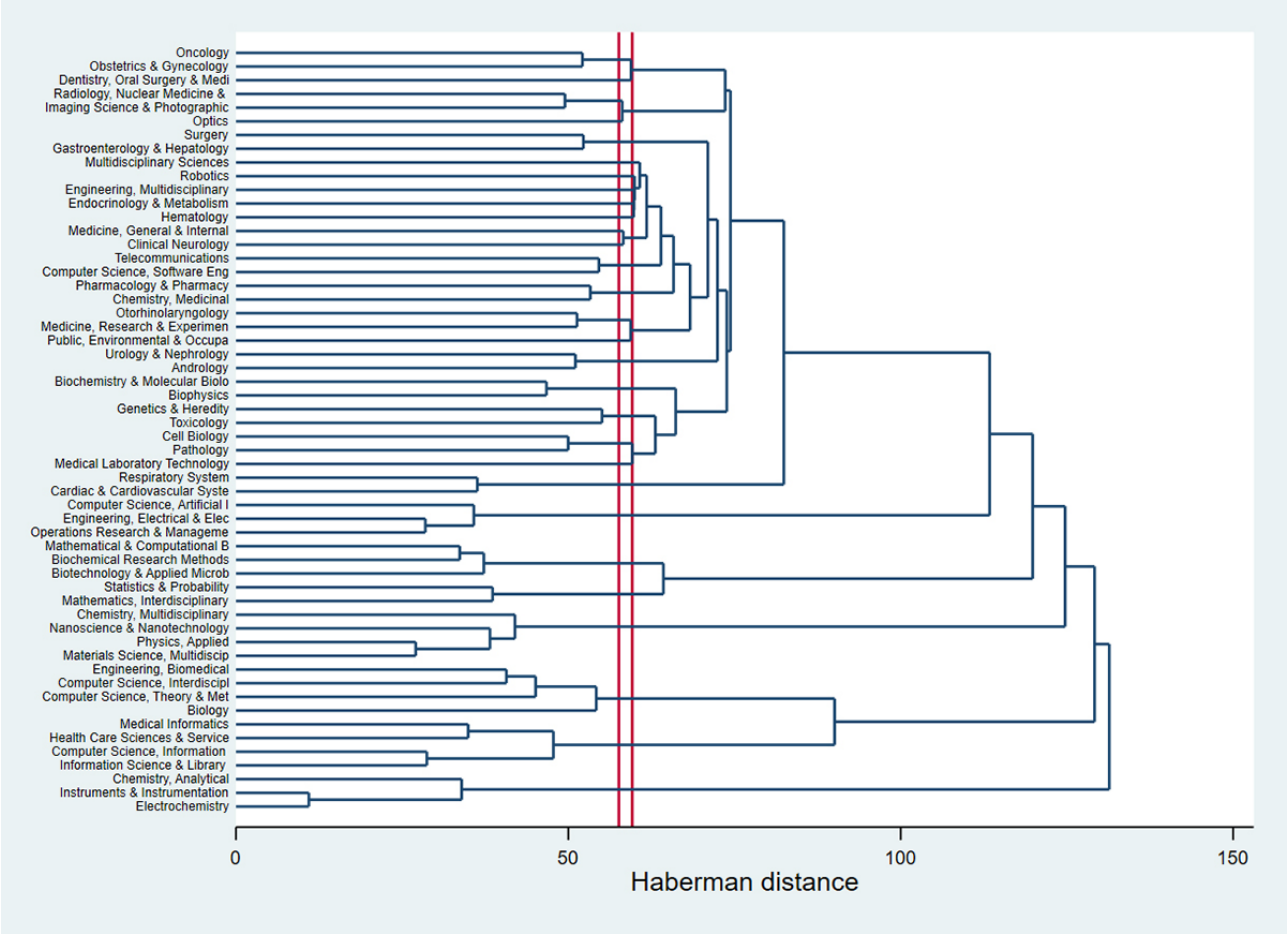


Figure 5 presents the hierarchical clustering of research disciplines used in AI and cancer research. The horizontal axis of the dendrogram represents the distance or dissimilarity between clusters. The vertical axis represents the research disciplines. It shows that AI applications in cancer care are rooted in the following disciplines: robotics, multidisciplinary engineering, and multidisciplinary sciences. Imaging science and photography was very close to oncology, obstetrics and gynecology, dentistry, radiology, and optics. Those biomedical and clinical aspects account for the major areas of AI application; meanwhile, health service–focused areas, for example, operations and management, are rather distant.

XSL•FO
RenderX

Figure 5. Dendogram of coincidence of research areas using the Web of Science classifications.



Discussion

Principal Findings

By systematically synthesizing and analyzing the bibliography of AI applications in cancer studies, we have characterized the development of its research landscape over the period from 1991 to 2018. The findings illustrate the rapidly growing research productivity and expansion of multidisciplinary approaches, largely driven by machine learning, artificial neural networks, and AI in various clinical practices. Our analysis highlights the most frequent areas of research and the paucity of research in other areas. The research topics and landscapes constructed show that the development of AI in cancer care is focused on improving prediction in cancer screening and AI-assisted therapeutics and corresponding areas of precision and personalized medicine. Our findings show the rapid growth in these areas over the past decade. Although cancer outcomes of interest covering clinical and physical functioning and mental and quality of life measures are on the rise, our analysis indicates the relative paucity of research focusing on cancer outcomes and survivorship. This is of special relevance, considering the continuously growing cancer survivor population [22].

Comparison With Past Work

This study supplements the previous global mapping on AI in medicine by analyzing the content and characteristics of studies of specific applications of AI in cancer research and clinical practice [2]. Compared with previous reviews, this study is more comprehensive in describing the research trends by applying content analysis and topic modeling [4-10]. Therefore, the findings are helpful to inform the design and priority of the settings of future studies. Classifying information sources and content in corresponding topics to identify priorities for interventions has been widely applied in many studies. For example, previous authors have analyzed newspaper and social media content to understand topics of interest related to breast cancer and secondhand smoking [23-28]. However, none of the previous studies have analyzed the scientific bibliography to determine the development of research landscapes in AI applied in cancer care. Li et al proposed a text-mining framework using LDA to construct topics that were helpful for supporting systematic reviews [29]. In this study, we applied this approach to classify topics that a paper belongs to. Moreover, we further analyzed the frequency of concurrence of terms and their associated clusters using factor analysis. These clusters of terms enrich the understanding of scopes of each topic, especially for diseases involving the development of multidisciplinary research.

The findings from this study help inform the future development of AI applications in cancer research and clinical practices of cancer control and management. First, the difference in citation rates between very recent articles and older articles demonstrates the speed of knowledge accumulation in this area. Understanding the scope of research landscapes helps inform the selection of variables and topics to develop an application or conduct a study. Moreover, the previous bibliometric analysis could only

XSL•FO
RenderX

distinguish and determine trends in the applications of AI techniques in cancer care, whereas this study showed that research trends have also expanded to encompass the comparative effectiveness of these innovations compared with traditional practices [2]. In addition, research landscapes have expanded beyond clinics to evaluate the functioning and performance of the patients being treated, in addition to their mental well-being and quality of life. To support this research topic, there should be more exploration of different study settings and incorporation of individual characteristics to improve the validity of AI techniques. One important question is how to integrate and scale-up AI-based applications in cancer care into clinical practice and community prevention. Currently, little is known on the adaptation and integration of AI applications into health systems and communities; future implementation research should be conducted.

Limitations

One of the shortcomings of this study is that we used only WOS databases. Although the WOS covers the greatest proportion of the literature in the field of AI research, it might not be fully representative of all databases. Another limitation is that only documents in English were selected for this study. Finally, the content analysis included only abstracts instead of full texts. Nonetheless, this topic modeling serves to expand, improve, and supplement previous systematic reviews in this field.

Conclusions

In conclusion, AI applications have been rapidly growing in cancer clinical practices, including prediction, diagnosis, enhanced therapeutics, and optimal selection. As interest in AI in medicine continues to grow, it will be increasingly critical to better understand the incremental effectiveness of these innovations and their validities in supporting the performance and quality of life of individuals after getting treated.

Conflicts of Interest

None declared.

References

1. GBD 2017 DALYs and HALE Collaborators. Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018 Nov 10;392(10159):1859-1922 [FREE Full text] [doi: 10.1016/S0140-6736(18)32335-3] [Medline: 30415748]
2. Tran BX, Vu GT, Ha GH, Vuong QH, Ho MT, Vuong TT, et al. Global evolution of research in artificial intelligence in health and medicine: a bibliometric study. J Clin Med 2019 Mar 14;8(3):E360 [FREE Full text] [doi: 10.3390/jcm8030360] [Medline: 30875745]
3. Bashiri A, Ghazisaeedi M, Safdari R, Shahmoradi L, Ehtesham H. Improving the prediction of survival in cancer patients by using machine learning techniques: experience of gene expression data: a narrative review. Iran J Public Health 2017 Feb;46(2):165-172 [FREE Full text] [Medline: 28451550]
4. Lisboa PJ, Taktak AF. The use of artificial neural networks in decision support in cancer: a systematic review. Neural Netw 2006 May;19(4):408-415. [doi: 10.1016/j.neunet.2005.10.007] [Medline: 16483741]
5. Spelt L, Andersson B, Nilsson J, Andersson R. Prognostic models for outcome following liver resection for colorectal cancer metastases: a systematic review. Eur J Surg Oncol 2012 Jan;38(1):16-24. [doi: 10.1016/j.ejso.2011.10.013] [Medline: 22079259]
6. Jalalian A, Mashohor SB, Mahmud HR, Saripan MI, Ramli AR, Karasfi B. Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. Clin Imaging 2013;37(3):420-426. [doi: 10.1016/j.clinimag.2012.09.024] [Medline: 23153689]
7. Tucker SR, Speer SA, Peters S. Development of an explanatory model of sexual intimacy following treatment for localised prostate cancer: a systematic review and meta-synthesis of qualitative evidence. Soc Sci Med 2016 Aug;163:80-88. [doi: 10.1016/j.socscimed.2016.07.001] [Medline: 27421074]
8. Ray PP, Dash D, De D. A systematic review of wearable systems for cancer detection: current state and challenges. J Med Syst 2017 Oct 2;41(11):180. [doi: 10.1007/s10916-017-0828-y] [Medline: 28971278]
9. Sadoughi F, Kazemy Z, Hamedan F, Owji L, Rahmanikatigari M, Azadboni TT. Artificial intelligence methods for the diagnosis of breast cancer by image processing: a review. Breast Cancer (Dove Med Press) 2018;10:219-230 [FREE Full text] [doi: 10.2147/BCTT.S175311] [Medline: 30555254]
10. Marka A, Carter JB, Toto E, Hassanpour S. Automated detection of nonmelanoma skin cancer using digital images: a systematic review. BMC Med Imaging 2019 Feb 28;19(1):21 [FREE Full text] [doi: 10.1186/s12880-019-0307-7] [Medline: 30819133]
11. Waltman L, van Eck NJ, Noyons EC. A unified approach to mapping and clustering of bibliometric networks. J Informetr 2010 Oct;4(4):629-635. [doi: 10.1016/j.joi.2010.07.002]
12. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010 Aug;84(2):523-538 [FREE Full text] [doi: 10.1007/s11192-009-0146-3] [Medline: 20585380]

XSL•FO
RenderX

13. Waltman L, van Eck NJ. A smart local moving algorithm for large-scale modularity-based community detection. Eur Phys J B 2013 Nov 13;86(11):471. [doi: 10.1140/epjb/e2013-40829-0]
14. van Eck NJ, Waltman L. Visualizing bibliometric networks. In: Ding Y, Rousseau R, Wolfram D, editors. Measuring Scholarly Impact: Methods and Practice. New York City: Springer; 2014:285-320.
15. de Araújo CC, Pedron CD, Picoto WN. What's behind CRM research? A bibliometric analysis of publications in the CRM research field. J Relatsh Mark 2018 Apr 2;17(1):29-51. [doi: 10.1080/15332667.2018.1440139]
16. Huang A. Similarity Measures for Text Document Clustering. In: Proceedings of the New Zealand Computer Science Research Student Conference. 2008 Presented at: NZCSRSC'08; April 28-29, 2008; Christchurch, New Zealand.
17. Li Y, Rapkin B, Atkinson TM, Schofield E, Bochner BH. Leveraging latent dirichlet allocation in processing free-text personal goals among patients undergoing bladder cancer surgery. Qual Life Res 2019 Jun;28(6):1441-1455 [FREE Full text] [doi: 10.1007/s11136-019-02132-w] [Medline: 30798421]
18. Valle D, Albuquerque P, Zhao Q, Barberan A, Fletcher Jr RJ. Extending the latent dirichlet allocation model to presence/absence data: a case study on North American breeding birds and biogeographical shifts expected from climate change. Glob Chang Biol 2018 Nov;24(11):5560-5572. [doi: 10.1111/gcb.14412] [Medline: 30058746]
19. Chen C, Zare A, Trinh HN, Omotara GO, Cobb JT, Lagaunne TA. Partial membership latent dirichlet allocation for soft image segmentation. IEEE Trans Image Process 2017 Dec;26(12):5590-5602. [doi: 10.1109/TIP.2017.2736419] [Medline: 28792897]
20. Lu HM, Wei CP, Hsiao FY. Modeling healthcare data using multiple-channel latent Dirichlet allocation. J Biomed Inform 2016 Apr;60:210-223 [FREE Full text] [doi: 10.1016/j.jbi.2016.02.003] [Medline: 26898516]
21. Gross A, Murthy D. Modeling virtual organizations with latent Dirichlet allocation: a case for natural language processing. Neural Netw 2014 Oct;58:38-49. [doi: 10.1016/j.neunet.2014.05.008] [Medline: 24930023]
22. Miller KD, Nogueira L, Mariotto AB, Rowland JH, Yabroff KR, Alfano CM, et al. Cancer treatment and survivorship statistics, 2019. CA Cancer J Clin 2019 Jun 11 (epub ahead of print)(forthcoming)(forthcoming) [FREE Full text] [doi: 10.3322/caac.21565] [Medline: 31184787]
23. Liu Q, Chen Q, Shen J, Wu H, Sun Y, Ming WK. Data analysis and visualization of newspaper articles on thirdhand smoke: a topic modeling approach. JMIR Med Inform 2019 Jan 29;7(1):e12414 [FREE Full text] [doi: 10.2196/12414] [Medline: 30694199]
24. Tang C, Zhou L, Plasek J, Rozenblum R, Bates D. Comment topic evolution on a cancer institution's Facebook page. Appl Clin Inform 2017 Aug 23;8(3):854-865 [FREE Full text] [doi: 10.4338/ACI-2017-04-RA-0055] [Medline: 28832069]
25. Nzali MD, Bringay S, Lavergne C, Mollevi C, Opitz T. What patients can tell us: topic analysis for social media on breast cancer. JMIR Med Inform 2017 Jul 31;5(3):e23 [FREE Full text] [doi: 10.2196/medinform.7779] [Medline: 28760725]
26. Westmaas JL, McDonald BR, Portier KM. Topic modeling of smoking- and cessation-related posts to the american cancer society's cancer survivor network (CSN): implications for cessation treatment for cancer survivors who smoke. Nicotine Tob Res 2017 Aug 1;19(8):952-959. [doi: 10.1093/ntr/ntx064] [Medline: 28340059]
27. Thackeray R, Burton SH, Giraud-Carrier C, Rollins S, Draper CR. Using Twitter for breast cancer prevention: an analysis of breast cancer awareness month. BMC Cancer 2013 Oct 29;13:508 [FREE Full text] [doi: 10.1186/1471-2407-13-508] [Medline: 24168075]
28. Huang Z, Dong W, Ji L, Gan C, Lu X, Duan H. Discovery of clinical pathway patterns from event logs using probabilistic topic models. J Biomed Inform 2014 Feb;47:39-57 [FREE Full text] [doi: 10.1016/j.jbi.2013.09.003] [Medline: 24076435]
29. Li D, Wang Z, Wang L, Sohn S, Shen F, Murad MH, et al. A text-mining framework for supporting systematic reviews. Am J Inf Manag 2016 Nov;1(1):1-9 [FREE Full text] [doi: 10.11648/j.infomgmt.20160101.11] [Medline: 29071308]

Abbreviations

AI: artificial intelligence
ANN: artificial neural network
AUC: area under the curve
CT: computed tomography
HPV: human papilloma virus
LDA: Latent Dirichlet Allocation
MRI: magnetic resonance imaging
PSA: prostate specific antigen
PT: prothrombin time
RALP: robot assisted laparoscopic prostatectomy
RARC: remittance advice remark code
RARP: robotic-assisted radical prostatectomy
SBRT: stereotactic body radiation therapy
SVM: support vector machine
TORS: transoral robotic surgery

XSL•FO
RenderX

WOS: Web of Science

Edited by G Eysenbach; submitted 16.04.19; peer-reviewed by C Short, F Coelho; comments to author 25.06.19; revised version received 27.06.19; accepted 19.07.19; published 15.09.19.

Please cite as:
Tran BX, Latkin CA, Sharafeldin N, Nguyen K, Vu GT, Tam WWS, Cheung NM, Nguyen HLT, Ho CSH, Ho RCM
Characterizing Artificial Intelligence Applications in Cancer Research: A Latent Dirichlet Allocation Analysis
JMIR Med Inform 2019;7(4):e14401
URL: https://medinform.jmir.org/2019/4/e14401
doi:10.2196/14401
PMID:31573929

©Bach Xuan Tran, Carl A Latkin, Noha Sharafeldin, Katherina Nguyen, Giang Thu Vu, Wilson WS Tam, Ngai-Man Cheung, Huong Lan Thi Nguyen, Cyrus SH Ho, Roger CM Ho. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 15.09.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Fast Prediction of Deterioration and Death Risk in Patients With Acute Exacerbation of Chronic Obstructive Pulmonary Disease Using Vital Signs and Admission History: Retrospective Cohort Study

Mi Zhou[1*], MD; Chuan Chen[2*], PhD; Junfeng Peng[2], MSc; Ching-Hsing Luo[2], MD, PhD; Ding Yun Feng[3], MD; Hailing Yang[3], MD; Xiaohua Xie[2], PhD; Yuqi Zhou[3], MD, PhD

[1]Surgical Intensive Care Unit, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

[2]School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

[3]Department of Respiratory and Critical Care Medicine, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Yuqi Zhou, MD, PhD
Department of Respiratory and Critical Care Medicine
The Third Affiliated Hospital of Sun Yat-sen University
No 600 Tianhe Road
Tianhe District
Guangzhou, 510640
China
Phone: 86 13533489943
Email: zhouyuqi@mail.sysu.edu.cn

Abstract

Background: Chronic obstructive pulmonary disease (COPD) has 2 courses with different options for medical treatment: the acute exacerbation phase and the stable phase. Stable patients can use the Global Initiative for Chronic Obstructive Lung Disease (GOLD) to guide treatment strategies. However, GOLD could not classify and guide the treatment of acute exacerbation as acute exacerbation of COPD (AECOPD) is a complex process.

Objective: This paper aimed to propose a fast severity assessment and risk prediction approach in order to strengthen monitoring and medical interventions in advance.

Methods: The proposed method uses a classification and regression tree (CART) and had been validated using the AECOPD inpatient's medical history and first measured vital signs at admission that can be collected within minutes. We identified 552 inpatients with AECOPD from February 2011 to June 2018 retrospectively and used the classifier to predict the outcome and prognosis of this hospitalization.

Results: The overall accuracy of the proposed CART classifier was 76.2% (83/109 participants) with 95% CI 0.67-0.84. The precision, recall, and F-measure for the mild AECOPD were 76% (50/65 participants), 82% (50/61 participants), and 0.79, respectively, and those with severe AECOPD were 75% (33/44 participants), 68% (33/48 participants), and 0.72, respectively.

Conclusions: This fast prediction CART classifier for early exacerbation detection could trigger the initiation of timely treatment, thereby potentially reducing exacerbation severity and recovery time and improving the patients' health.

(JMIR Med Inform 2019;7(4):e13085) doi:10.2196/13085

KEYWORDS

chronic obstructive pulmonary disease; clinical decision support systems; health risk assessment

XSL•FO
RenderX

Introduction

Background

Chronic obstructive pulmonary disease (COPD) is characterized by incomplete reversible airflow obstruction. Patients with COPD may experience exacerbations of the disease, which are associated with significant morbidity and mortality as well as reduced quality of life. COPD is a serious long-term condition that progressively restricts airflow from the lungs and imposes a significant burden on patient's daily lives [1]. Currently, it is the fourth leading cause of death in the world but is projected to be the third by 2030 [2-4]. As one of the most common and frequently occurring diseases, COPD has 2 different courses: the acute exacerbation phase and the stable phase. An acute exacerbation of COPD (AECOPD) has been described as an acute worsening of respiratory symptoms associated with a variable degree of physiological deterioration [5]. Sudden deterioration because of any cause requires critical medical care and may require hospitalization. Previous studies have shown that early intervention on these COPD patients decreases morbidity of acute exacerbation and mortality [6].

Since 2001, according to the Global Initiative for Chronic Obstructive Lung Disease (GOLD) guideline, patients with stable COPD have been classified as mild, moderate, severe, and extremely severe depending on lung function. The 2011 GOLD guideline has been revised to divide patients with COPD into grades A, B, C, and D. This classification method has been improved several times and is still in use today, which is based on lung function, frequency of acute exacerbations, symptom scores, and risk factors [7]. However, AECOPD patients are highly heterogeneous. According to the differences in basic conditions, causes, and complications, the acute exacerbations of different patients in the same grade may be different, and even the 2 consecutive acute exacerbations of the same patient may be very different. Patients with mild acute exacerbation may be discharged after several days of treatment; however, patients with severe acute exacerbation may require longer hospital stay, higher costs, ICU admission, and even mechanical ventilation. In the worst case, a small number of patients may eventually die without remission. Therefore, it is important to assess the severity of acute exacerbations in patients with COPD, which can determine what treatments are needed to improve prognosis and reduce mortality [8,9]. However, there is currently no consensus on the assessment of the severity of acute exacerbations.

There are some attempts to predict the course of disease using machine learning in general and deep learning models in particular. Most of the studies analyzed the correlation between clinical treatment and prognosis. Amalakuhan et al [10] took advantage of random forest (RF) algorithm to research which patients were at high risk for multiple COPD exacerbations and hospital readmission within a single year. The study included 60 indicators in 106 patients, such as medical history, general conditions, and medication, and the prediction accuracy is 0.72. However, because patients have many influencing factors outside the hospital, such as weather changes, environmental pollution, treatment compliance, and pathogen epidemics, this may affect the accuracy of prediction. Yang et al [11] used 3 methods to predict the risk of 30-day readmission of patients. The study used a public database with a total of 323,813 patients and 100 features, and COPD patients were a subgroup among them. The precision rate was 0.257, and the recall rate was 0.786. Zheng et al [12] proposed a hesitant fuzzy linguistic complex proportional assessment method to solve the decision-making problems under hesitant fuzzy linguistic environment. The study assessed the severity of COPD patients by outpatient doctors' description of patient symptoms and risk factors, but it was difficult to verify the accuracy of the evaluation because of lack of follow-up and prognostic data. Swaminatha et al [13] collected vital signs, symptoms, and comorbidities data of patients with COPD. The study used physician opinion in a statistically and clinically comprehensive set of patient cases to train a supervised prediction algorithm. After 2400 training sessions, the gradient-enhanced RF algorithm was 88% identical to the physician's judgment in 101 validated cases. However, the study also lacked follow-up outcome data, making it difficult to verify whether the subjective judgment of the physician met the objective prognosis.

Objectives

Although scholars have worked on predicting the severity of AECOPD with various machine learning algorithms, none of the abovementioned studies examined the fast severity assessment approach, which only requires the patient's vital signs and admission history data that can be collected within minutes after admission. In this study, we propose a fast severity assessment and risk prediction approach by exploring the usefulness of the classification and regression tree (CART) for fast predicting the severity of AECOPD once the patient is admitted to the hospital. CART as a decision tree algorithm was introduced by Leo Breiman in 1985, which is successfully used for classification or regression predictive modeling problems [14]. The proposed fast assessment system can help the doctors to obtain the severity assessment of the patients quickly within minutes after admission. The fast prediction CART classifier is a promising research tool for the identification of at-risk populations with COPD. Therefore, it is necessary to establish a rapid classification method to predict the outcome and prognosis of patients with AECOPD.

Methods

Data Acquisition

The data of AECOPD patients were obtained from the Department of Respiratory and Critical Care Medicine of the Third Affiliated Hospital, Sun Yat-sen University (TAHSYU). TAHSYU is a comprehensive third-grade class-A hospital directly managed by the National Health Commission of the People's Republic of China. We searched for medical records of all inpatients from 2011 to 2018, screening out patients with a major diagnosis of AECOPD by using International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) code for AECOPD (J44.100, J44.101). Patients needed to have a pulmonary function test record with forced expiratory volume in 1 second/forced vital capacity <0.7, and the main complaint in this hospitalization included a description

XSL•FO
RenderX

of increased cough or shortness of breath. We excluded the patients who were discharged without medical advice or had missing variables. Finally, 552 hospitalized AECOPD patients were included. For statistical purposes, the triage was labeled as mild group and severe group according to the situation of the patient during hospitalization. Here, mild group means the patient was stable and no intensive care was required, eventually got better, and was discharged. Severe group means the patient had a notable deterioration, needed intensive care, was dead, dying, incurable, and automatically discharged. The distribution of AECOPD patients with mild and severe symptoms is shown in Table 1. The research was performed under the guidance of the TAHSYU Institutional Review Board, protocol #2019-02-334-01.

Table 1. Distribution of mild and severe groups in patients with acute exacerbation of chronic obstructive pulmonary disease.

Characteristic	Mild group (low risk)	Severe group (high risk)
Number of cases, n (%)	308 (55.8)	244 (44.2)
Outcome, n (%)		
Dead	0 (0.0)	24 (9.8)
Deteriorating discharge	0 (0.0)	50 (20.4)
Sex, n (%)		
Male	240 (77.9)	201 (82.4)
Female	68 (22.1)	43 (17.6)
Smoking history, n (%)	215 (69.8)	167 (68.4)
Age (year), mean (SD)	75.06 (7.94)	76.66 (9.49)
Number of hospitalizations, mean (SD)	4.21 (4.07)	5.97 (6.87)
Days in hospital, mean (SD)	8.01 (2.64)	73.96 (195.21)
Costs (RMB Yuan), mean (SD)	9428 (2921)	68,860 (98,109)

Data Analysis

The GOLD guideline is only for the classification of patients in stable phase, and there is no consistent classification for patients with acute exacerbations. Some scholars have proposed a 2-axis and 4-group classification by considering the pathobiological and clinical heterogeneity of AECOPD [15], but it takes a long time to get results and also requires more clinical validation. Thus, we propose a fast assessment indicator system to make it more reasonable and practical by the advice of the clinician.

One of the important missions is the variable selection in the process of fast assessing the severity of the COPD. In this paper, the process of predictor selection is shown as follows:

- Step 1: Find some predictors from the perspective of system engineering;
- Step 2: Make sure the predictors can be collected quickly after the patient is admitted to the hospital;
- Step 3: Verify the reasonability of the above predictors from the clinical experience of professional pulmonary physicians;
- Step 4: Predictors with too many missing values (more than 10% over 552 rows of records) are discarded directly to avoid inaccurate predictions;
- Step 5: Laboratory testing data are not included because most laboratory testing results are not available within 10 min, such as blood gas analysis, sputum culture, and so on;
- Step 6: Text-based features like common chief complaints that need to be processed by natural language processing are left for subsequent processing and are not included now.

From the abovementioned steps, we can establish a fast assessment system that only includes these 7 variables at the beginning of admission. In particular, this indicator can be obtained within minutes after admission.

1. Respiratory rate (RR): RR is one of the most important predictors of the COPD, and excessive breathing rate is the main factor causing the patients feeling anxious with the loss of physical ability [16]. Normal respiration rate is between 12 and 18 breaths per minute. Typical COPD patients describe excessive breathing rate as a sense of shortness of breath, wheezing, or needing great effort to breathe.
2. Systolic blood pressure (SBP) and diastolic blood pressure (DBP): Blood pressure is usually expressed in terms of SBP over DBP and is measured in millimeters of mercury (mmHg), reflects the stability of the blood circulation. Blood pressure in patients with severe COPD may be affected by hypoxemia or cardiac insufficiency.
3. Pulse rate (PR): Pulse is also one of the important indicators for doctors to diagnose COPD. The PR changes obviously when the patient is in critical condition. Therefore, measuring PR is an indispensable examination item for patients.
4. Number of hospitalizations (NOH): NOH is defined as the total number of hospitalizations of patients at TAHSYU. NOH is proportional to the severity of the disease. Generally, the greater the number of admissions, the more severe the COPD patient will be.
5. Temperature (TEMP): Body temperature is an important indicator of body metabolism, which is dynamically balanced within a certain range. COPD patients often

XSL•FO
RenderX

develop fever because of inflammation. Especially, the measurement of the patients' TEMP is relatively simple and rapid.

6. Smoking: Define a patient who has smoked for 6 consecutive months as having a history of smoking. Smoking is one of the most common risk factors of COPD and will worsen the severity of COPD.

Mode Selection

CART is a nonparametric statistical procedure containing classification procedure and regression procedure. It is formed by using a set of if-then-else logical conditions to assign an unknown vector of feature values (or predictors) to a predefined class or category. CART methodology has been increasingly applied to health sciences and clinical research and has been applied to a much lesser extent in COPD condition monitoring. Algorithms for constructing a CART usually work top down, by choosing a variable at each step that best splits the set of items [17]. Gini impurity, information gain, and variance reduction are often applied to each candidate subset, and the resulting values are combined (eg, averaged) to provide a measure of the quality of the split [18].

The results calculated by CART techniques are straightforward to interpret. Compared with the black box model, such as neural network algorithm, CART is a highly interpretive model. Compared with the white box model, such as linear regression, CART does not need data to satisfy the linear priori hypothesis. In addition, CART analysis has the statistical advantage of being a nonparametric technique that does not invoke assumptions about the functional form of the data. Furthermore, CART can process multiclass problem easily. Finally, CART is good at processing categorical and missing features easily and nonlinear test efficiently.

Classification Using a Classification and Regression Tree

At this stage, 7 predictors collected from 552 COPD patients' records included NOH, smoking history, RR per minute, TEMP, PR, SBP, and DBP. From the available dataset, each of the N observations is denoted by the 2-tuple, (x, y), where $x \in \{x_1,x_2,...,x_7\}$ is the vector containing all the 7 features. $y \in \{1,2\}$ represents the categories of low risk and high risk.

In the process of mode training, we use 80% of the observations for model training and the remaining 20% for mode validation. A cross-validated grid-search approach is employed to tune the hyperparameters of the CART. To avoid overtraining the CART, we first estimate the optimal depth of the CART. The tree depth is defined as the maximum number of branches (a branch joins 2 nodes) on the path from any leaf node to the root node. The tree construction algorithm is described as follows: (1) search the best predictor as the root node of the tree according to gini index. The node is then split using the best predictor to create 2 leaf nodes; for multivariate classification, all variables are evaluated by gini values to find the variable with the minimum gini values as the root node of the CART. Gini index is usually selected as the measurement for the classification problem to reduce a chosen global measure of impurity for the tree; the messier the category overall, the bigger the gini index; (2) if the node is no longer separable, then the node is stored as a leaf node; a completely pure node contains only instances from 1 class; (3) the splitting process is repeated (binary splitting) until all leaf nodes reside no greater than the predefined depth from the root node for all existing leaf nodes [19,20]; (4) create left and right subtrees recursively.

In the process of the model testing, we measure the classification performance of CART model by precision, recall, and F-Measure. We check the prediction performance of the model on the training set and the test set to choose the best model by avoiding overfitting and underfitting. We implemented CART classifier on the development platform of R 3.5.1. R is available as free software in source code form. It was originally developed at Bell Laboratories by John Chambers and colleagues, which provides a wide variety of statistical and graphical techniques and is highly extensible.

Results

The Accuracy of Classifier

Precision, recall, and F-measure are the measures widely used in the field of information retrieval and statistical classification to evaluate the quality of results. Precision is defined as the ratio of the correct number predicted by the model to the actual correct number. Recall is defined as the ratio of the actual correct number to the correct number predicted by the model. F-measure is the weighted average of precision and recall. The larger the parameters are, the better the prediction performance will be. In particular, 1 is the ideal state.

The overall accuracy on the test dataset of the proposed CART classifier was 76.2%, with 95% CI 0.67-0.84. The evaluation of the fast prediction CART classifier is shown in Table 2. The receiver operating characteristic curve of the CART classifier is shown in Figure 1. The optimal tipping point is 1.50 (0.82, 0.69). The area under the curve is 0.75.

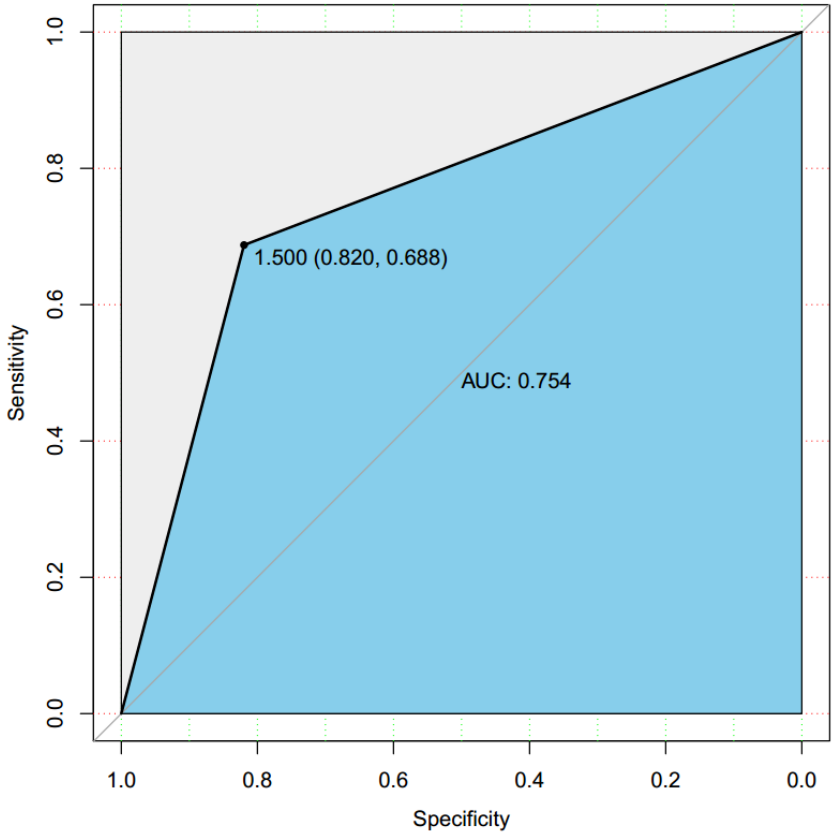
Currently, the proposed CART classifier can achieve the same performance on a similar test dataset. However, to improve the generalization of the model, it is necessary to provide a wide variety of training samples to gain more comprehensive knowledge. By working with external data sources, we can provide a more comprehensive set of training for the model, allowing the model to gain more comprehensive knowledge and continuously improve predictive performance. In addition, we use the K-fold cross verification method to estimate the depth of the CART tree.

Table 2. Evaluation of fast prediction classification and regression tree classifier on test dataset. The overall accuracy was 76.2%.

Group	Precision, %	Recall, %	F-measure
Mild group (low risk)	76	82	0.79
Severe group (high risk)	75	68	0.72

XSL•FO
RenderX

Figure 1. Receiver operating characteristic curve in the classification and regression tree classifier. AUC: area under the curve.



The Importance of Variables

To understand the contribution of each predictor to the CART model, we computed the variable importance in the tree model. Table 3 shows the variable importance of the fast prediction CART classifier. The predictor is more important if the value on the x-axis is bigger. We can find that RR per minute, SBP, PR, DBP, NOH, TEMP, and smoking history (Smoking) were important predictors. The RR per minute reflects the severity of dyspnea and may be a good indicator of prognosis [21]. Other vital signs also reflect the overall condition of the patient. For example, elevated TEMP may mean more serious infections, increased heart rate may represent severe dyspnea, or heart failure. Hypertension is one of the most common comorbidities in COPD patients; all of the above are related to the prognosis of the patient [22,23].

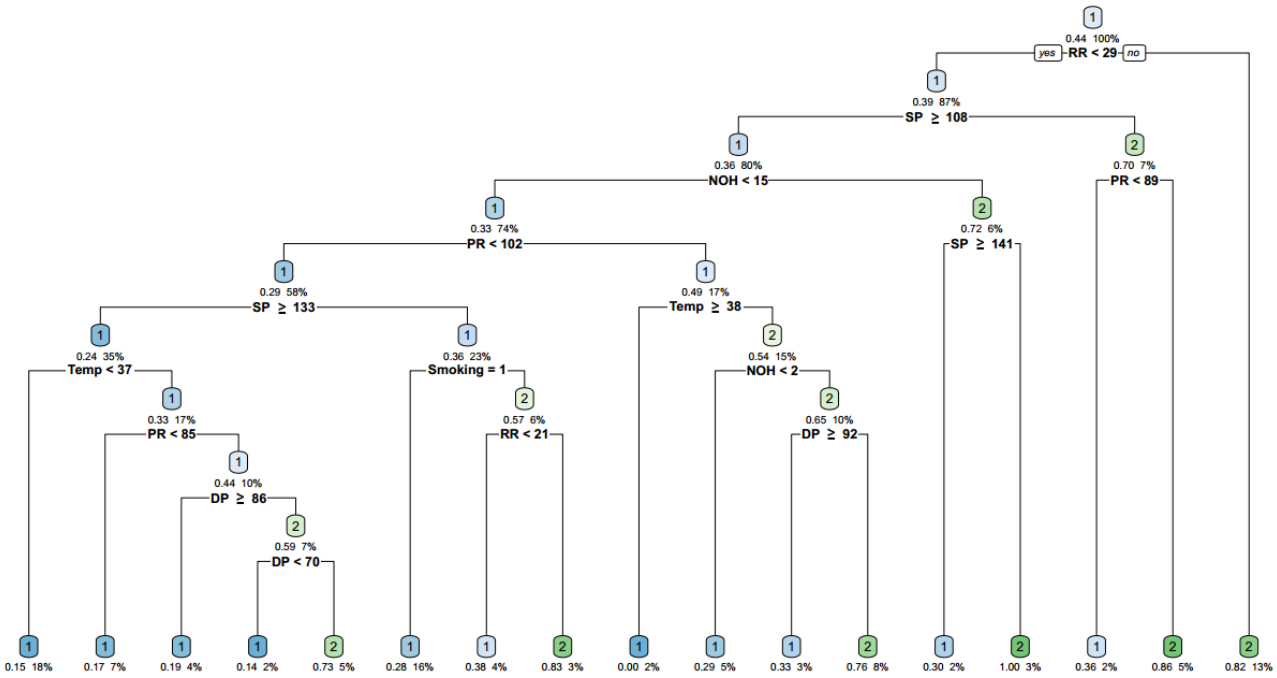
This CART model for early detection could trigger the initiation of timely treatment, thereby potentially reducing exacerbation severity and recovery time and improving the patients' health. Figure 2 is an illustration of a CART constructed from one run. A leaf node in the tree, 1 represents a low-risk health prediction and 2 represents high-risk one. Each decision node represents the corresponding feature used and the decision threshold. At each decision node, the left branch is chosen if the feature is less than the threshold value. RR is the root node of the tree. The closer the root node is, the more important the feature is. We can see from Figure 2 that if the RR of a patient with COPD is greater than 29 breaths per minute, the patient is a high-risk patient; otherwise, it can be judged according to other features, and so on.

Table 3. The relative importance of each variable to the prediction, with respiratory rate as 100%.

Variables	Relative importance (%)
Respiratory rate	100.0
Systolic blood pressure	85.8
Pulse rate	68.6
Diastolic blood pressure	66.1
Number of hospitalizations	59.2
Temperature	49.0
Smoking	22.8

XSL•FO
RenderX

Figure 2. The classification and regression tree constructed from one run. DP: diastolic pressure; NOH: number of hospitalizations; PR: pulse rate; RR: respiratory rate; SP: systolic pressure; Temp: temperature.



Discussion

Clinical Significance

In this study, we investigated approaches to fast predict the severity and risk of patients with AECOPD within minutes after admission. Fast predicting can serve as a useful tool for physicians to assess the risk of deterioration, thereby strengthening monitoring and medical interventions in advance. CART classifier proposed in this paper predicts 76.2% of instances correctly.

The clinical presentation and disease progression of AECOPD patients are significantly heterogeneous, that means the severity of patients is quite different. Severe patients may need to be admitted to the ICU with systemic glucocorticoids, broad-spectrum antibiotics, or even mechanical ventilation [24,25]. Therefore, it is important to judge the severity and prognosis of patients with AECOPD early. However, the GOLD guideline is only for the treatment of patients in stable phase, and there is no consistent classification for patients with acute exacerbations. Some scholars have proposed a 2-axis and 4-group classification by considering the pathobiological and clinical heterogeneity of AECOPD [15], but it also requires more clinical validation.

Machine learning provides a powerful tool for the classification and prediction of COPD patients, but the pathogenesis of COPD is not completely clear, the course of the disease is extremely complex. Therefore, the data need to be properly selected and analyzed to obtain more accurate and credible conclusions. The patients in this study had an exact outcome and were hospitalized for standardized treatment, effectively reducing the impact of factors outside the hospital. The selection of features and the time span of observation are also very important, and there are many factors that influence the prognosis of patients with AECOPD. Multiple studies have attempted to predict risk factors that affect mortality and readmission rates in AECOPD patients, such as acute physiology and chronic health evaluation scores, C-reactive protein, blood carbon dioxide partial pressure, and blood urea nitrogen [26,27]. Obviously, incorporating more features and increasing observation time is beneficial to improve the accuracy of the forecast, but it also increases the cost and complexity of the assessment. If a large number of examinations and several days of time are needed for prediction, the clinical significance will become very poor.

The 7 indicators we selected are simple, fast, noninvasive, and objective. Measurements only require watches, sphygmomanometers, and thermometers. In clinical work, usually the nurses measure vital signs, ask for general information, and register after admission, which takes 7 to 10 min. If we only specifically acquire the 7 indicators and use an electronic sphygmomanometer and an infrared thermometer, the time can be shortened to less than 3 min. Doctors, nurses, and even trainees can quickly grasp the assessment method. This is very helpful in assessing the severity and risk of patients before the senior physician arrives or in scheduling the intensive care unit resources faster. Although this study included hospitalized patients, it can be applied to outpatients or even patients for self-assessment because features can be easily and quickly obtained.

Limitations

In addition to the 7 indicators of this study, there are some other indicators that can be quickly obtained and may be helpful in predicting prognosis. Cough, dyspnea, and increased sputum are criteria for judging acute exacerbations, and the severity of these symptoms correlates with prognosis. However, most of the hospitalized patients in this study have these symptoms to

XSL•FO
RenderX

varying degrees, and it is difficult to quantify the changes and severity of these symptoms in the medical records. Some studies use a sound monitoring system to continuously record a patient's cough and perform an automated analysis to assess the severity of cough [28]. However, it is still difficult to guide clinical practice in the short term.

Complications, such as acute heart failure and diabetes, may significantly increase mortality in patients with AECOPD [29]. However, for newly hospitalized patients, multiple tests and several days may be required to diagnose the comorbidities, which limits the rapid judgment of prognosis. For patients with repeated hospitalizations and chronic comorbidities with a clear history, this may be more meaningful. We will use text mining methods to improve data and further study the impact of chronic comorbidities on the prognosis of patients with AECOPD.

There are also some shortcomings in this study. Due to the single-center study, the number of cases is small. The research also lacks oxygen saturation data, which is a simple, noninvasive indicator of oxygenation in patients. This is because in the past few years, not all patients, especially those with mild symptoms, have routinely measured blood oxygen saturation. If oxygen saturation is included, the amount of data will be significantly reduced, while causing bias. Now, with the popularity of portable finger oximeters, the vast majority of patients measure blood oxygen saturation on admission, which can be used in subsequent studies.

In summary, this study shows that the use of machine learning methods to analyze the vital signs and other indicators of newly hospitalized patients may help clinicians to judge the severity of patients more quickly, so as to carry out early medical intervention for patients with severe AECOPD. In spite of this, the results are still valid when some of the variables are not included, as this study is not a causal analysis, but an exploratory data analysis. The proposed model is generic enough to cope with similar medical scenarios, provided that these data can be obtained as long as COPD patients are hospitalized.

Conclusions

In this study, we developed a fast severity assessment and risk prediction approach, which only requires the patient's vital signs and admission history data that can be collected within minutes, and showed that it can rapidly predict the severity of COPD patients. The overall accuracy of the proposed CART classifier is 76.2% with 95% CI 0.67-0.84. It is concluded that CART classifier can be used as a forecasting tool for COPD inpatients. As CART is a nonlinear system, it is found that its performance is better than previous classifiers or regression techniques. Further work can be done on similar lines by adding predictors, or optimizing the classifier parameters, or using other fusion learning algorithms, such as RF [30].

Acknowledgments

This work was supported by Sun Yat-sen University, China, under Scientific Initiation Project No. 67000-18821109 for High-level Experts.

Authors' Contributions

MZ and YZ participated in the design of clinical data and feature selection in this study. CC and JP designed a clinically meaningful and effective classifier algorithm, and XX provided constructive algorithm guidance. DF and HY collected and compiled clinical data. MZ, JP, and CC drafted the manuscript. CL reviewed and guided the revision of the paper. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Seemungal TA, Donaldson GC, Paul EA, Bestall JC, Jeffries DJ, Wedzicha JA. Effect of exacerbation on quality of life in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 1998 May;157(5 Pt 1):1418-1422. [doi: 10.1164/ajrccm.157.5.9709032] [Medline: 9603117]
2. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. Lancet 2012 Dec 15;380(9859):2095-2128. [doi: 10.1016/S0140-6736(12)61728-0] [Medline: 23245604]
3. Lopez AD, Shibuya K, Rao C, Mathers CC, Hansell AL, Held LS, et al. Chronic obstructive pulmonary disease: current burden and future projections. Eur Respir J 2006 Feb;27(2):397-412 [FREE Full text] [doi: 10.1183/09031936.06.00025805] [Medline: 16452599]
4. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med 2006 Nov;3(11):e442 [FREE Full text] [doi: 10.1371/journal.pmed.0030442] [Medline: 17132052]
5. Seemungal TA, Donaldson GC, Bhowmik A, Jeffries DJ, Wedzicha JA. Time course and recovery of exacerbations in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2000 May;161(5):1608-1613. [doi: 10.1164/ajrccm.161.5.9908022] [Medline: 10806163]

XSL•FO
RenderX

6. Wilkinson TM, Donaldson GC, Hurst JR, Seemungal TA, Wedzicha JA. Early therapy improves outcomes of exacerbations of chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2004 Jun 15;169(12):1298-1303. [doi: 10.1164/rccm.200310-1443OC] [Medline: 14990395]
7. Global Initiative for Chronic Obstructive Lung Disease - GOLD. 2018. Global Strategy for the Diagnosis, Management and Prevention of Chronic Obstructive Pulmonary Disease: 2019 Report URL: https://goldcopd.org/wp-content/uploads/2018/11/GOLD-2019-v1.7-FINAL-14Nov2018-WMS.pdf [accessed 2019-05-28]
8. Chandra D, Tsai C, Camargo CA. Acute exacerbations of COPD: delay in presentation and the risk of hospitalization. COPD 2009 Apr;6(2):95-103. [doi: 10.1080/15412550902751746] [Medline: 19378222]
9. Seemungal TA, Wedzicha JA. Acute exacerbations of COPD: the challenge is early treatment. COPD 2009 Apr;6(2):79-81. [doi: 10.1080/15412550902806011] [Medline: 19378218]
10. Amalakuhan B, Kiljanek L, Parvathaneni A, Hester M, Cheriyath P, Fischman D. A prediction model for COPD readmissions: catching up, catching our breath, and improving a national problem. J Community Hosp Intern Med Perspect 2012;2(1):99-115 [FREE Full text] [doi: 10.3402/jchimp.v2i1.9915] [Medline: 23882354]
11. Yang C, Delcher C, Shenkman E, Ranka S. Predicting 30-Day All-Cause Readmissions From Hospital Inpatient Discharge Data. In: Proceedings of the 18th International Conference on e-Health Networking, Applications and Services. 2016 Presented at: Healthcom'16; September 14-16, 2016; Munich, Germany p. 1-6. [doi: 10.1109/HealthCom.2016.7749452]
12. Zheng Y, Xu Z, He Y, Liao H. Severity assessment of chronic obstructive pulmonary disease based on hesitant fuzzy linguistic COPRAS method. Appl Soft Comput 2018 Aug;69:60-71. [doi: 10.1016/j.asoc.2018.04.035]
13. Swaminathan S, Qirko K, Smith T, Corcoran E, Wysham NG, Bazaz G, et al. A machine learning approach to triaging patients with chronic obstructive pulmonary disease. PLoS One 2017;12(11):e0188532 [FREE Full text] [doi: 10.1371/journal.pone.0188532] [Medline: 29166411]
14. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. Boca Raton, Florida, United States: CRC Press; 1985.
15. Lopez-Campos JL, Agustí A. Heterogeneity of chronic obstructive pulmonary disease exacerbations: a two-axes classification proposal. Lancet Respir Med 2015 Sep;3(9):729-734. [doi: 10.1016/S2213-2600(15)00242-8] [Medline: 26165134]
16. Jones PW, Nadeau G, Small M, Adamek L. Characteristics of a COPD population categorised using the GOLD framework by health status and exacerbations. Respir Med 2014 Jan;108(1):129-135 [FREE Full text] [doi: 10.1016/j.rmed.2013.08.015] [Medline: 24041746]
17. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Ann Behav Med 2003 Dec;26(3):172-181. [doi: 10.1207/S15324796ABM2603_02] [Medline: 14644693]
18. Rokach L, Maimon O. Top-down induction of decision trees classifiers—a survey. IEEE Trans Syst, Man, Cybern 2005 Nov;35(4):476-487. [doi: 10.1109/TSMCC.2004.843247]
19. Loh W. Classification and regression trees. WIREs Data Mining Knowl Discov 2011 Jan 6;1(1):14-23. [doi: 10.1002/widm.8]
20. Deconinck E, Zhang MH, Coomans D, Heyden YV. Classification tree models for the prediction of blood-brain barrier passage of drugs. J Chem Inf Model 2006;46(3):1410-1419. [doi: 10.1021/ci050518s] [Medline: 16711761]
21. Flattet Y, Garin N, Serratrice J, Perrier A, Stirnemann J, Carballo S. Determining prognosis in acute exacerbation of COPD. Int J Chron Obstruct Pulmon Dis 2017;12:467-475 [FREE Full text] [doi: 10.2147/COPD.S122382] [Medline: 28203070]
22. Terzano C, Conti V, di Stefano F, Petroianni A, Ceccarelli D, Graziani E, et al. Comorbidity, hospitalization, and mortality in COPD: results from a longitudinal study. Lung 2010 Aug;188(4):321-329. [doi: 10.1007/s00408-009-9222-y] [Medline: 20066539]
23. Sethi S, Murphy TF. Infection in the pathogenesis and course of chronic obstructive pulmonary disease. N Engl J Med 2008 Nov 27;359(22):2355-2365. [doi: 10.1056/NEJMra0800353] [Medline: 19038881]
24. Wedzicha JA, Miravitlles M, Hurst JR, Calverley PM, Albert RK, Anzueto A, et al. Management of COPD exacerbations: a European Respiratory Society/American Thoracic Society guideline. Eur Respir J 2017 Mar;49(3):1600791 [FREE Full text] [doi: 10.1183/13993003.00791-2016] [Medline: 28298398]
25. Vollenweider DJ, Frei A, Steurer-Stey CA, Garcia-Aymerich J, Puhan MA. Antibiotics for exacerbations of chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2018 Oct 29;10:CD010257. [doi: 10.1002/14651858.CD010257.pub2] [Medline: 30371937]
26. Sivakumar S, McNally A, Tobin J, Williamson J, Aneman A. Observational cohort study of outcomes in patients admitted to the intensive care unit for acute exacerbation of chronic obstructive pulmonary disease. Intern Med J 2018 Aug;48(8):944-950. [doi: 10.1111/imj.13805] [Medline: 29582542]
27. Faner R, Tal-Singer R, Riley JH, Celli B, Vestbo J, MacNee W, ECLIPSE Study Investigators. Lessons from ECLIPSE: a review of COPD biomarkers. Thorax 2014 Jul;69(7):666-672. [doi: 10.1136/thoraxjnl-2013-204778] [Medline: 24310110]
28. Crooks MG, den Brinker A, Hayman Y, Williamson JD, Innes A, Wright CE, et al. Continuous cough monitoring using ambient sound recording during convalescence from a COPD exacerbation. Lung 2017 Jun;195(3):289-294 [FREE Full text] [doi: 10.1007/s00408-017-9996-2] [Medline: 28353117]

XSL•FO
RenderX

29. Rodríguez LA, Wallander MA, Martín-Merino E, Johansson S. Heart failure, myocardial infarction, lung cancer and death in COPD patients: a UK primary care study. Respir Med 2010 Nov;104(11):1691-1699 [FREE Full text] [doi: 10.1016/j.rmed.2010.04.018] [Medline: 20483577]
30. Miao F, Cai YP, Zhang YX, Fan X, Li Y. Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest. IEEE Access 2018;6:7244-7253. [doi: 10.1109/ACCESS.2018.2789898]

Abbreviations

AECOPD: acute exacerbation of chronic obstructive pulmonary disease
CART: classification and regression tree
COPD: chronic obstructive pulmonary disease
DBP: diastolic blood pressure
GOLD: Global Initiative for Chronic Obstructive Lung Disease
NOH: number of hospitalizations
PR: pulse rate
RF: random forest
RR: respiratory rate
SBP: systolic blood pressure
TAHSYU: Third Affiliated Hospital, Sun Yat-sen University
TEMP: temperature

Edited by G Eysenbach; submitted 19.12.18; peer-reviewed by S Shah, I Yang, S Kamalakannan, M Ghajarzadeh; comments to author 29.04.19; revised version received 22.06.19; accepted 19.08.19; published 21.10.19.

Please cite as:
Zhou M, Chen C, Peng J, Luo CH, Feng DY, Yang H, Xie X, Zhou Y
Fast Prediction of Deterioration and Death Risk in Patients With Acute Exacerbation of Chronic Obstructive Pulmonary Disease Using Vital Signs and Admission History: Retrospective Cohort Study
JMIR Med Inform 2019;7(4):e13085
URL: http://medinform.jmir.org/2019/4/e13085/
doi:10.2196/13085
PMID:31638595

©Mi Zhou, Chuan Chen, Junfeng Peng, Ching-Hsing Luo, Ding Yun Feng, Hailing Yang, Xiaohua Xie, Yuqi Zhou. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 21.10.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploiting Machine Learning Algorithms and Methods for the Prediction of Agitated Delirium After Cardiac Surgery: Models Development and Validation Study

Hani Nabeel Mufti[1,2,3], MD, MSc, CIP, FRCSC; Gregory Marshal Hirsch[4], MD; Samina Raza Abidi[5], MBBS, PhD; Syed Sibte Raza Abidi[6], PhD

[1]Division of Cardiac Surgery, Department of Cardiac Sciences, King Faisal Cardiac Center, King Abdulaziz Medical City, Ministry of National Guard Health Affairs - Western Region, Jeddah, Saudi Arabia

[2]College of Medicine-Jeddah, King Saud bin Abdulaziz University for Health, Ministry of National Guard Health Affairs, Jeddah, Saudi Arabia

[3]King Abdullah International Medical Research Center, Jeddah, Saudi Arabia

[4]Department of Surgery, Faculty of Medicine, Dalhousie University, Halifax, NS, Canada

[5]Department of Community Health and Epidemiology, Faculty of Medicine, Dalhousie University, Halifax, NS, Canada

[6]kNowledge Intensive Computing for Healthcare Enterprise Research Group, Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

Corresponding Author:
Hani Nabeel Mufti, MD, MSc, CIP, FRCSC
Division of Cardiac Surgery, Department of Cardiac Sciences, King Faisal Cardiac Center
King Abdulaziz Medical City
Ministry of National Guard Health Affairs - Western Region
PO Box 9515
Mail code: 6599
Jeddah, 21423
Saudi Arabia
Phone: 966 122266666 ext 25805
Email: muftihn@ngha.med.sa

Abstract

Background: Delirium is a temporary mental disorder that occasionally affects patients undergoing surgery, especially cardiac surgery. It is strongly associated with major adverse events, which in turn leads to increased cost and poor outcomes (eg, need for nursing home due to cognitive impairment, stroke, and death). The ability to foresee patients at risk of delirium will guide the timely initiation of multimodal preventive interventions, which will aid in reducing the burden and negative consequences associated with delirium. Several studies have focused on the prediction of delirium. However, the number of studies in cardiac surgical patients that have used machine learning methods is very limited.

Objective: This study aimed to explore the application of several machine learning predictive models that can pre-emptively predict delirium in patients undergoing cardiac surgery and compare their performance.

Methods: We investigated a number of machine learning methods to develop models that can predict delirium after cardiac surgery. A clinical dataset comprising over 5000 actual patients who underwent cardiac surgery in a single center was used to develop the models using logistic regression, artificial neural networks (ANN), support vector machines (SVM), Bayesian belief networks (BBN), naïve Bayesian, random forest, and decision trees.

Results: Only 507 out of 5584 patients (11.4%) developed delirium. We addressed the underlying class imbalance, using random undersampling, in the training dataset. The final prediction performance was validated on a separate test dataset. Owing to the target class imbalance, several measures were used to evaluate algorithm's performance for the delirium class on the test dataset. Out of the selected algorithms, the SVM algorithm had the best F1 score for positive cases, kappa, and positive predictive value (40.2%, 29.3%, and 29.7%, respectively) with a *P*=.01, .03, .02, respectively. The ANN had the best receiver-operator area-under the curve (78.2%; *P*=.03). The BBN had the best precision-recall area-under the curve for detecting positive cases (30.4%; *P*=.03).

Conclusions: Although delirium is inherently complex, preventive measures to mitigate its negative effect can be applied proactively if patients at risk are prospectively identified. Our results highlight 2 important points: (1) addressing class imbalance on the training dataset will augment machine learning model's performance in identifying patients likely to develop postoperative delirium, and (2) as the prediction of postoperative delirium is difficult because it is multifactorial and has complex pathophysiology,

XSL•FO
RenderX

applying machine learning methods (complex or simple) may improve the prediction by revealing hidden patterns, which will lead to cost reduction by prevention of complications and will optimize patients' outcomes.

(JMIR Med Inform 2019;7(4):e14993) doi:10.2196/14993

KEYWORDS

delirium; cardiac surgery; machine learning; predictive modeling

Introduction

Background

Delirium or acute confusion is a temporary mental disorder that occurs among hospitalized patients [1]. The Society of Thoracic Surgeons defines delirium as a mental disturbance marked by illness, confusion, and cerebral excitement, with a comparatively short course [2], developing over a short period (usually from hours to days) and which tends to fluctuate during the day [3]. Delirium symptoms range from a disturbance in consciousness (eg, coma) to cognitive disorders involving disorientation and hallucinations. Delirium has a wide range of presentations, from extremely dangerous agitation to depression-like isolation and, on the basis of its presentation, it has 3 distinct subclasses—that is, hyperactive, hypoactive, and mixed [4]. This diversity of possible presentations, along with its sudden onset and unpredictable course, makes early detection challenging. Royston and Cox state that "from the patient's point of view, delirium and subsequent cognitive decline are among the most feared adverse events following surgery" [5]. The diversity of delirium's presentation, along with its sudden onset and unpredictable course, makes its early detection difficult; however, the ability to predict delirium in patients can play a fundamental role in initiating preventive measures that can significantly improve outcomes.

Patients undergoing cardiac surgery are at higher risk of developing delirium [6-9]. Several studies demonstrated a negative association between postoperative delirium and an increased morbidity and mortality [7-10]. Of particular concern is the strong relationship between delirium and postoperative infections in cardiac surgery patients [7,9,11].

Given the undesirable consequences of delirium on surgical outcomes, it is deemed useful to predict the potential incidence of delirium in patients to pre-emptively administer and plan for therapeutic interventions to deal with delirium and in turn improve the surgical outcomes. Typically, predictive models for delirium use a range of clinical variables, applied to conventional statistical methods, mainly logistic regression (LR) [12-14]. The current predictive models for delirium generally present a simplified linear weighted representation of the statistical significance of the clinical variables toward the prediction of delirium [15].

However, we argue that the prediction of delirium is quite complex given the multiplicity of reasons and confounding factors contributing to the manifestation of delirium in patients. Data mining methods can be used to uncover underlying relationships between variables to develop predictive models that can categorize the patient population into ones that have the propensity to develop delirium versus those that are less likely to develop delirium. Sometimes, these relationships or patterns cannot be easily explained yet appear to be essential and have a significant contribution to the improvement of the predictive model's performance, even if it is minimal (eg, a 0.01% improvement in a model's performance means that for every 1000 patients, 1 extra life is saved or a complication is prevented or an accident is avoided).

Artificial intelligence in health care, particularly the use of machine learning methods, provides a purposeful opportunity to discover such underlying patterns and correlations by mining the data leading to the *learning* of data-driven prediction models. Machine learning models have been successfully applied in medical data [16-22] to solve a wide range of clinical issues, such as myocardial infarction [23], atrial fibrillation [24], trauma [25], breast cancer [26-28], Alzheimer [29-31], cardiac surgery [22,32], and others [20,21,33-35].

The main objective of this study was to develop predictive models to pre-emptively predict the manifestation of agitated delirium in patients after cardiac surgery. Although discovering underlying hidden patterns is interesting and can be done using the data mining methods used in this work, this was not our main objective as the pathophysiology of delirium is considered multifactorial and complex to start with. The rationale is that if we can identify based on preoperative clinical parameters which patients are likely to develop postoperative delirium, then clinicians can initiate preventive and therapeutic measures in a timely fashion, to mitigate the undesirable effects of delirium. Our approach for predictive modeling is to investigate machine learning methods to *learn* the prediction models using retrospective clinical data for around 5500 patients over a 7-year period who received cardiac surgery at Queen Elizabeth II Health Sciences Center (QEII HSC) in Halifax, Canada. In this paper, several machine learning models were explored, including artificial neural networks (ANN), Bayesian belief networks (BBN), decision trees (DT), naïve Bayesian (NB), LR, random forest (RF), and support vector machines (SVM).

Related Work

Although the prevalence of postoperative delirium is low (10%-25%), it is associated with cognitive deterioration coupled with a set of complications in surgical patients. The complexity of delirium comes from its relation to multiple risk factors and the accompanying uncertainty of its pathophysiology [10,11,36]; this leads to challenges in pre-emptively identifying patients that are likely to develop postoperative delirium. Several authors have indicated that delirium is associated with adverse outcomes and advocate early recognition to ensure preventive measures can be applied in a timely and effective manner [3,7,9,10,13,14,37]. Some of the proposed preventive interventions that have been shown to reduce the incidence of

XSL•FO
RenderX

delirium in high-risk patients include early mobilization and use of patient's personal aids (reading glasses, hearing aid, etc) [38]. However, the pre-emptive identification of postoperative delirium is clinically challenging.

A structured PubMed search using the PubMed Advanced Search Builder with the structure ("delirium") AND "predictive model", will result in only 38 items. If we direct our attention to all the research published focusing on delirium and cardiac surgery, query structure ("delirium") AND "cardiac surgery", we will get 485 items. If we combine all the 3 terms, query structure (("delirium") AND "cardiac surgery") AND "predictive model", we will narrow the results down to 4 items.

In recognition of the importance of delirium within the cardiac surgical population, some have attempted to develop a predictive model. In this work, we decided to focus on articles that were published in English and focused on developing a predictive model for the prediction of delirium after cardiac surgery in adult patients. The initial search resulted in 38 articles. After reviewing the articles' abstracts, we excluded articles that were not written in English, not about cardiac surgery patients, and in which no statistical model was developed. We ended up with 16 articles that were available for review. Multimedia Appendix 1 represents a summary of most relevant studies that attempted to develop a model for the prediction of delirium after cardiac surgery on adult patients.

For patients who underwent cardiac surgery, Afonso et al [12] conducted a prospective observational study on 112 consecutive adult cardiac surgical patients. Patients were evaluated twice daily for delirium using Richmond Agitation-Sedation Scale (RASS) and confusion assessment method for the intensive care unit (CAM-ICU), and the overall incidence of delirium was 34%. Increased age and the surgical procedure duration were found to be independently associated with postoperative delirium. Similarly, Bakker et al [13] prospectively enrolled 201 cardiac surgery patients aged 70 years and above. They found that a low Mini-Mental State Exam score and a higher preoperative creatinine were independent predictors of postoperative delirium [13]. Unfortunately, both of these models were based on a small sample size (<250 patients) and did not have a validation cohort.

Research in the use of machine learning–based prediction models to detect delirium is rather limited, especially for cardiac surgery. Kramer et al [39] developed predictive models using a large dataset comprising medical and geriatrics patients that had the diagnosis of delirium in their discharge code and a control group of randomly selected patients from the same period who did not develop delirium. The prediction models performed well with the highest performance achieved by the RF model (receiver operating characteristic-area under the curve [ROC-AUC]≈91%). Although they argue that their data were imbalanced, they used the ROC-AUC as their evaluation metric, which does not consider the class imbalance. Davoudi et al [40] applied 7 different machine learning methods on data extracted from the electronic health (eHealth) record of patients undergoing major surgery in a large tertiary medical center to predict delirium; they found an incidence of 3.1%. They were able to achieve a ROC-AUC ranging from 71% to 86%. Owing to the class imbalance secondary to the low incidence of delirium and to improve the model's performance, they applied data-level manipulation using over- and undersampling, which did not result in a significant improvement (ROC-AUC ranging from 79% to 86%). Lee et al [41] published a nice systematic review and identified 3 high-quality ICU delirium risk prediction models: the Katznelson model, the original PRE-DELIRIC (PREdiction of DELIRium in ICu patients), and the international recalibrated PRE-DELIRIC model. All of these models used LR modeling as the primary technique for creating the predictive model. In the same paper by Lee et al [41], they externally validated these models on a prospective cohort of 600 adult patients that underwent cardiac surgery in a single institution. After updating, recalibrating, and applying decision curve analysis (DCA) to the models, they concluded that the recalibrated PRE-DELIRIC risk model is slightly more helpful. They argue that available models of predicting delirium after cardiac surgery have only modest accuracy. The current models are suboptimal for routine clinical use. Corradi et al [42] developed a predictive model using a large dataset (~78,000 patients) over 3 years in a single center using a good number of feature set (~128 variables). Their model had very good accuracy and the ROC-AUC ~90% on their test dataset. They used the CAM to detect delirium in the intensive care (CAM-ICU) and regular patient wards. Lee et al [41] conducted a systematic review in search for prediction models for delirium specifically designed for cardiac surgery patients. They found only 3 high-quality models and externally validated them on a local population of 600 patients. They used several metrics to evaluate the recalibrated models on the validation cohort (ROC-AUC, Hosmer–Lemeshow test, Nagelkerke's R2, Brier score, and DCA). In their analysis, the recalibrated PRE-DELIRIC prediction model performed better when compared with the Katznelson model. However, based on the DCA and the expected net benefit of both models, there appears to be limited clinical utility of any of the models.

Methods

Data Sources and Study Population

This single-center retrospective cohort study included patients who underwent cardiac surgery at the QEII HSC in Halifax, Canada, between January 2006 and December 2012. Over those 7 years, 7209 patients underwent cardiac surgery. The Maritime Heart Center (MHC) registry was used to create the dataset. The MHC registry is a prospectively collected, detailed clinical database on all cardiac surgical cases performed at the MHC since March 1995 with more than 20,000 patients and 500 different variables. The final dataset included 5584 patients who met our inclusion criteria and were successfully discharged (home, other institution closer to home, nursing home, or rehabilitation facility).

Delirium in the acquired database is coded as a binary outcome (Yes/No) and is defined as short-lived mental disturbance marked by illusions, confusion, or cerebral excitement, requiring temporary medical and/or physical intervention or a consultation, or extending the patient's hospital stay. Intraoperative management varied depending on the anesthetist

XSL•FO
RenderX

preferences and the patient clinical status. Although most patients were managed in a systematic approach based on standard of care, in the ICU, CAM-ICU was used to trigger further investigations if delirium was suspected. If delirium had been suspected after transfer from the ICU, the diagnosis was confirmed using different diagnostic criteria and screening tools (eg, Mini-Mental State Exam and CAM).

Full ethics approval was obtained from the Capital Health Research Ethics Board, in keeping with the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans. Informed consent was waved by the ethics board as the study did not involve therapeutic interventions or potential risks to the involved subjects.

Predictive Modeling: Methodology and Methods

Our aim was to develop a prediction model that can identify patients who are at risk of developing delirium after cardiac surgery. We investigated relevant machine learning methods, each with a specific learning algorithm to correlate the patient presurgery variables with a probabilistic determination of delirium as per the observations noted in the cardiac surgery dataset. The rationale for working with multiple machine learning methods was to determine the effectiveness of the different methods and then to select the best performing model that can be used in a clinical setting to predict postoperative delirium in new patients.

We pursued the standard data mining methodology comprising 6 steps as shown in Multimedia Appendix 2. These steps are as follows: (1) *data acquisition*: This step involved the procurement of the required dataset from the source (in this case from the MHC), while complying with data access and secondary data usage protocols; (2) *data preprocessing:* This step involved the cleaning of the dataset by removing incomplete records and next identifying the significant features/variables to develop the prediction models; (3) *modeling strategy set-up*: This step involved the formulation of the modeling strategy in terms of data partitioning into training dataset (N=4476; 80% of original) and test dataset (N=1117; 20% of original), data presentation during training, model evaluation criteria; (4) *class imbalance and training dataset class optimization:* This step was introduced to address the target class imbalance in the original dataset, so as to minimize the effect of the dominant class on the performance of the predictive models. We explored data level techniques, such as over- and undersampling, to address the class imbalance in the final training dataset, resulting in the final balanced training dataset (n=1014). (5) *model learning:* This step involved setting up different model configurations—that is, setting up the model parameters for the candidate machine learning methods—and learning the models by presenting the preprocessed training data (step 2) as per the modeling strategy (step 3). As model learning is an exploratory exercise where different model configurations and multiple instantiations of the model are pursued to account for the probabilistic nature of machine learning methods and to avoid overfitting, 10-fold stratified cross-validation was used; and (6) *model evaluation:* In this step, the learnt models are evaluated (against the predefined criteria) for their effectiveness to predict delirium using the test data.

Data Preprocessing and Variables Selection

Characteristics of patients who developed delirium postoperatively were compared with patients who did not. The mean and standard deviation were used for continuous variables that had a normal distribution and were compared using the 2-sided *t* test. Continuous variables that were not normally distributed were reported using the median and interquartile range and were compared using the Wilcoxon rank sum test. Categorical variables were reported as frequencies and percentages and were analyzed by ^{2} (Chi-square) or Fisher exact test as appropriate. The Kruskal-Wallis test was used for ordinal variables. Next, exploratory data analysis followed by univariate LR analysis was applied to isolate key perioperative variables with significant influence on postoperative agitated delirium.

All measures of significance are 2-tailed, and a *P* value <.05 was considered statistically significant. Statistical analysis and the assessment of model's performance was conducted using the R-Software, version 3.1.0 (R Project for Statistical Computing) [43]. On the basis of univariate LR analysis, 22 variables were used to generate the machine learning–based predictive models.

The basic premise of any DT model is that it recursively split features based on the target variable's purity. The ultimate goal of the algorithm is to optimize each split on maximizing the homogeneity of the grouping at each split (also known as purity) [44-46]. A node having multiple classes is impure, whereas a node having only 1 class is pure. One of the useful features of RF is its ability to identify relevant variables by assigning variable importance measure to the input variables [44-46]. Variable importance in RF can be measured using either misclassification error, Gini index, or cross-entropy. Most machine learning experts discourage the use of misclassification error in tree-based models because it is not differentiable and, hence, less amenable to numerical optimization [44,46]. In addition, cross-entropy and the Gini index are more sensitive to changes in the node probabilities than the misclassification rate. Both Gini index and cross-entropy apply probability to gauge the disorder of grouping by the target variable. However, they are a bit different, and the results can vary. The Gini index measures how often a randomly chosen element from the set would be incorrectly labeled, starting with the assumption that the node is impure (Gini index=1) and subtracting the probabilities of the target variable. If the node is composed of a single class (also known as pure), then the Gini index will be 0. On the other hand, cross-entropy is more computationally heavy because of the log in the equation. Instead of utilizing simple probabilities, this method takes the log of the probabilities (usually the log base 2; any log base can be used, but it has to be consistent for the sake of comparison between different tree-based models). The entropy equation uses logarithms because of many advantageous properties (mainly the additive property) that can be very beneficial in imbalanced class distributions and multiclass target variables [44,46]. A cross-entropy of 1 indicates a highly disorganized node (impure node), whereas a cross-entropy of 0 indicates a highly organized node (pure node).

XSL•FO
RenderX

In RF, each tree in the forest is grown fully (unpruned) using bootstrap samples of the original dataset, the out-of-bag (OOB) samples are used as test samples. A random subset of variables *k* from the original input variables space *K (where k<K)* is used at each node. On the basis of a specific measure (eg, mean decrease in impurity, Gini index, and mean decrease in accuracy), variables are selected, and the process is repeated to the end of the tree. The performance of each tree is computed over the corresponding OOB sample. For each variable, its importance is calculated as the mean relative decrease across the forest of trees performance when the observations of this variable in the OOB sample are randomly permuted. As the Waikato Environment for Knowledge Acquisition software (WEKA) was used in this work to develop the RF model, it applies the cross-entropy method as its default method for variables importance ranking.

The Issue of Outcome Class Imbalance

In our dataset, the outcome class distribution is notably imbalanced (only 11.4% of patients developed delirium). Typically, classification algorithms tend to predict the majority class very well but perform poorly on the minority class due to 3 main reasons [47-49]: (1) the goal of minimizing the overall error (maximize accuracy), to which the minority class contributes very little; (2) algorithm's assumption that classes are balanced; and (3) the assumption that impact of making an error is equal.

Several data manipulation techniques can be applied to reduce the impact of this class imbalance: at the data level (oversampling minority class or undersampling the majority class) or at the algorithm level (applying different costs to each class) [47-49]. Although data manipulation methods can improve a model's performance, these methods do have some drawbacks [49]. At the data-level manipulation, oversampling tends to artificially increase the number of the minority class by creating modified copies; it tends to overfit the results to the training set and consequently is likely to poorly generalize. On the other hand, because undersampling discards some of the majority class observations, it essentially bears the risk of losing some potentially important hidden information. Algorithm level manipulation involves some trial and error and can be sensitive to training data changes.

In real life, class imbalance cannot be avoided as it is a result of the nature of the problem and domain (eg, natural disasters and patient death). In our dataset, oversampling led to overfitting on the training dataset with suboptimal generalization when applied to the imbalanced dataset. As postoperative delirium is linked with a wide range of complications (from a minor temporary confusion that totally resolves with no sequalae to the other extreme of sepsis and death), it is very hard to associate it to a specific cost. As such, given the intent of this study, we decided to apply random subsampling to balance the training dataset and have equal representation of outcome classes, thus optimizing the training dataset for the models. We used the *SpreadSubSample* filter in WEKA [46] to produce a random subsample by undersampling the majority class (which can be done by either specifying a ratio or the number of observations). In our case, we specified a ratio of 1:1. By doing so, the filter generates a new balanced dataset by decreasing the number of the majority class instances, which reduces the difference between the minority and the majority classes. Undersampling is considered an effective method for dealing with class imbalance [50]. In this approach, a subset of the majority class is used to learn the model. Many of the majority class examples are ignored; the training set becomes more balanced, which makes the training more efficient. The most common type of undersampling is random majority undersampling (RUS). In RUS, observations from the majority class are randomly removed. The final balanced training dataset (N=1014, 1:1 delirium) was used to develop the models.

Training With 10-Fold Cross-Validation and Test Datasets

In predictive modeling, it is a common practice to separate the data into training and test dataset. In an effort to avoid overfitting and overestimating the model's performance, the test dataset is only used to evaluate the performance of the prediction model [44,46,51,52]. The problem of evaluating the model on the training dataset is that it may exhibit high prediction ability (overfitting), yet it fails when asked to predict new observations. To address this issue, cross-validation is commonly used to (1) estimate the generalizability of an algorithm and (2) optimize the algorithm performance by adjusting the parameters [44,46,51-53]. We applied stratified 10-fold cross-validation on the balanced training dataset (50% delirium). The test dataset was preserved imbalanced to simulate the real clinical scenario and evaluate the behavior of different methods. Several metrics were used, that are immune to class imbalance, to appraise the final model's performance on the test dataset [44,46,47,49,51,52].

Results

Development of Prediction Models: Experiments and Results

We investigated a range of relevant predictive modeling methods—that is, function-based models (LR, ANN, and SVM), Bayesian models (NB and BBN), and tree-based models (C4.5 DT and RF)—to generate 7 prediction models (all developed using the same balanced dataset). All models were generated and tested using the WEKA software, version 3.7.10 [54]. The setting of the prediction models and the optimization steps that were applied in this research are available in Multimedia Appendix 3. These predictive modeling algorithms were chosen based on 2 main reasons: (1) their noted effectiveness in solving medical-related classification problems and (2) a strong theoretical background that supports predictive modeling via data classification [11,16,19,20,22,23,25,30,31,39,46,52,55-63]. Experiments were conducted on a MacBook Pro (Apple Inc; 15-inch, 2017) with a 3.1-GHz Intel Core i7 processor and a 16 GB RAM 2133 MHz, running a MacOS High Sierra Version 10.13.

General Patients' Characteristics and Important Variables in the Dataset

Given the above definitions and procedures, agitated delirium was documented in 11.4% patients (n=661). The majority of

XSL•FO
RenderX

patients were men (74%). Coronary artery bypass graft (CABG) was the most commonly performed procedure (67%). Almost 56% stayed in the ICU for 24 hours or less. Only 2% suffered a permanent stroke. Patients who developed postoperative agitated delirium were older and had a significantly higher incidence of comorbid diseases. A higher proportion of patients who developed agitated delirium underwent a combined procedure (CABG plus valve). The median stay in the cardiovascular intensive care unit in hours was 4 times higher for patients who developed agitated delirium postoperatively, compared with patients who did not (P<.001). Univariate analysis of in-hospital mortality did not show any statistical significance (in-hospital mortality: 4.1% vs 3.6%; P=.57; Table 1).

Univariate LR analysis of all pre-, intra-, and postoperative variables that can contribute to the development of delirium was performed using appropriate statistical tests in the R-Software. Univariate LR was applied on all candidate variables with a P value of less than .05 in univariate LR analysis to extract odds ratio (OR) with 95% CI generated for each candidate variable. The candidate variables were ranked based on the how low is the actual P value, the Akaike information criterion (lower is better), and impact of variable on postoperative delirium (signified by the OR). Then WEKA was used to generate variable importance using the RF model. WEKA applies the cross-entropy method to assess purity of the candidate variables with the RF algorithm as its default method, as it is more sensitive to class imbalance. Variables that appear higher at the trees are considered more relevant [44,51,52,63]. This is represented by the percentage of decrease of impurity (or increase of purity) of the final model based on adding this specific attribute. The number of times the candidate variable appeared in any location in all of the created tree models through the RF ensemble model process is also a criterion used in WEKA. The more times a variable is being selected in the RF creation process, the higher likelihood of it being important for the classification of the final target variable. This is also reflected in the decrease of impurity measure as the more decrease in impurity, the higher number of times that variable appears, which can imply its importance. Table 2 displays the importance of each input variable used in our RF model and its rank compared with the univariate LR analysis.

XSL•FO
RenderX

Table 1. Patient characteristics (N=4467).

Patient characteristics	Delirium		*P* value
	No (n=3960)	Yes (n=507)	
Preoperative characteristics			
Age (years)			**<.001**
Mean (SD)	66 (11)	72 (10)	
Range	19-95	25-91	
Male gender, n (%)	2942 (74.3)	386 (76.1)	.36
Hypertension, n (%)	2970 (75)	401 (79)	.04
Diabetes mellitus, n (%)	1426 (36)	223 (44)	<.001
Cerebrovascular disease, n (%)	436 (11)	112 (22)	<.001
Chronic obstructive pulmonary disease, n (%)	531 (13.4)	104 (20.5)	<.001
Frail, n (%)	238 (6)	49 (9.7)	.002
Ejection fraction <30%, n (%)	436 (11)	106 (21)	<.001
Preoperative atrial fibrillation, n (%)	424 (10.7)	102 (20.1)	<.001
EURO II[a] score >5%, n (%)	717 (18.1)	231 (45.6)	<.001
Urgency, n (%)			**<.001**
Elective (admitted from home)	1901 (48)	198 (39)	
Need surgery during hospitalization	1742 (44)	223 (44)	
Urgent/emergent (life threatening)	317 (8)	91 (18)	
Intraoperative characteristics, n (%)			**<.001**
Procedure			
Coronary artery bypass graft	2744 (69.3)	291 (57.4)	
Aortic valve replacement	622 (15.7)	93 (18.3)	
Mitral valve surgery[b]	170 (4.3)	20 (4)	
CABG+AVR[c]	325 (8.2)	79 (15.6)	
CABG+MV[d] surgery	51 (1.3)	3.4 (17)	
Repeat sternotomy	230 (5.8)	59 (11.6)	<.001
In-hospital morbidity, n (%)			**<.001**
Reintubation	79 (2)	48 (9.5)	
New postoperative atrial fibrillation	1247 (31.5)	217 (42.8)	
Pneumonia	174 (4.4)	101 (20)	
Sepsis	40 (1)	35 (6.9)	
Deep sternal wound infection	24 (0.6)	15 (3)	
Blood products transfusion within 48 hours from surgery	990 (25)	269 (53)	
Length of stay after surgery <1 week	2257 (57)	66 (13)	
Discharged home	3513 (88.7)	301 (59.4)	

[a]EURO II: European System for Cardiac Operative Risk Evaluation II.

[b]Mitral valve replacement or repair.

[c]CABG+AVR: coronary artery bypass graft + aortic valve replacement.

[d]CABG+MV: coronary artery bypass graft + mitral valve.

XSL•FO
RenderX

Table 2. List of candidate variables based on univariate logistic regression analysis compared with random forest.

Variable	Type	Unit	Univariate logistic regression analysis[a]			Random forest		
			OR (95% CI)	*P* value	Rank	Decrease of impurity, %	Nodes using that attribute, n	Rank
Age (years)	Continuous	Years	1.1 (1.03-1.07)	<.001	1[b]	43	3238	1
Mechanical ventilation >24 hours	Categorical	Yes/no	5.8 (3.9-8.6)	<.001	3	21	297	21
Preoperative creatinine clearance	Continuous	µmol/L	0.97 (0.96-0.98)	<.001	1[b]	39	2544	4
Length of stay in the ICU[c]	**Ordinal**	—[d]	—	—	**2**	**26**	**590**	**20**
>72 hours	—	—	7.6 (4.9-11.9)	<.001	—	—	—	—
24-72 hours	—	—	1.7 (0.9-2.8)	<.001	—	—	—	—
Procedure other than isolated CABG[e]	Categorical	Yes/no	2.9 (1.8-2.5)	<.001	6	28	370	15
Blood product within 48 hours	Categorical	Yes/no	2.9 (2.0-4.2)	<.001	5	28	452	14
Intraoperative TEE[f]	Categorical	Yes/No	2.0 (1.3-3.1)	.002	10	27	568	18
EURO II[g] score	Continuous	Percent	1.07 (1.05-1.09)	<.001	17	41	2716	2
Preoperative hemoglobin	Continuous	gm/dL	0.98 (0.97-0.99)	<.001	1[b]	40	2766	3
Preoperative A-Fib[h]	Categorical	Yes/no	2.3 (1.4-3.6)	<.001	7	35	486	6
Timing of IABP[i]	**Ordinal**	—	—	—	**4**	**29**	**329**	**12**
Preoperative	—	—	1.4 (0.6-2.9)	.42	—	—	—	—
Intraoperative	—	—	6.8 (1.9-23.1)	.002	—	—	—	—
Intraoperative inotropes	Categorical	Yes/no	2.1 (1.4-3.0)	<.001	8	27	514	17
COPD[j]	Categorical	Yes/no	1.7 (1.1-2.7)	.02	14	33	689	9
CVD[k]	Categorical	Yes/no	1.8 (1.1-2.9)	.01	13	29	516	13
DM[l]	Categorical	Yes/no	0.9 (0.6-1.4)	.79	16	39	995	5
Frail	Categorical	Yes/no	2.0 (1.1-3.5)	.03	12	30	381	11
History of turn down	Categorical	Yes/no	8.2 (2.8-24.3)	<.001	1[b]	21	93	22
EF[m] categories	**Ordinal**	—	—	—	**9**	**33**	**89**	**10**
30%-50%	—	—	1.4 (0.9-2.1)	.18	—	—	—	—
<30%	—	—	2.1 (1-4.2)	.04	—	—	—	—
Gender	Categorical	Yes/no	1.2 (0.8-1.9)	.47	16	35	752	7
Aortic stenosis	**Ordinal**	—	—	—	**14**	**26**	**899**	**16**
Moderate	—	—	1.4 (0.6-2.8)	.43	—	—	—	—
Severe	—	—	1.6 (1.1-2.5)	.01	—	—	—	
Mitral insufficiency	**Ordinal**	—	—	—	**15**	**26**	**899**	**19**
Moderate	—	—	1.4 (0.9-2.1)	.07	—	—	—	—
Severe	—	—	2.3 (1.02-4.6)	.03	—	—	—	—
Postoperative arrhythmias	Categorical	Yes/no	1.8 (1.3-2.8)	.002	11	34	746	8

[a]Analysis was done using univariate logistic regression with a *P* value of <.05 considered to be statistically significant.

[b]These variables were all equally ranked as 1st because they had almost equal odds ratios and a *P* value of <.001.

[c]ICU: intensive care unit.

[d]Not applicable.

[e]CABG: coronary artery bypass graft.

[f]TEE: transesophageal echo.

[g]EURO II: European System for Cardiac Operative Risk Evaluation II.

XSL•FO
RenderX

[h]A-Fib: atrial fibrillation.

[i]IABP: intra-aortic balloon pump.

[j]COPD: chronic obstructive pulmonary disease.

[k]CVD: cerebrovascular disease.

[l]DM: diabetes mellitus.

[m]EF: ejection fraction.

Prediction Model's Performance Evaluation

There exist several metrics to evaluate the performance of a predictive model, whereby predictive accuracy is the most commonly used metric as it relates a model's ability to correctly identify observation assignments, irrespective of the class distribution. However, in the presence of a noted class imbalance in the dataset, this measure can be misleading because the minority class (positive cases in our dataset) has a smaller influence of the model's output, and as such the model will tend to favor the majority class [47]. In our dataset, there is a significant imbalance of the outcome of interest distribution (delirium: 11.4% positive cases).

To provide a more robust evaluation of the prediction model's performance, in the presence of the class imbalance in our dataset, we used the evaluation measures of F1 measure, ROC-AUC, and precision-recall curve area under the curve (PRC-AUC) [44,46,47,51,52]. The ROC-AUC was primarily used to assess the classifier's general performance (model discrimination=how well the predicted risks distinguish between patients with and without disease) [64]. The F1 score was primarily used as the harmonic mean of precision and recall [46,52]. The F1 score provides the most reliable assessment of a model's prediction performance, while considering the worst-case prediction scenario for a classifier (model calibration=evaluates the reliability of the estimated risks: if we predict 10%, on average 10/100 patients should have the disease) [64].

Sensitivity (recall) is considered a measure of completeness (the percentage of positive cases that have been correctly identified as positive). Positive predictive value (precision, PPV) is considered a measure of exactness (the percentage of cases labeled by the classifier as positive that are indeed positive) [46,52]. The PRC-AUC is a useful measure in the presence of class imbalance, and the outcome of interest is to identify the minority class [65,66]. The PRC identifies the PPV for each corresponding value on the sensitivity scale (model calibration). As the PRC is dependent on the class representation in the dataset, it provides a simple visual representation of the model's performance across the whole spectrum of sensitivities. By doing so, it can aid in identifying the best model (based on the trade of being either exact vs complete, ideally optimizing both) [66]. In addition, the PRC enables comparing models at predetermined recall thresholds (eg, the best precision at 50% recall). This adds more fixability in choosing the best model based on the domain and problem in hand.

As our primary interest was to identify patients who were more likely to develop delirium (minority class) while accounting for the class imbalance in the test dataset, we decided to evaluate the models using the ROC-AUC as a measure of the model discrimination in conjunction with F1 score and PRC-AUC as measures of the model calibration. Tables 3 and 4 present the prediction performance of all prediction models based on the test data. Figure 1 illustrates the ROC-AUCs and PRC-AUCs for the developed models.

When comparing the prediction performance using the ROC-AUC (Figure 1) for the test dataset, it may be noted that the prediction performance of all the prediction models on the test dataset is quite similar, except for DT, which was lower. This indicates that there is no obvious difference in the discriminative power of the classification models—that is, the ability of a model to distinguish between positive cases from negative ones. However, given the class imbalance in our dataset, this result might not be representative of a model's true predictive power; hence, a further examination of the results was needed to identify the best performing model given the class imbalance.

As LR was the most commonly used algorithm to predict the manifestation of postoperative delirium in the medical literature [8,12,13,40,41,67-74], we developed a multivariate step-wise LR model that identified 8 variables as significant predictors of postoperative agitated delirium (Multimedia Appendix 2). The main purpose of developing the LR model was to give medical experts, who are not familiar with machine learning algorithms, an algorithm that they are acquainted with and use as a comparator.

In our study, for every 100 patients who developed delirium, the RF model had the best sensitivity and was able to correctly identify 72 patients (see Tables 3 and 4). The SVM model had the best PPV (out of 100 patients who were labeled positive by SVM, 30 were actually positive) and the best accuracy, specificity, and kappa. The PRC-AUC and F1 scores for SVM were the best out of all models (29.2% and 40.2%, respectively), with moderate discrimination (ROC-AUC=77.2 %). We also examined the relationship between precision (PPV) and recall (sensitivity) at different thresholds (see Table 5). At 50% sensitivity (recall), the RF model had the best precision, 37%). At 75% sensitivity (recall), RF was the best model with a precision of 25% followed by ANN with a PPV of 24%.

XSL•FO
RenderX

Table 3. Comparison of model's performance metrics applied on the balanced training dataset using 10-fold cross-validation and the imbalanced test dataset to predict delirium after cardiac surgery. Performance metrics: accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and Cohen kappa. All measures are reported out of 100% with standard deviation in brackets as a measure of variability.

Model	Accuracy	Δ[a]	Sensitivity	Δ	Specificity	Δ	PPV[b]	Δ	NPV[c]	Δ	Kappa	Δ
Dataset: 10-fold cross-validation applied on the balanced training dataset (N=1014, delirium=50%)												
ANN[d]	71.7 (4.3)	ns[e]	71.8 (7)	+[f]	71.6 (7)	–[g]	71.7 (5)	−	71.7 (7)	−	43.3 (9)	ns
BBN[h]	71.3 (4.4)	ns	72.2 (7)	+	71.2 (7)	−	69.9 (5)	−	71.3 (7)	−	43.1 (9)	ns
DT[i]	70.1 (4.3)	ns	68.1 (7)	ns	72.9 (9)	ns	72.9 (5)	ns	72.6 (9)	ns	43.3 (8)	ns
LR[j]	73.3 (4.4)	B[k]	69.8 (7)	B	76.7 (7)	B	75 (5)	B	75.6 (6)	B	44.5 (9)	B
NB[l]	73.0 (4.2)	ns	64.8 (7)	ns	79.5 (5)	+	74.4 (5)	ns	79.5 (5)	+	42.9 (8)	ns
RF[m]	72.5 (4.4)	ns	74.3 (7)	+	71.7 (7)	−	72.1 (4)	ns	72.8 (7)	−	45.7 (9)	ns
SVM[n]	71.3 (4.5)	ns	60.2 (8)	−	83.8 (5)	+	77.8 (5)	+	83.1 (5)	+	43.2 (9)	ns
Dataset: Imbalanced test dataset (N=1117, delirium=11.4%)												
ANN	74.3 (3.2)	ns	67.7 (5)	+	72.9 (5)	ns	24.3 (14)	ns	94.6 (5)	ns	22.85 (9)	ns
BBN	74.1 (3.8)	ns	68.7 (9)	+	70.8 (9)	−	22.9 (15)	ns	94.5 (6)	ns	21.81 (11)	ns
DT	74.4 (5.4)	ns	66.9 (10)	+	75.4 (10)	ns	25.8 (17)	ns	94.7 (10)	ns	24.97 (13)	ns
LR	75.6 (4.7)	B	64.6 (9)	B	77.1 (7)	B	26.5 (16)	B	94.4 (8)	B	22.6 (13)	B
NB	71.7 (3.1)	−	66.1 (12)	ns	72.4 (8)	−	23.5 (18)	ns	94.3 (9)	ns	21.55 (10)	ns
RF	75.4 (3.4)	ns	72.4 (4)	+	72.4 (4)	−	25.2 (8)	+	95.3 (4)	+	24.69 (7)	ns
SVM	78.9 (2.1)	+	62.2 (4)	ns	81.1(3.2)	+	29.7 (12)	+	94.4 (6)	ns	29.33 (9)	+

[a]Change compared to base model (B).

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

[d]ANN: artificial neural networks.

[e]ns: not a statistically significant change in performance ($P \geq .05$).

[f]Statistically significant improvement of performance metric ($P < .05$).

[g]Statistically significant deterioration of performance metric ($P < .05$).

[h]BBN: Bayesian belief networks.

[i]DT: J48 decision tree.

[j]LR: logistic regression.

[k]B: base comparator (reference) algorithm.

[l]NB: naïve Bayesian.

[m]RF: random forest.

[n]SVM: support vector machines.

XSL•FO
RenderX

Table 4. Comparison of model's performance metrics applied on the balanced training dataset using 10-fold cross-validation and the imbalanced test dataset to predict delirium after cardiac surgery. Performance metrics: receiver operator curve-area under the curve, harmonic mean of precision and recall, and precision-recall curve-area under the curve. All measures are reported out of 100% with standard deviation in brackets as a measure of variability.

Model	ROC-AUC[a]		F1 score[b]						PRC-AUC[c]					
			Yes[d]	Δ[e]	No[f]	Δ	Avg[g]	Δ	Yes	Δ	No	Δ	Avg	Δ
Dataset: 10-fold cross-validation applied on the balanced training dataset (N=1014, delirium=50%)														
ANN[h]	80.4 (4)	ns[i]	71.7 (5)	ns	71.7 (5)	ns	71.7 (5)	ns	78.5 (5)	ns	80.1 (5)	ns	79.3 (5)	ns
BBN[j]	77.4 (4)	–[k]	70.1 (5)	ns	69.1 (5)	ns	69.6 (5)	ns	75.3 (5)	ns	77.3 (5)	ns	76.3 (5)	–
DT[l]	77.2 (4)	ns	70.9 (4)	ns	72.4 (4)	ns	71.7 (4)	ns	74.4 (5)	ns	73.8 (5)	ns	73.8 (5)	–
LR[m]	81.4 (4)	B[n]	72.3 (5)	B	74.2 (5)	B	73.2 (5)	B	79.8 (5)	B	81 (5)	B	80.4 (5)	B
NB[o]	79.9 (4)	ns	72.7 (5)	ns	73.2 (5)	ns	73 (5)	ns	78.1 (5)	ns	79.8 (5)	ns	78.9 (5)	ns
RF[p]	81.3 (4)	ns	74.1 (5)	ns	72.6 (5)	ns	73.3 (5)	ns	78.8 (5)	ns	81 (5)	ns	79.9 (5)	ns
SVM[q]	81.1 (5)	ns	67.2 (6)	–	74.4 (6)	ns	71.1 (6)	–	80.4 (5)	ns	80.5 (5)	ns	80.4 (5)	ns
Dataset: Imbalanced test dataset (N=1117, delirium=11.4%)														
ANN	78.2 (6)	ns	35.8 (9)	ns	82.4 (9)	ns	77.1 (9)	ns	30.4 (9)	+[r]	96.2 (9)	ns	88.7 (9)	ns
BBN	77.3 (6)	ns	34.3 (8)	ns	82.9 (8)	ns	76.6 (8)	ns	30.7 (8)	+	95.8 (8)	ns	88.4 (8)	ns
DT	74.6 (7)	–	37.3 (8)	ns	83.9 (8)	ns	78.6 (8)	ns	25.3 (8)	ns	94.3 (8)	ns	86.5 (8)	ns
LR	77.5 (5)	B	37.6 (11)	B	84.9 (11)	B	79.5 (11)	B	27.1 (10)	B	97.1 (10)	B	88.4 (10)	B
NB	75.6 (8)	ns	34.7 (10)	ns	81.9 (10)	ns	76.6 (10)	ns	28.7 (9)	ns	95.6 (9)	ns	88.0 (9)	ns
RF	78.0 (4)	ns	37.4 (8)	ns	82.3 (8)	ns	77.2 (8)	ns	28.3 (8)	ns	96.3 (8)	ns	88.6 (8)	ns
SVM	77.2 (6)	ns	40.2 (7)	+	87.2 (7)	+	81.9 (7)	+	29.6 (9)	+	96.0 (9)	ns	88.4 (9)	ns

[a]ROC-AUC: receiver operator curve-area under the curve.

[b]F1 score: harmonic mean of precision and recall.

[c]PRC-AUC: precision-recall curve-area under the curve.

[d]Yes: positive instances or patients who developed delirium.

[e]Change compared to base model (B)

[f]No: negative instances or patients who did not develop delirium.

[g]Avg: weighted average measured as the sum of all values in that metric, each weighted according to the number of instances with that particular class label by multiplying that value by the number of instances in that class, then divided by the total number of instances in the dataset.

[h]ANN: artificial neural networks.

[i]ns: not a statistically significant change in performance ($P \geq .05$).

[j]BBN: Bayesian belief networks.

[k]Statistically significant deterioration of performance metric ($P < .05$).

[l]DT: J48 decision tree.

[m]LR: logistic regression.

[n]B: base comparator (reference) algorithm.

[o]NB: naïve Bayesian.

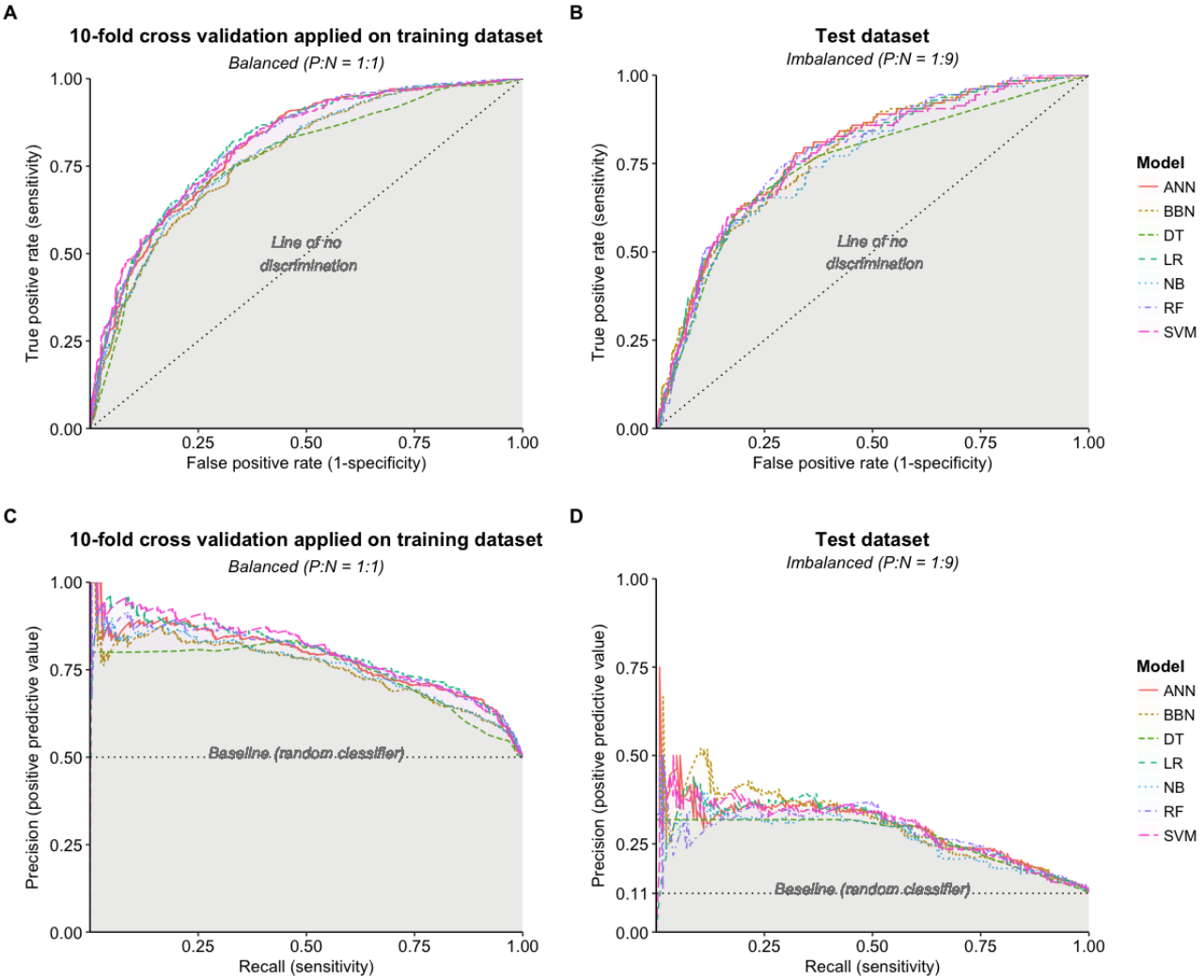
[p]RF: random forest.

[q]SVM: support vector machines.

[r]Statistically significant improvement of performance metric ($P < .05$).

XSL•FO
RenderX

Figure 1. Receiver-operator curves (ROC) and precision-recall curves (PRC) for the training dataset using 10-fold cross-validation and test datasets. (A) ROC for training using 10-fold cross-validation. (B) ROC for test dataset. (C) PRC for training using 10-fold cross-validation. (D) PRC for test dataset. ANN: artificial neural networks; BBN: Bayesian belief networks; DT: J48 decision tree; LR: logistic regression; NB: naïve Bayesian; RF: random forest, SVM: support vector machines; P:N: positive to negative ratio.



XSL•FO
RenderX

Table 5. Precision of each model for all datasets at different recall thresholds.

Recall threshold (%)	Model precision (%)						
	ANN[a]	BBN[b]	DT[c]	LR[d]	NB[e]	RF[f]	SVM[g]
Dataset: Training with 10-fold cross-validation							
~25	87	83	81	88	85	87	91
~50	80	78	81	82	78	83	83
~75	71	70	69	73	70	72	72
Dataset: Test							
~25	35	40	31	36	32	32	36
~50	34	33	30	34	31	37	34
~75	24	22	21	23	20	25	23

[a]ANN: artificial neural networks.

[b]BBN: Bayesian belief networks.

[c]DT: J48 decision tree.

[d]LR: logistic regression.

[e]NB: naïve Bayesian.

[f]RF: random forest.

[g]SVM: support vector machine.

On the basis of our experiments using the PRC-AUC and PRC analysis, the RF and ANN models demonstrated the ability to distinguish patients at risk of developing delirium (minority class) when compared with the other models. ANN is considered to be a *black box* as it is difficult to explain, especially to people who are nonexperts, not familiar with the principles and motivation behind the ANN algorithm, and do not know how the algorithm reaches its decision and activation thresholds. However, major work has been conducted over the last decade and is still ongoing to enhance the expandability of ANN by unlocking the black box to allow accountability [75-77]. Numerous techniques have been developed and were successfully applied [78-80], giving some transparency to the model and making it more human interpretable.

Discussion

Principal Findings

Patients undergo high-risk interventions with the expectation of improving their quality of life. It is highly undesirable that any medical intervention, inadvertently, negatively impacts their cognitive functions and in turn quality of life, especially if an adverse outcome is preventable.

With the paradigm shift in health care emphasizing the patient's quality of life after an intervention [81], innovative approaches are needed to both pre-emptively identify and effectively treat delirium. Given the availability of long-term surgical outcome data and advance machine learning methods, it is now possible to investigate the formulation of data-driven prediction models to pre-emptively identify patients susceptible to postsurgery delirium. LR-based prediction models to detect delirium have been developed using patient data from electronic medical records—in one study advanced text mining has been applied to abstract relevant data from clinical notes [82], and in another study attribute-based triggers were used [57]. We contend that with the availability of large volumes of patient data (before, during, and after the medical intervention), there are practical opportunities to develop data-driven prediction models to detect postoperative delirium in patients. Such artificial intelligence–based machine learning–based models are quite capable of identifying hidden yet important relations among variables and representing them in terms of a mathematical model that can be applied to classify/predict the output for new scenario. The artificial intelligence-based machine learning approach is rather different from the traditional statistical data analysis approaches; however, recently such methods have been applied to improve early and precise detection of diseases [16,21,25,27-29], including the prediction of outcomes after cardiac surgery [22,32,83].

In our study, we investigated the development of delirium prediction models using long-term (over 5 years) surgical outcomes data for over 5000 patients. We developed several prediction models, while addressing the underlying class imbalance issue, and compared their performance on an independent test set. Except for SVM (ROC-AUC=71.7%), the ROC-AUC of the predictive models was at least 75%, indicating a good general performance by predicting the correct classification most of the time [84,85]. Using the F1 score and the PRC-AUC, which are more sensitive to class imbalance, we were able to demonstrate that the SVM followed by the BBN models offered the best prediction performance in correctly identifying adult patients at risk of developing agitated delirium after cardiac surgery (F1 score: 40.2 and 34.4 and PRC-AUC: 30.7 and 29.6; respectively).

Our predictive models had a worse performance when compared with the findings of Kumar et al [39] (ROC-AUC of the RF model ~91%). Although they argue that their data were imbalanced, they used the ROC-AUC as their evaluation metric, which does not consider the class imbalance. On the other hand, PRC-AUC inherently accounts for class distribution (the

XSL•FO
RenderX

probability is conditioned on the model estimate of the class label, which will vary if the model is applied on a population with different baseline distributions). It is more useful if the goal is improving the prediction of *positive* class in an imbalanced population with known baseline probability (eg, document retrieval, fraud detection, and medical complications) [44,46,48,51,66].

Compared with the findings of Corradi et al [42] (ROC-AUC of the RF model ~91% and PRC-AUC ~61 %), our model was worse. Although they included a lot of physiological parameters, they did not include any laboratory parameters. In addition, they applied the algorithm on all patients within the study period (medical and surgical). Most of the variables used were correlated—that is, they were a function of each other (eg, RASS and mechanical ventilation, RASS score and vasopressors, and dementia and the Charleston Comorbidity Index)—which likely impacted the generalizability of the model.

The paper published by Davoudi et al [40] is the only paper that is closely related to our work as they were specifically addressing the question of predicting delirium after major surgery and had a large cohort of patients who underwent cardiothoracic surgery (6890 patients, 13%). They were able to achieve an ROC-AUC ranging from 79% to 86%, which was close to the ROC-AUC we were able to achieve (71.7%-78%). Unfortunately, it is not clear what type of delirium they were capturing and the urgency of surgery these patients were undergoing. Also, only 13% of these patients underwent cardiothoracic surgery. They mainly relied on the ROC-AUC to compare the model's performance, which is insensitive to the target class imbalance.

Lee et al [41] conducted a unique systematic review in 2017, addressing the issue of predictive models for discovering delirium after cardiac surgery. They were only able to identify 3 high-quality models (Katznelson, Original PRE-DELIRIC, and the recalibrated PRE-DELIRIC). As the original PRE-DELIRIC was recently externally validated, they externally validated the Katznelson and recalibrated PRE-DELIRIC model on a local population dataset of 600 patients. Several metrics were used to evaluate the model's discrimination and calibration. All metrics for recalibrated PRE-DELIRIC model outperformed the Katznelson model (see Multimedia Appendix 1). However, these metrics cannot distinguish clinical utility. To identify clinical utility of these models, they performed DCA to ascertain the clinical utility of each model. The main advantage of DCA is that it incorporates preferences (patient or physician) represented as threshold probability of choosing or denying a treatment, across a range of probabilities [41]. The net benefit (the expected benefit of offering or denying a treatment at that threshold) of each algorithm was evaluated. Based on the DCA analysis, both models had limited clinical utility, with the recalibrated PRE-DELIRIC having marginally better performance at low thresholds between 20% and 40%. Regrettably, they used already validated models that are based on LR. They mentioned very limited information about the validation cohort (such as mean age, gender distribution, and type of cardiac surgery). In addition, they did not address the significant class imbalance (delirium=13.8%). Finally, the use of DCA to evaluate clinical utility of the models is very innovative but it can be only applied to evaluate models that were developed by the same algorithm but have different parameters. Its applicability across different modeling algorithms is still not clear. One of the essential assumptions of DCA is that the predicted probability and threshold probability are independent. In the case of delirium, it would be very difficult to assert that independence, as delirium is multifactorial, and there is no clear mechanism to its development. Violating this assumption might significantly affect the results and interpretation of the DCA.

To our knowledge, this is the first paper that explicitly attempts to develop several predictive models using machine learning methodology and compare their performance for the sole purpose of proactively predicting agitated delirium in adult patients undergoing cardiac surgery. A notable aspect of our work is the use of multiple performance evaluation measures to evaluate the different facets of a prediction model with respect to its prediction performance. We demonstrated the importance of using different metrics when analyzing model's performance (eg, F1 score and PRC-AUC) and the importance of visual analysis of the curves across different probabilities (eg, PRC). Using a static or single measure, like ROC-AUC or accuracy, might lead to false assumptions and incorrect decisions, especially in the presence of class imbalance in the dataset [66].

An important factor in the selection of a prediction model is its interpretability (clarity) to the users (especially health care providers) who are particularly keen to know the basis for a recommendation/decision when it is derived from a computational model. One of the drawbacks of ANN and SVM is that they are not easy to explain, that is, how the output was produced (ie, they are regarded as black box models). This inability to explain the model and its predictions tends to raise a degree of skepticism among health care practitioners regarding the prediction produced [46,52,56]. However, the application of additional methods to decipher the ANN and SVM models' decision logic in terms of understandable production rules that illustrate a correlation between clinical attribute values and the output class can increase their acceptance and subsequent use by medical practitioners [75-80]. Other machine learning methods, such as the BBN model provides a simple but elegant graphical representation of the problem space that can be interpreted by health care professionals.

Predicting delirium is a challenging problem, but with a significant health outcome and system use impact. Given the complexity of how and why delirium manifests in certain patients, the ability to correctly identify if not all but even a fair number of the potential patients who are at risk of developing delirium will be a significant improvement from the current state where patients are diagnosed with delirium only after it starts, and hence, the administration of appropriate interventions is delayed. To address this challenging problem, we investigated the application of machine learning methods to predict postoperative delirium after cardiac surgery. Our methodology involved addressing the target class imbalance and employing appropriate evaluation metrics to measure the prediction performance from a clinical utility perspective. We argue that with the increased use of eHealth records and auxiliary data collection tools, the volume of health data being collected is

XSL•FO
RenderX

reaching the level of *big data*. This brings relief to the need to apply advance machine learning techniques to analyze the data for improved and effective data-driven decision support [86,87] that would enable timely intervention for negative outcomes [5,56,87] to improve health outcomes and in turn enhance patient safety and satisfaction.

Limitations

We recognize that our study has certain limitations. First, as postoperative complications (including delirium) in our database are captured as binary outcomes (yes/no) but without a time stamp, it was hard to determine if agitated delirium was a secondary phenomenon (eg, because of infection, uncontrolled pain, and prolonged mechanical ventilation) or because of a pre-existing medical comorbidity. Second, the prevalence of agitated delirium was only 11.4%. This low representation is most likely because of the definition of delirium in the source database (only agitated subtype). This can potentially limit the ability to generalize the developed models to other types of delirium [10,11]. Third, there exist more advance machine learning software than what were available in WEKA, but we chose WEKA because of its open source, flexibility, and ease of use [54]; and finally, the study is based on a retrospective design and hence may suffer from the pitfalls associated with such a design.

Clinical Equipoise and Key Messages

The key messages of this paper are as follows:

- From a clinical standpoint:
  - Patients undergoing cardiovascular surgical procedures are at higher risk of developing agitated delirium due to several factors, including surgical complexity, comorbidities, and age [7,8].
  - Preventing delirium should be the goal, especially if patients at risk were identified. This will mitigate its negative sequalae and improve the patient's quality of life. Some of the proposed preventive interventions that have been shown to reduce the incidence of delirium in high-risk patients include early mobilization, use of patient's personal aids (reading glasses, hearing aid, etc), pharmacological interventions (the use of less sedatives and addressing pain), and improving sleep environment especially in the intensive care [38,88-91].
- From a predictive modeling perspective:
  - Addressing class imbalance on the training dataset (a common feature of medical datasets) could enhance the machine learning model's performance in identifying patients likely to develop postoperative delirium.
  - Keeping an open mind and exploring different modeling methodologies will enable the selection of the most appropriate model that can generate the best results.
  - The PRC offers a more intuitive and direct measure of the model performance that is representative of its true performance, especially in the presence of class imbalance.

Conclusions and Future Research

Postoperative agitated delirium is associated with major morbidity that impacts the patient postoperative recovery. Cardiac surgery patients are at high risk of developing postoperative delirium. To improve health outcomes of cardiac surgery, the current approach to address the effects of delirium is a preventive program of care [88-91], such as ABCDE, which involves awakening and breathing coordination for liberation from sedation and mechanical ventilation, choosing sedatives that are less likely to increase risk of delirium, delirium management, and finally, early mobility and exercise [36]. As much as the ABCDE approach provides a road map of how to manage delirium, it does not provide mechanisms to identify patients at risk of developing delirium. Hence, the ABCDE approach serves as an after-the-event management strategy, while leaving a gap in terms of a proactive prevention strategy for delirium. Our ability to predict delirium in patients, and in turn proactively administer therapeutic and behavioral therapies to mitigate the negative effects of delirium, will lead to significant improvements in health outcomes, patient satisfaction and quality of life, and health system cost saving.

In this study, we pursued the development of prediction models using preoperative clinical data to establish a mapping between the patient's preoperative clinical variables and the onset of postoperative delirium. We investigated machine learning methods to develop a viable postoperative delirium prediction model which can be operationalized in a clinical setting as a delirium screening tool to proactively identify patients at risk of developing postcardiac surgery agitated delirium. We posit that the use and operationalization of delirium predictive model can significantly reduce the incidence of delirium by enabling the administration of preventive measures in a timely manner. In this paper, we presented work detailing the development of data-driven delirium prediction models with a reasonable accuracy. Furthermore, the work contributes 3 findings that are useful for future efforts to develop advanced delirium prediction models—that is, (1) addressing class imbalance on the training dataset will enhance the machine learning model's performance in identifying patients likely to develop postoperative delirium, (2) when evaluating the model's performance, selecting unsuitable measures can influence model interpretation and its utility, and (3) the PRC offers a more intuitive and direct measure of the model's performance that is representative of its true performance, especially in the presence of class imbalance.

In our future research, we will attempt to apply feature extraction to identify key features to enhance the model's performance. At the same time, we will attempt to isolate modifiable features that are clinically relevant so that personalized interventions can be started in a timely fashion. We will also attempt to apply evolutionary computations to optimize classifiers parameters. Another interesting application is the use of deep learning methods to create new features or feature sets to boost the model's performance and accuracy.

In conclusion, we argue that any improvement in our ability to predict delirium using prediction models, even if numerically small, is of consequential clinical significance—this situation

XSL•FO
RenderX

is like 2 drugs that have the same treatment profile, but one drug has fewer side effects, and hence, the ability to precisely select the right drug has an impact on patient safety. When dealing with complex medical problems, such as delirium, we posit that the application of advanced machine learning methods might actually improve disease prediction capabilities which in turn will enhance opportunities for preventive, personalized, and precise medical interventions that would improve the patient's quality of life after surgery.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' Contributions

HNM conceived the presented research question, devised the project and the main conceptual ideas, conducted the literature review, applied for research ethics, designed and performed the experiments, analyzed the data, derived the predictive models, assessed their performance on the test dataset, and took the lead in writing the manuscript. SSRA and SRA verified the analytical methods. GMH verified clinical relevance and literature review. SSRA, SRA, and GMH equally cosupervised this work. All authors discussed the results and reviewed to the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of relevant related work.
[DOCX File , 43 KB - medinform_v7i4e14993_app1.docx]

Multimedia Appendix 2
Machine learning methodology and forest plot for logistic regression.
[DOCX File , 3549 KB - medinform_v7i4e14993_app2.docx]

Multimedia Appendix 3
Setting of the prediction models and the optimization steps that were applied on the used algorithms.
[DOCX File , 28 KB - medinform_v7i4e14993_app3.docx]

References

1. Koster S, Oosterveld FG, Hensens AG, Wijma A, van der Palen J. Delirium after cardiac surgery and predictive validity of a risk checklist. Ann Thorac Surg 2008 Dec;86(6):1883-1887. [doi: 10.1016/j.athoracsur.2008.08.020] [Medline: 19022003]
2. The Society of Thoracic Surgeons. STS National Database URL: https://www.sts.org/registries-research-center/sts-national-database [accessed 2019-10-08]
3. American Psychiatric Association. Practice Guideline for the Treatment of Patients with Delirium. Washington, DC: American Psyciatric Assosiation; 2010.
4. American Psychiatric Association. Diagnostic and statistical manual of mental disorders : DSM-52013. In: Diagnostic And Statistical Manual Of Mental Disorders. Fifth Edition. Washington, DC: American Psychiatric Publishing; 2013.
5. Royston D, Cox F. Anaesthesia: the patient's point of view. Lancet 2003 Nov 15;362(9396):1648-1658. [doi: 10.1016/S0140-6736(03)14800-3] [Medline: 14630448]
6. Martin B, Buth KJ, Arora RC, Baskett RJ. Delirium as a predictor of sepsis in post-coronary artery bypass grafting patients: a retrospective cohort study. Crit Care 2010;14(5):R171 [FREE Full text] [doi: 10.1186/cc9273] [Medline: 20875113]
7. Martin B, Buth KJ, Arora RC, Baskett RJ. Delirium: a cause for concern beyond the immediate postoperative period. Ann Thorac Surg 2012 Apr;93(4):1114-1120. [doi: 10.1016/j.athoracsur.2011.09.011] [Medline: 22200370]
8. Gottesman R, Grega M, Bailey M, Pham L, Zeger S, Baumgartner W, et al. Delirium after coronary artery bypass graft surgery and late mortality. Ann Neurol 2010 Mar;67(3):338-344 [FREE Full text] [doi: 10.1002/ana.21899] [Medline: 20373345]
9. Smulter N, Lingehall HC, Gustafson Y, Olofsson B, Engström KG. Delirium after cardiac surgery: incidence and risk factors. Interact Cardiovasc Thorac Surg 2013 Nov;17(5):790-796 [FREE Full text] [doi: 10.1093/icvts/ivt323] [Medline: 23887126]
10. Cavallazzi R, Saad M, Marik PE. Delirium in the ICU: an overview. Ann Intensive Care 2012 Dec 27;2(1):49 [FREE Full text] [doi: 10.1186/2110-5820-2-49] [Medline: 23270646]

XSL•FO
RenderX

11. Andrejaitiene J, Sirvinskas E. Early post-cardiac surgery delirium risk factors. Perfusion 2012 Mar;27(2):105-112. [doi: 10.1177/0267659111425621] [Medline: 22170877]
12. Afonso A, Scurlock C, Reich D, Raikhelkar J, Hossain S, Bodian C, et al. Predictive model for postoperative delirium in cardiac surgical patients. Semin Cardiothorac Vasc Anesth 2010 Sep;14(3):212-217. [doi: 10.1177/1089253210374650] [Medline: 20647262]
13. Bakker RC, Osse R, Tulen J, Kappetein A, Bogers A. Preoperative and operative predictors of delirium after cardiac surgery in elderly patients. Eur J Cardiothorac Surg 2012 Mar;41(3):544-549. [doi: 10.1093/ejcts/ezr031] [Medline: 22345177]
14. Stransky M, Schmidt C, Ganslmeier P, Grossmann E, Haneya A, Moritz S, et al. Hypoactive delirium after cardiac surgery as an independent risk factor for prolonged mechanical ventilation. J Cardiothorac Vasc Anesth 2011 Dec;25(6):968-974. [doi: 10.1053/j.jvca.2011.05.004] [Medline: 21741272]
15. Stoltzfus JC. Logistic regression: a brief primer. Acad Emerg Med 2011 Oct;18(10):1099-1104 [FREE Full text] [doi: 10.1111/j.1553-2712.2011.01185.x] [Medline: 21996075]
16. Koh HC, Tan G. Data mining applications in healthcare. J Healthc Inf Manag 2005;19(2):64-72. [Medline: 15869215]
17. Lisboa PJ. A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Netw 2002 Jan;15(1):11-39. [doi: 10.1016/s0893-6080(01)00111-3] [Medline: 11958484]
18. Bertsimas D, Bjarnadóttir MV, Kane MA, Kryder JC, Pandey R, Vempala S, et al. Algorithmic prediction of health-care costs. Oper Res 2008;56(6):1382-1392. [doi: 10.1287/opre.1080.0619]
19. Bell LM, Grundmeier R, Localio R, Zorc J, Fiks AG, Zhang X, et al. Electronic health record-based decision support to improve asthma care: a cluster-randomized trial. Pediatrics 2010 Apr;125(4):e770-e777. [doi: 10.1542/peds.2009-1385] [Medline: 20231191]
20. Tomar D, Agarwal S. A survey on Data Mining approaches for Healthcare. Int J Bio Sci Bio Tech 2013 Oct 31;5(5):241-266. [doi: 10.14257/ijbsbt.2013.5.5.25]
21. Fei Y, Hu J, Gao K, Tu J, Li W, Wang W. Predicting risk for portal vein thrombosis in acute pancreatitis patients: a comparison of radical basis function artificial neural network and logistic regression models. J Crit Care 2017 Jun;39:115-123. [doi: 10.1016/j.jcrc.2017.02.032] [Medline: 28246056]
22. Santelices LC, Wang Y, Severyn D, Druzdzel MJ, Kormos RL, Antaki JF. Development of a hybrid decision support model for optimal ventricular assist device weaning. Ann Thorac Surg 2010 Sep;90(3):713-720 [FREE Full text] [doi: 10.1016/j.athoracsur.2010.03.073] [Medline: 20732482]
23. Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann Intern Med 1991 Dec 1;115(11):843-848. [doi: 10.7326/0003-4819-115-11-843] [Medline: 1952470]
24. Artis SG, Mark R, Moody G. Detection of Atrial Fibrillation Using Artificial Neural Networks. In: Proceedings Computers in Cardiology. 1991 Presented at: CinC'91; September 23-26, 1991; Venice, Italy, Italy. [doi: 10.1109/cic.1991.169073]
25. Eftekhar B, Mohammad K, Ardebili HE, Ghodsi M, Ketabchi E. Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC Med Inform Decis Mak 2005 Feb 15;5:3 [FREE Full text] [doi: 10.1186/1472-6947-5-3] [Medline: 15713231]
26. Belciug S, Gorunescu F, Salem AB, Gorunescu M. Clustering-Based Approach for Detecting Breast Cancer Recurrence. In: Proceedings of the 2010 10th International Conference on Intelligent Systems Design and Applications. Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on; 2010 Presented at: ISDA'10; November 29-December 1 2010; Cairo, Egypt. [doi: 10.1109/isda.2010.5687211]
27. Salama GI, Abdelhalim M, Zeid MA. Breast cancer diagnosis on three different datasets using multi-classifiers. Int J Comp Inf Technol 2012;1(1):36-43 [FREE Full text]
28. Ayer T, Chhatwal J, Alagoz O, Kahn CE, Woods RW, Burnside ES. Informatics in radiology: comparison of logistic regression and artificial neural network models in breast cancer risk estimation. Radiographics 2010 Jan;30(1):13-22 [FREE Full text] [doi: 10.1148/rg.301095057] [Medline: 19901087]
29. Joshi S, Shenoy D, Vibhudendra SG, Rrashmi P, Venugopal K, Patnaik L. Classification of Alzheimer's Disease and Parkinson's Disease by Using Machine Learning and Neural Network Methods. In: Proceedings of the 2010 Second International Conference on Machine Learning and Computing. 2010 Presented at: ICMLC'10; February 9-11, 2010; Singapore p. 218-222. [doi: 10.1109/icmlc.2010.45]
30. Escudero J, Zajicek J, Ifeachor E. Early detection and characterization of Alzheimer's disease in clinical scenarios using Bioprofile concepts and K-means. Conf Proc IEEE Eng Med Biol Soc 2011;2011:6470-6473. [doi: 10.1109/IEMBS.2011.6091597] [Medline: 22255820]
31. Ramani RG, Sivagami G. Parkinson disease classification using data mining algorithms. Int J Comp App 2011;32(9):17-22. [doi: 10.5120/3932-5571]
32. Nilsson J, Ohlsson M, Thulin L, Höglund P, Nashef SA, Brandt J. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. J Thorac Cardiovasc Surg 2006 Jul;132(1):12-19 [FREE Full text] [doi: 10.1016/j.jtcvs.2005.12.055] [Medline: 16798296]
33. Kazemi Y, Mirroshandel SA. A novel method for predicting kidney stone type using ensemble learning. Artif Intell Med 2018 Jan;84:117-126. [doi: 10.1016/j.artmed.2017.12.001] [Medline: 29241659]

XSL•FO
RenderX

34. Edelstein P. Emerging directions in analytics. Predictive analytics will play an indispensable role in healthcare transformation and reform. Health Manag Technol 2013 Jan;34(1):16-17. [Medline: 23420986]
35. Moradi M, Ghadiri N. Different approaches for identifying important concepts in probabilistic biomedical text summarization. Artif Intell Med 2018 Jan;84:101-116. [doi: 10.1016/j.artmed.2017.11.004] [Medline: 29208328]
36. Brummel NE, Girard TD. Preventing delirium in the intensive care unit. Crit Care Clin 2013 Jan;29(1):51-65 [FREE Full text] [doi: 10.1016/j.ccc.2012.10.007] [Medline: 23182527]
37. Reade MC, Finfer S. Sedation and delirium in the intensive care unit. N Engl J Med 2014 Jan 30;370(5):444-454. [doi: 10.1056/NEJMra1208705] [Medline: 24476433]
38. Ettema RG, van Koeven H, Peelen LM, Kalkman CJ, Schuurmans MJ. Preadmission interventions to prevent postoperative complications in older cardiac surgery patients: a systematic review. Int J Nurs Stud 2014 Feb;51(2):251-260. [doi: 10.1016/j.ijnurstu.2013.05.011] [Medline: 23796313]
39. Kramer D, Veeranki S, Hayn D, Quehenberger F, Leodolter W, Jagsch C, et al. Development and validation of a multivariable prediction model for the occurrence of delirium in hospitalized gerontopsychiatry and internal medicine patients. Stud Health Technol Inform 2017;236:32-39. [Medline: 28508776]
40. Davoudi A, Ozrazgat-Baslanti T, Ebadi A, Bursian A, Bihorac A, Rashidi P. Delirium Prediction using Machine Learning Models on Predictive Electronic Health Records Data. In: Proceedings of the 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering. 2017 Presented at: BIBE'17; October 23-25, 2017; Washington, DC, USA. [doi: 10.1109/bibe.2017.00014]
41. Lee A, Mu J, Joynt G, Chiu C, Lai V, Gin T, et al. Risk prediction models for delirium in the intensive care unit after cardiac surgery: a systematic review and independent external validation. Br J Anaesth 2017 Mar 1;118(3):391-399 [FREE Full text] [doi: 10.1093/bja/aew476] [Medline: 28186224]
42. Corradi JP, Thompson S, Mather JF, Waszynski CM, Dicks RS. Prediction of incident delirium using a random forest classifier. J Med Syst 2018 Nov 14;42(12):261. [doi: 10.1007/s10916-018-1109-0] [Medline: 30430256]
43. Global Biodiversity Information Facility. 2014. R: a language and environment for statistical computing URL: https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing [accessed 2019-10-08]
44. Hastie T, Tibshirani R, Friedman J. The Elements Of Statistical Learning: Data Mining, Inference, And Prediction. New York: Springer; 2008.
45. Breiman L. Random forests. Mach Learn 2001;45(1):5-32 [FREE Full text]
46. Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools And Techniques. Third Edition. Burlington, Massachusetts: Morgan Kaufmann; 2011.
47. Ganganwar V. An overview of classification algorithms for imbalanced datasets. Int J Emerg Technol Adv Eng 2012;2:42-47 [FREE Full text]
48. Chawla NW. Data mining for imbalanced datasets: An overview. In: Data Mining and Knowledge Discovery Handbook. New York: Springer; 2005:853-867.
49. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009;21(9):1263-1284. [doi: 10.1109/tkde.2008.239]
50. Pyle D. Data Preparation for Data Mining. Burlington, Massachusetts: Morgan Kaufmann; 1999.
51. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R. New York: Springer; 2013.
52. Han J, Kamber M, Jian PP. Data Mining Concepts And Techniques. Burlington, Massachusetts: Morgan Kaufmann; 2011.
53. Refaeilzadeh P, Tang L, Liu H. Cross-validation. Encyclopedia Database Syst 2016:1-7. [doi: 10.1007/978-1-4899-7993-3_565-2]
54. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software. SIGKDD Explor Newsl 2009;11(1):10. [doi: 10.1145/1656274.1656278]
55. Cooper CG, Dash D, Levander J, Wong W, Hogan W, Wagner M. Bayesian biosurveillance of disease outbreaks. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence. 2004 Presented at: UAI'04; July 7-11, 2004; Banff, Canada p. 94-103.
56. Tufféry S. Data Mining and Statistics for Decision Making. Hoboken, New Jersey: Wiley; 2011.
57. Moon K, Jin Y, Jin T, Lee S. Development and validation of an automated delirium risk assessment system (Auto-DelRAS) implemented in the electronic health record system. Int J Nurs Stud 2018 Jan;77:46-53. [doi: 10.1016/j.ijnurstu.2017.09.014] [Medline: 29035732]
58. Mao Y, Chen Y, Hackmann G, Chen M, Lu C, Kollef M. Early Deterioration Warning for Hospitalized Patients by Mining Clinical Data. Int J Knowl Discov Bioinformatics 2011;2:1-20. [doi: 10.4018/jkdb.2011070101]
59. Watt EW, Bui AA. Evaluation of a dynamic bayesian belief network to predict osteoarthritic knee pain using data from the osteoarthritis initiative. AMIA Annu Symp Proc 2008 Nov 6:788-792 [FREE Full text] [Medline: 18999030]
60. Krishna G, Kumar B, Orsu N, B S. Performance analysis and evaluation of different data mining algorithms used for cancer classification. Int J Adv Res Artif Intell 2013;2(5). [doi: 10.14569/ijarai.2013.020508]
61. Zhou F, Jin L, Dong J. Premature ventricular contraction detection combining deep neural networks and rules inference. Artif Intell Med 2017 Jun;79:42-51. [doi: 10.1016/j.artmed.2017.06.004] [Medline: 28662816]

XSL•FO
RenderX

62. Haddawy P, Hasan AI, Kasantikul R, Lawpoolsri S, Sa-Angchai P, Kaewkungwal J, et al. Spatiotemporal Bayesian networks for malaria prediction. Artif Intell Med 2018 Jan;84:127-138. [doi: 10.1016/j.artmed.2017.12.002] [Medline: 29241658]
63. Boucekine M, Loundou A, Baumstarck K, Minaya-Flores P, Pelletier J, Ghattas B, et al. Using the random forest method to detect a response shift in the quality of life of multiple sclerosis patients: a cohort study. BMC Med Res Methodol 2013 Feb 15;13:20 [FREE Full text] [doi: 10.1186/1471-2288-13-20] [Medline: 23414459]
64. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer; 2008.
65. Murphy KP. Machine Learning: A Probabilistic Perspective. Cambridge: MIT Press; 2012.
66. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
67. Inouye SK, Charpentier PA. Precipitating factors for delirium in hospitalized elderly persons. Predictive model and interrelationship with baseline vulnerability. J Am Med Assoc 1996 Mar 20;275(11):852-857. [doi: 10.1001/jama.1996.03530350034031] [Medline: 8596223]
68. Inouye SK, Viscoli CM, Horwitz RI, Hurst LD, Tinetti ME. A predictive model for delirium in hospitalized elderly medical patients based on admission characteristics. Ann Intern Med 1993 Sep 15;119(6):474-481. [doi: 10.7326/0003-4819-119-6-199309150-00005] [Medline: 8357112]
69. O'Keeffe ST, Lavan JN. Predicting delirium in elderly patients: development and validation of a risk-stratification model. Age Ageing 1996 Jul;25(4):317-321. [doi: 10.1093/ageing/25.4.317] [Medline: 8831879]
70. van den Boogaard M, Pickkers P, Slooter AJ, Kuiper MA, Spronk PE, van der Voort PH, et al. Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICu patients) delirium prediction model for intensive care patients: observational multicentre study. Br Med J 2012 Feb 9;344:e420 [FREE Full text] [doi: 10.1136/bmj.e420] [Medline: 22323509]
71. Katznelson R, Djaiani GN, Borger MA, Friedman Z, Abbey SE, Fedorko L, et al. Preoperative use of statins is associated with reduced early delirium rates after cardiac surgery. Anesthesiology 2009 Jan;110(1):67-73. [doi: 10.1097/ALN.0b013e318190b4d9] [Medline: 19104172]
72. Isfandiaty R, Harimurti KF, Setiati SF, Roosheroe A. Incidence and predictors for delirium in hospitalized elderly patients: a retrospective cohort study. Acta Med Indones 2012 Oct;44(4):290-297 [FREE Full text] [Medline: 23314969]
73. Carrasco MP, Villarroel L, Andrade M, Calderón J, González M. Development and validation of a delirium predictive score in older people. Age Ageing 2014 May;43(3):346-351. [doi: 10.1093/ageing/aft141] [Medline: 24064236]
74. Chaiwat O, Chanidnuan M, Pancharoen W, Vijitmala K, Danpornprasert P, Toadithep P, et al. Postoperative delirium in critically ill surgical patients: incidence, risk factors, and predictive scores. BMC Anesthesiol 2019 Mar 20;19(1):39 [FREE Full text] [doi: 10.1186/s12871-019-0694-x] [Medline: 30894129]
75. Lisboa PJ. Interpretability in Machine Learning – Principles and Practice. In: Proceedings of the International Workshop on Fuzzy Logic and Applications. 2013 Presented at: WILF'13; November 19-22, 2013; Genoa, Italy p. 15-21.
76. Goodman B, Flaxman S. European union regulations on algorithmic decision-making and a 'Right to Explanation'. AI Mag 2017;38(3):50-57. [doi: 10.1609/aimag.v38i3.2741]
77. Vellido A, Martín-Guerrero J, Lisboa PJ. CiteSeer. 2012. Making Machine Learning Models Interpretable URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.431.5382 [accessed 2019-10-08]
78. Intrator O, Intrator N. Interpreting neural-network results: a simulation study. Comput Stat Data Anal 2001;37(3):373-393. [doi: 10.1016/s0167-9473(01)00016-0]
79. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv 2019;51(5):1-42. [doi: 10.1145/3236009]
80. Montavon G, Samek W, Müller K. Methods for interpreting and understanding deep neural networks. Digit Signal Process 2018;73:1-15. [doi: 10.1016/j.dsp.2017.10.011]
81. Buth KJ, Gainer RA, Legare J, Hirsch GM. The changing face of cardiac surgery: practice patterns and outcomes 2001-2010. Can J Cardiol 2014 Feb;30(2):224-230. [doi: 10.1016/j.cjca.2013.10.020] [Medline: 24373760]
82. Mikalsen K, Soguero-Ruiz C, Jensen K, Hindberg K, Gran M, Revhaug A, et al. Using anchors from free text in electronic health records to diagnose postoperative delirium. Comput Methods Programs Biomed 2017 Dec;152:105-114. [doi: 10.1016/j.cmpb.2017.09.014] [Medline: 29054250]
83. Wise ES, Hocking KM, Brophy CM. Prediction of in-hospital mortality after ruptured abdominal aortic aneurysm repair using an artificial neural network. J Vasc Surg 2015 Jul;62(1):8-15 [FREE Full text] [doi: 10.1016/j.jvs.2015.02.038] [Medline: 25953014]
84. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27(8):861-874. [doi: 10.1016/j.patrec.2005.10.010]
85. Fawcett T. ROC graphs: Notes and practical considerations for researchers. Mach Learn 2004;31:1-38 [FREE Full text]
86. O'Connor PJ, Sperl-Hillen JM, Rush WA, Johnson PE, Amundson GH, Asche SE, et al. Impact of electronic health record clinical decision support on diabetes care: a randomized trial. Ann Fam Med 2011;9(1):12-21 [FREE Full text] [doi: 10.1370/afm.1196] [Medline: 21242556]

XSL•FO
RenderX

87. Berwick DM, Hackbarth AD. Eliminating waste in US health care. J Am Med Assoc 2012 Apr 11;307(14):1513-1516. [doi: 10.1001/jama.2012.362] [Medline: 22419800]
88. Hsieh SJ, Ely EW, Gong MN. Can intensive care unit delirium be prevented and reduced? Lessons learned and future directions. Ann Am Thorac Soc 2013 Dec;10(6):648-656 [FREE Full text] [doi: 10.1513/AnnalsATS.201307-232FR] [Medline: 24364769]
89. O'Hanlon S, O'Regan N, Maclullich AM, Cullen W, Dunne C, Exton C, et al. Improving delirium care through early intervention: from bench to bedside to boardroom. J Neurol Neurosurg Psychiatry 2014 Feb;85(2):207-213. [doi: 10.1136/jnnp-2012-304334] [Medline: 23355807]
90. Cerejeira J, Mukaetova-Ladinska E. A clinical update on delirium: from early recognition to effective management. Nurs Res Pract 2011;2011:875196 [FREE Full text] [doi: 10.1155/2011/875196] [Medline: 21994844]
91. Trogrlić Z, van der Jagt M, Bakker J, Balas MC, Ely EW, van der Voort PH, et al. A systematic review of implementation strategies for assessment, prevention, and management of ICU delirium and their effect on clinical outcomes. Crit Care 2015 Apr 9;19:157 [FREE Full text] [doi: 10.1186/s13054-015-0886-9] [Medline: 25888230]

Abbreviations

ANN: artificial neural networks
BBN: Bayesian belief networks
CABG: coronary artery bypass graft
CAM: confusion assessment method
CAM-ICU: confusion assessment method for the intensive care unit
DCA: decision curve analysis
DT: decision trees
eHealth: electronic health
ICU: intensive care unit
LR: logistic regression
MHC: Maritime Heart Center
NB: naïve Bayesian
OOB: out-of-bag
OR: odds ratio
PPV: positive predictive value (precision)
PRC-AUC: precision-recall curve-area under the curve
QEII HSC: Queen Elizabeth II Health Sciences Center
RASS: Richmond Agitation-Sedation Scale
RF: random forest
ROC-AUC: receiver operator curve-area under the curve
RUS: random majority undersampling
SVM: support vector machines
WEKA: Waikato Environment for Knowledge Acquisition

Edited by G Eysenbach; submitted 11.06.19; peer-reviewed by A Davoudi, D Carvalho, D Surian, B Polepalli Ramesh; comments to author 09.07.19; revised version received 02.09.19; accepted 24.09.19; published 23.10.19.

Please cite as:
Mufti HN, Hirsch GM, Abidi SR, Abidi SSR
Exploiting Machine Learning Algorithms and Methods for the Prediction of Agitated Delirium After Cardiac Surgery: Models Development and Validation Study
JMIR Med Inform 2019;7(4):e14993
URL: http://medinform.jmir.org/2019/4/e14993/
doi:10.2196/14993
PMID:31558433

©Hani Nabeel N Mufti, Gregory Marshal Hirsch, Samina Raza Abidi, Syed Sibte Raza Abidi. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 23.10.2019. 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

XSL•FO
RenderX

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

XSL•FO
RenderX

Original Paper

A Bayesian Network Analysis of the Diagnostic Process and its Accuracy to Determine How Clinicians Estimate Cardiac Function in Critically Ill Patients: Prospective Observational Cohort Study

Thomas Kaufmann[1*], MD; José Castela Forte[1,2,3,4*], BSc; Bart Hiemstra[1,2], MD, PhD; Marco A Wiering[4], PhD; Marco Grzegorczyk[4], PhD; Anne H Epema[1], MD, PhD; Iwan C C van der Horst[2,5], MD, PhD; SICS Study Group

[1]Department of Anesthesiology, University Medical Center Groningen, University of Groningen, Groningen, Netherlands

[2]Department of Critical Care, University Medical Center Groningen, University of Groningen, Groningen, Netherlands

[3]Department of Clinical Pharmacology, University Medical Center Groningen, University of Groningen, Groningen, Netherlands

[4]Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Groningen, Netherlands

[5]Department of Intensive Care, Maastricht University Medical Center+, Maastricht University, Maastricht, Netherlands

[*]these authors contributed equally

Corresponding Author:
Thomas Kaufmann, MD
Department of Anesthesiology
University Medical Center Groningen
University of Groningen
Hanzeplein 1
PO Box 30.001
Groningen, 9700 RB
Netherlands
Phone: 31 503616161
Email: t.kaufmann@umcg.nl

Abstract

Background: Hemodynamic assessment of critically ill patients is a challenging endeavor, and advanced monitoring techniques are often required to guide treatment choices. Given the technical complexity and occasional unavailability of these techniques, estimation of cardiac function based on clinical examination is valuable for critical care physicians to diagnose circulatory shock. Yet, the lack of knowledge on how to best conduct and teach the clinical examination to estimate cardiac function has reduced its accuracy to almost that of "flipping a coin."

Objective: The aim of this study was to investigate the decision-making process underlying estimates of cardiac function of patients acutely admitted to the intensive care unit (ICU) based on current standardized clinical examination using Bayesian methods.

Methods: Patient data were collected as part of the Simple Intensive Care Studies-I (SICS-I) prospective cohort study. All adult patients consecutively admitted to the ICU with an expected stay longer than 24 hours were included, for whom clinical examination was conducted and cardiac function was estimated. Using these data, first, the probabilistic dependencies between the examiners' estimates and the set of clinically measured variables upon which these rely were analyzed using a Bayesian network. Second, the accuracy of cardiac function estimates was assessed by comparison to the cardiac index values measured by critical care ultrasonography.

Results: A total of 1075 patients were included, of which 783 patients had validated cardiac index measurements. A Bayesian network analysis identified two clinical variables upon which cardiac function estimate is conditionally dependent, namely, noradrenaline administration and presence of delayed capillary refill time or mottling. When the patient received noradrenaline, the probability of cardiac function being estimated as reasonable or good $P(E_{R,G})$ was lower, irrespective of whether the patient was mechanically ventilated ($P[E_{R,G}|$ventilation, noradrenaline$]=0.63$, $P[E_{R,G}|$ventilation, no noradrenaline$]=0.91$, $P[E_{R,G}|$no ventilation, noradrenaline$]=0.67$, $P[E_{R,G}|$no ventilation, no noradrenaline$]=0.93$). The same trend was found for capillary refill time or mottling. Sensitivity of estimating a low cardiac index was 26% and 39% and specificity was 83% and 74% for students

XSL•FO
RenderX

and physicians, respectively. Positive and negative likelihood ratios were 1.53 (95% CI 1.19-1.97) and 0.87 (95% CI 0.80-0.95), respectively, overall.

Conclusions: The conditional dependencies between clinical variables and the cardiac function estimates resulted in a network consistent with known physiological relations. Conditional probability queries allow for multiple clinical scenarios to be recreated, which provide insight into the possible thought process underlying the examiners' cardiac function estimates. This information can help develop interactive digital training tools for students and physicians and contribute toward the goal of further improving the diagnostic accuracy of clinical examination in ICU patients.

Trial Registration: ClinicalTrials.gov NCT02912624; https://clinicaltrials.gov/ct2/show/NCT02912624

(JMIR Med Inform 2019;7(4):e15358) doi:10.2196/15358

KEYWORDS

cardiac function; physical examination; Bayesian network; critical care; ICU; medical education; educated guess; cognition; clinical decision-support; cardiology

Introduction

Background

In hemodynamically unstable patients admitted to the intensive care unit (ICU) for circulatory shock, the diagnosis and treatment decisions initially rely on accurate assessment of clinical examination [1,2]. Shock is the clinical expression of circulatory failure that results in inadequate cellular oxygen utilization and is often accompanied by systemic arterial hypotension, clinical signs of tissue hypoperfusion, and hyperlactatemia [3]. About one-third of critically ill patients experience circulatory shock, which is associated with increased morbidity and mortality [4].

Hemodynamic assessment of critically ill patients is challenging; depending on the type of shock, patients present with highly variable states of circulating blood volume, cardiac contractility, sympathetic nervous activity, vascular tone, and microcirculatory dysfunction. In addition, assessment is even more difficult if comorbidities are present [5]. Currently, hemodynamic estimates based on clinical examination show poor association with cardiac index in both univariate and multivariate analyses, and these estimates are no better than flipping a coin [6]. Due to this limited ability to assess a patient's hemodynamic status using clinical examination, physicians often base changes in treatment primarily on information obtained through advanced monitoring techniques [7]. However, advanced monitoring techniques are currently advised and desired when clinical examination does not lead to a clear diagnosis, or when a patient does not respond to initial therapy [2,8]. Therefore, it is important to place emphasis on improving hemodynamic estimates made with clinical examination, to avoid inappropriate overuse of technological aid [9].

The first step in developing improved clinical examination structures for hemodynamic estimates is to study the current clinical practice. To understand how students and physicians diagnosed low cardiac index, Bayesian networks can be used to gain insight into the thought process behind the educated guess on hemodynamic status.

Bayesian networks have been frequently used to model domain knowledge in the context of decision support in other fields of medicine, given their ability to be interpreted as causal networks when no confounders are present [10-13]. By combining prior knowledge and the uncertainty in data, Bayesian networks allow for inference tasks to be performed, which establish conditional, possibly causal, dependencies between variables [14]. Conditional probabilities queries are interesting tools to study clinical reasoning, which are seen as an additive thought process where, at every step, information is interpreted conditioned on previously acquired information.

Objectives

The aim of this study was to use Bayesian networks to investigate the decision-making process underlying estimates of cardiac function of patients acutely admitted to the ICU, based on current standardized clinical examination using Bayesian methods. Additionally, we aimed to determine the diagnostic accuracy of the current standardized clinical examination for estimating cardiac function in patients acutely admitted to the ICU.

Methods

Design, Setting, and Participants

This study was a predefined substudy of the prospective observational cohort Simple Intensive Care Studies-I (SICS-I) (ClinicalTrial.gov trial registration: NCT02912624) [15]. The study was approved by the local institutional review board (METc M15.168207). In SICS-I, all consecutive, acutely admitted adults expected to stay beyond 24 hours were included on their first day of admission to the ICU. Written informed consent was obtained from all patients or their relatives. This study is reported following the Standards for the Reporting of Diagnostic Accuracy Studies guidelines [16].

Aims

The primary aim was to determine the conditional probabilities relating the variables measured during clinical examination to the cardiac function estimate made by the examiners.

The secondary aim of this study was to assess the diagnostic accuracy of cardiac function estimates made by the examiners and compare them to the cardiac index measured by critical care ultrasonography (CCUS).

XSL•FO
RenderX

Bayesian Network Analysis

Bayesian networks are probabilistic models that represent the conditional (in)dependence relations between a set of variables in the form of a directed acyclic graph. In the graph, each variable is represented as a node and the directed edges (arcs) connecting the nodes represent the conditional dependency relations among the variables. Given the conditional (in)dependencies implied by the directed acyclic graph, the joint probability distribution of all variables can be factorized into a product of simpler local probability distributions.

From the initial set of variables registered during clinical examination, 14 clinical variables available from bedside monitors and patient record files, perfusors, physical examination, and the cardiac function estimate were included for modeling (Multimedia Appendix 1). All continuous variables were discretized according to the definitions provided in the study protocol. The correlation coefficients between variables after discretization were calculated with the Cramér V test for correlation strength.

The network structure was learned using the Max-Min Hill-Climbing algorithm with the Bayesian-Dirichlet equivalent scoring metric, as implemented in the R package "bnlearn" [17]. The Max-Min Hill-Climbing algorithm searches for the best network structure (ie, the best directed acyclic graph) that maximizes the Bayesian-Dirichlet equivalent scoring metric. To this end, the algorithm starts with an initial directed acyclic graph and then improves the Bayesian-Dirichlet equivalent score by iteratively adding, deleting, and reversing individual edges until the Bayesian-Dirichlet equivalent score does not improve further [18].

A set of restrictions can be applied to enforce certain connections between arcs in the network, so that prior knowledge is implemented *a priori* [13]. Arcs representing known dependencies can be whitelisted (ie, forced to appear in the directed acyclic graph), while arcs that represent impossible dependencies can be blacklisted (ie, excluded from the directed acyclic graph). In this network, *age* and *gender* are not determined by any other variables, so all arcs from other variables to these two were blacklisted. Similarly, as *estimate* does not influence any clinical variable, any arc from *estimate* to other variables was also blacklisted.

After the restrictions are defined, to obtain a confidence measure for the presence and directionality of the individual network edges, the bootstrap technique was applied. R=2000 bootstrap samples were generated from the original data, and the Max-Min Hill-Climbing algorithm was used to search for the best network for each bootstrap data set. This gives R=2000 best networks, and the confidence on the presence of an edge ranges from 0 (learned from 0 bootstrap samples) to 1 (learned from all bootstrap samples) [13]. To further increase the robustness of the final or consensus network, we defined the minimum significance threshold for arc strength as 0.700 if the calculated significance threshold was lower and accepted the calculated threshold otherwise. Regarding directionality, arcs with a direction coefficient below 0.666 after bootstrapping were considered undirected.

To determine the distributions of the variables and calculate the associated probabilities of the network, the adjacency matrix of the average bootstrapped directed acyclic graph was reproduced using the Bayesian network function, and belief propagation was carried out using the *gRain* package [13,19]. Belief propagation allows for inference tasks (probability queries) to be performed on the learned Bayesian networks, thereby providing a calculation of the distribution of values of a certain variable and the marginal and conditional probabilities of these values occurring based on the known value of an observed variable. Given a certain distribution, the marginal probability of a certain value occurring is calculated by integrating out all other variables, while the conditional probability is the probability of a value occurring for one variable, given a known, fixed value for at least one other variable [20]. These probability queries will allow for multiple relevant clinical scenarios to be recreated, based on the consensus network and the properties of the Markov blanket. When carrying out a query for *estimate*, if the values of its parent nodes are known, no other node can influence the conditional distribution of *estimate* [21]. If only some of its parent nodes are known, however, then some of the ancestors upstream of the undefined parent nodes can still influence the conditional probability of *estimate* [21]. To validate the structure learning process beyond the bootstrapping strategy used in learning a consensus network, two steps were taken. First, an ad hoc expert analysis was conducted to assess the plausibility and accuracy of the physiological relationships identified in the network. Second, 10-fold cross-validation was used to determine its predictive accuracy. Using the consensus network, the accuracy of the cross-validated predictions was determined by dichotomizing the estimates as described below and by calculating the area under the receiver operating curve, specificity, and sensitivity of the predictions made for patients, from which a validated cardiac index measurement was available.

Definitions and Bias

Patients underwent a protocolized and standardized clinical examination and subsequent CCUS, as described in the SICS-I protocol [15]. The main variable of interest was cardiac function estimation made by the student or physician after clinical examination was performed but before CCUS was performed. Examiners could score cardiac function as "poor," "moderate," "reasonable," or "good." For diagnostic test analyses and the validation step of the network structure, the "poor" and "moderate" estimates were grouped as "low," and the "reasonable" or "good" estimates were grouped as "high." Quality of the CCUS images and measurements of cardiac index were validated by core laboratory technicians (Groningen Image Core Lab, Groningen, The Netherlands) who were blinded for the rest of the measurements. Cardiac index measurements were categorized in two groups: "low" for cardiac index≤2.2 L/min/m^2 and "high" for cardiac index>2.2 L/min/m^2 [22]. All patients for whom a validated cardiac index measurement and estimate of cardiac function were available were included in the Bayesian network analysis. Patients for whom CCUS images were of insufficient quality or cardiac index measurements were

XSL•FO
RenderX

not available, were excluded from the diagnostic accuracy analysis.

Statistical Analysis

Due to the observational nature of the study, a formal sample size calculation was not possible. Statistical analyses were performed in STATA 15.0 (StataCorp, College Station, Texas) and R version 3.5.1 (R Core Team, Vienna, Austria). Data are presented as mean with SD when normally distributed, or as median with interquartile range in case of skewed data. Dichotomous and categorical data are presented in proportions. Sensitivity and specificity for both the network's and the examiners' estimated guess were calculated by cross-tabulation of the respective predictions and the validated cardiac index measurements. Additionally, positive predictive values (PPV) and negative predictive values (NPV) and positive likelihood ratios (LR+) and negative likelihood ratios (LR-) were calculated with 95% CIs for the examiners' estimates. For these, the overall accuracy was further expressed as a proportion of correctly classified cardiac index measurements (true negative and true positive measures) among all measures.

Results

Participants

A total of 1075 patients fulfilled our inclusion criteria, of which 1073 patients had available cardiac function estimates and were therefore included in the Bayesian network analysis. Of the included patients, 783 (73%) had validated cardiac index measurements and were included in the diagnostic accuracy tests. Further, 569 patients (73%) were included by students and 214 patients (27%) were included by physicians.

Descriptive Measures

Characteristics of included patients according to availability of cardiac index measurements are shown in Table 1. Body mass index and Simplified Acute Physiology Score (SAPS) II score were significantly different between patients (Table 1).

Bayesian Network Analysis

The structure learned for the network identified two clinical variables, namely, noradrenaline administration and the presence of delayed capillary refill time or mottling (dCRT-M), upon which the estimates of cardiac function are directly conditionally dependent (Table 2).

As denoted in Figure 1 by the dotted line, the arc from elevated lactate to oliguria had the lowest strength coefficient (0.728). The average directionality coefficient was 0.909, indicating well-defined directionality. Only one edge (between mechanical ventilation and high respiratory rate) did not meet the threshold for directionality and was thereby left undirected in the consensus directed acyclic graph (for querying, however, a direction from high respiratory rate to mechanical ventilation was defined based on expert knowledge to comply with the formal computational requirements) [15]. Additionally, there was no difference in network structure when including only students (n=801) or only physicians (n=271) compared to the network obtained with all the participants' estimates.

The probability queries conducted with the conditional probabilities for *estimate* are presented in a tree diagram in Figure 2. Each of the pathways in the diagram represents a scenario that could occur during clinical examination. Since one of the main focuses of SICS-I was the collection and interpretation of information available at bedside during physical examination, we expanded the conditional probability queries to also include respiratory rate and mechanical ventilation. Tachypnea virtually did not influence the probability of cardiac pump function being estimated as reasonable or good $P(E_{R,G})$, whereas ventilation status did ($P[E_{R,G}|$not ventilated, no tachypnea$]=P[E_{R,G}|$not ventilated, tachypnea$)=0.85$; $P[E_{R,G}|$ventilated, tachypnea$]=0.69$ and $P[E_{R,G}|$ventilated, no tachypnea$]=0.63$). When the patient received noradrenaline, $P(E_{R,G})$ was lower irrespective of whether they were mechanically ventilated ($P[E_{R,G}|$ventilation, noradrenaline$]=0.63$, $P[E_{R,G}|$ventilation, no noradrenaline$]=0.91$, $P[E_{R,G}|$no ventilation, noradrenaline$]=0.67$, $P[E_{R,G}|$no ventilation, no noradrenaline$]=0.93$). The same trend was found for dCRT-M, with reasonable or good estimates being more likely in the absence of dCRT-M.

Finally, an area under the receiver operating characteristic curve of 0.58 was obtained for the 10-fold cross-validated predictions of cardiac function made by the consensus network, with a specificity of 36% and a sensitivity of 79% [23].

XSL•FO
RenderX

Table 1. Patient characteristics.

Variable	No cardiac index measurement (n=292)	Cardiac index measurement (n=783)	Total (N=1075)	*P* value
Age (years), mean (SD)	62 (14)	62 (15)	62 (15)	.75
Male gender, n (%)	188 (64)	486 (62)	674 (63)	.49
Body mass index (kg/m^2), mean (SD)	27.5 (5.4)	26.7 (5.6)	26.9 (5.5)	.04
Arterial pressure (mm Hg), mean (SD)	78 (14)	79 (14)	78 (14)	.30
Heart rate (bpm[a]), mean (SD)	87 (22)	88 (21)	88 (21)	.35
Irregular heart rhythm, n (%)	28 (10)	88 (11)	116 (11)	.44
Central venous pressure (mm Hg), median (IQR)	9 (5, 12)	9 (5, 13)	9 (5, 13)	.74
Patients administered noradrenaline, n (%)	142 (49)	386 (49)	528 (49)	.85
Urine output (mL/kg/h), median (IQR)	0.6 (0.3, 1.2)	0.7 (0.4, 1.2)	0.6 (0.4, 1.2)	.22
Respiratory rate (bpm), mean (SD)	18 (5)	18 (6)	18 (6)	.50
Mechanical ventilation, n (%)	179 (61)	452 (58)	631 (59)	.29
Positive end-expiratory pressure (cm H_2O), median (IQR)	7 (5, 8)	7 (5, 8)	7 (5, 8)	.41
Central temperature (°C), mean (SD)	37.0 (0.9)	36.9 (0.9)	36.9 (0.9)	.84
Difference between central temperature and temperature of the dorsum of the foot (°C), mean (SD)	7.7 (3.2)	7.8 (3.2)	7.8 (3.2)	.66
Subjective “cold” temperature, n (%)	109 (37.6)	289 (37.1)	398 (37.2)	.88
Capillary refill time				
Knee (s), median (IQR)	3.0 (2.0, 4.5)	3.0 (2.0, 4.5)	3.0 (2.0, 4.5)	.48
Sternum (s), median (IQR)	2.8 (2.0, 3.0)	3.0 (2.0, 3.0)	3.0 (2.0, 3.0)	.84
Finger (s), median (IQR)	3.0 (2.0, 4.0)	2.5 (2.0, 4.0)	2.5 (2.0, 4.0)	.37
Mottling rate, mean (SD)				.64
None	157 (58.8)	397 (56.8)	554 (57.3)	
Mild	24 (9.0)	79 (11.3)	103 (10.7)	
Moderate	75 (28.1)	201 (28.8)	276 (28.6)	
Severe	11 (4.1)	22 (3.1)	33 (3.4)	
Hemoglobin (mmol/L), mean (SD)	6.8 (1.5)	6.8 (1.4)	6.8 (1.4)	.90
Lactate (mmol/L)	1.4 (0.9, 2.4)	1.4 (0.9, 2.2)	1.4 (0.9, 2.2)	.79
ICU[b] length of stay (days)	3.5 (1.9, 6.9)	3.1 (1.9, 6.5)	3.2 (1.9, 6.6)	.29
SAPS[c] II (points)	47 (37, 58)	44 (34, 56)	45 (35, 57)	.037
APACHE[d] IV score (points)	77 (56, 92)	73 (55, 91)	74 (56, 92)	.14
90-day mortality, n (%)	81 (27.7)	217 (27.7)	298 (27.7)	.99
Cardiac function estimate, n (%)				.004
Poor	8 (2.8)	18 (2.3)	26 (2.4)	
Moderate	46 (15.9)	165 (21.1)	211 (19.7)	
Reasonable	164 (56.6)	349 (44.6)	513 (47.8)	
Good	72 (24.8)	251 (32.1)	323 (30.1)	

[a]bpm: beats per minute.

[b]ICU: intensive care unit.

[c]SAPS: Simplified Acute Physiology Score.

[d]APACHE: Acute Physiology and Chronic Health Evaluation.

XSL•FO
RenderX

Table 2. Strength and direction coefficients of the consensus directed acyclic graph.

From	To	Strength	Direction
Age	Irregular rhythm	0.983	1.00
Mechanically ventilated	High respiratory rate	0.994	0.504
Mechanically ventilated	dCRT-M[a]	0.875	0.884
Irregular rhythm	Tachycardia	0.848	0.954
Tachycardia	High respiratory rate	0.999	0.931
Tachycardia	Low SBP[b]	0.821	0.883
Tachycardia	Elevated lactate	0.832	0.821
Low SBP	Low MAP[c]	1	1
Low DBP[d]	Low MAP	1	1
Elevated lactate level	Oliguria	0.728	0.803
Elevated lactate level	Noradrenaline administration	1	1
Noradrenaline administration	Mechanically ventilated	1	0.957
Noradrenaline administration	Estimate	0.999	1
dCRT-M	Estimate	0.876	1

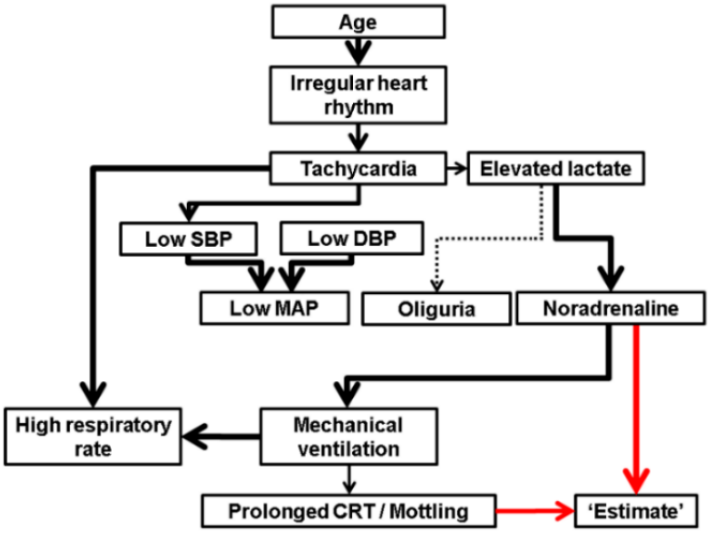
[a]dCRT-M: delayed capillary refill time or mottling.

[b]SBP: systolic blood pressure.

[c]MAP: mean arterial pressure.

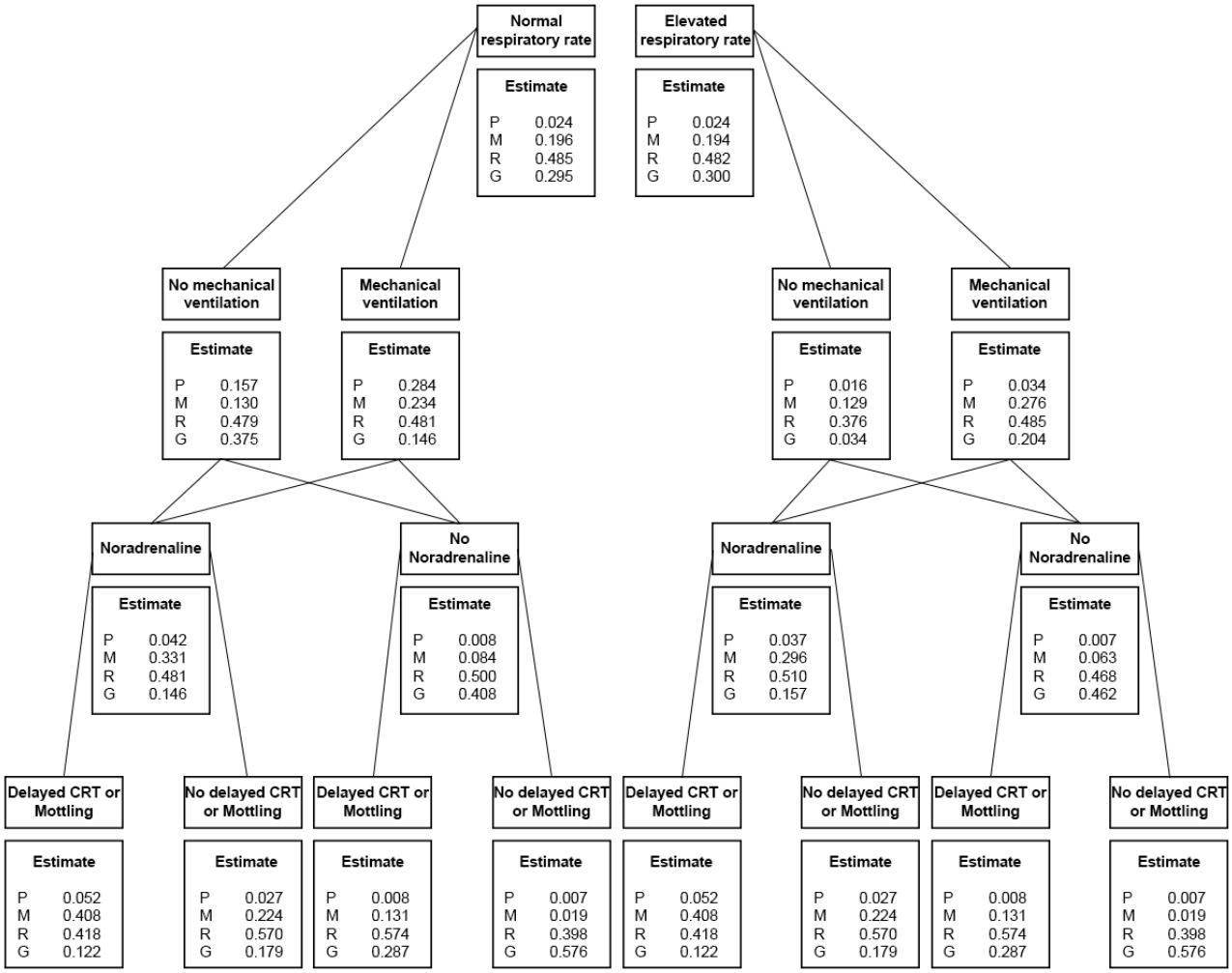
[d]DBP: diastolic blood pressure.

Figure 1. Consensus directed acyclic graph. Red lines represent direct conditional dependencies to estimate. Black lines represent direct conditional dependencies to other variables. Width of the line represents strength coefficient. The dotted line represents the weakest strength coefficient. DBP: diastolic blood pressure; SBP: systolic blood pressure; MAP: mean arterial pressure; CRT: capillary refill time.



XSL•FO
RenderX

Figure 2. Tree diagram showing the conditional probabilities queries for estimate associated with multiple scenarios during clinical examination. At each step, only the variables above the split are known and as more information becomes available, the conditional probabilities change. P=Poor; M=Moderate; R=Reasonable; G=Good; CRT: capillary refill time.



Diagnostic Accuracy

Diagnostic accuracy tests for estimating of a low cardiac index showed a sensitivity of 26% and 39%, a specificity of 83% and 74%, PPV of 45% and 48%, NPV of 67% and 66%, LR+ of 1.52 and 1.52, and LR- of 0.89 and 0.82 for students and physicians, respectively. The overall accuracy of cardiac index estimates was 63% and 61% for students and physicians, respectively. For all patients combined, sensitivity was 30%, specificity was 80%, PPV was 46%, NPV was 67%, LR+ was 1.53, LR- was 0.87, and the overall accuracy of diagnostic tests was 62% (Table 3).

Table 3. Accuracy, sensitivity, specificity, predictive values, and likelihood ratios for students' and physicians' estimates.

Variable	Students (n=569)	Physicians (n=214)	Overall (N=783)
Sensitivity, % (95% CI)	26 (20-33)	39 (28-50)	30 (25-36)
Specificity, % (95% CI)	83 (78-86)	74 (66-82)	80 (77-84)
Positive predictive value, % (95% CI)	45 (38-53)	48 (39-58)	46 (40-53)
Negative predictive value, % (95% CI)	67 (65-69)	66 (61-71)	67 (65-69)
Positive likelihood ratio, 95% CI	1.52 (1.10-2.09)	1.52 (1.02-2.25)	1.53 (1.19-1.97)
Negative likelihood ratio, 95% CI	0.89 (0.81-0.98)	0.82 (0.67-1.00)	0.87 (0.80-0.95)
Overall accuracy, % (95% CI)	63 (59-67)	61 (54-67)	62 (59-66)

XSL•FO
RenderX

Discussion

Principal Findings

Clinical examination is used daily by physicians as an easy, cheap, and noninvasive way of gathering information to guide interventions and further diagnostic testing. Clinical signs such as oliguria; altered consciousness; and cold, clammy skin are known possible indicators of organ hypoperfusion and are used to diagnose shock in critically ill patients [2]. However, the value of clinical examination has been questioned, and previous studies have shown physicians to perform poorly in diagnosing a low cardiac index based on physical signs alone [8,9]. In this study, we confirmed that the accuracy of these estimates remains low for both students and physicians. Surprisingly, we identified noradrenaline administration and delayed CRT or mottling as seemingly the major factors influencing cardiac function estimates using Bayesian network analysis. These findings may serve as the basis for improving the value of clinical examination (1) by identifying some of the biases clinicians may be subjected to, which causes them to overdiagnose compared to students, and (2) by clarifying some of the thought process behind the clinical examination. This allows the examiner to "think about how they think" when performing clinical examination and can help clinicians be trained to prioritize or leave out certain variables when making their assessment.

Bayesian Network Analysis

Validation and Limitations

Validation of the network structure was a crucial yet challenging step toward our goal of trying to obtain a plausible representation of the examiners' knowledge network and thought process at bedside. We believe to have tackled this challenge in the best way possible by validating it in three different ways: using the bootstrapping process to generate a consensus network; conducting expert validation of the plausibility of the arcs; and using the network as a predictor, as previously suggested [13]. We believe that the similarity in accuracy, sensitivity, and specificity between the network's predictions and the examiners' own estimates is further proof of the validity of its structure. It must be restated that the goal of this study was not to build and optimize a predictive model, in which case the predictive accuracy, sensitivity, and specificity we obtained would be subpar. In fact, had the network been able to make the estimates with a substantially higher accuracy than the examiners' estimates, we would be more reluctant to affirm that is parallel with the examiner's thought process.

As any exploratory study, however, we faced several limitations. The first was practical, as not all included patients had cardiac index measurements, since CCUS is not applicable for every ICU patient and views obtained by CCUS can be obstructed due to lines, wounds, or excess adiposity [24]. This prevented us from using the complete cohort and likely accounted for the difference in SAPS-II score and body mass index in the patients with and without CCUS measurements. Second, the discretization required by the parametric assumptions of Bayesian network algorithms comes with the inherent risk of useful information being discarded in the process, which does not guarantee that the dependence relationships involving the original variables are preserved. Last, for causality to be derived from Bayesian networks, there must be no unobserved variables influencing the variables included in the network that may act as confounding factors. In SICS-I, the focus was on examining and improving students' and physicians' educated guess, resorting primarily to bedside information, such as vasopressor and fluid perfusors, vital signs, and physical examination. Therefore, to best replicate this scenario, we opted to include in the network only variables that are readily available during the protocolized examination. Although this increases the risk of introducing bias in the causal network, the accuracy of the physiological dependencies identified gives us reason to believe that no substantial bias is present.

Do Probability Queries Help Explain the Modest Diagnostic Accuracy?

Previous studies on the diagnostic accuracy of clinical examination have found the performance of experienced physicians and students to be comparable [6]. Expert physicians are more often affected by multiple cognitive biases, such as confirmatory bias and premature closure, compared to students, who remain more open to new hypotheses and persist in collecting data [25,26]. Interestingly, while the diagnostic accuracy for individual physicians can be as low as 62.5%, there is a visible increase as the number of physicians involved increases (up to 85.6% for groups of nine physicians) [27]. Our results are in line with the literature, and we additionally showed that physicians had a higher sensitivity but lower specificity than students (39% and 26%, and 74% and 83%, respectively). These differences in sensitivity and specificity represent a tendency of physicians to overdiagnose, which has previously been related to confirmatory bias and premature closure. Indeed, two other findings support the idea already given by the direct dependence of *estimate* solely on noradrenaline and dCRT-M that premature closure was a common phenomenon. First, in the probability queries, while machine ventilation does not directly influence the *estimate*, considerable changes in the probability of the *estimate* are still observable, depending on whether the patient is ventilated, before noradrenaline use and dCRT-M are known. This could be due to the fact that mechanical ventilation is almost inevitably the first variable to be noted when the examiner approaches bedside. Second, a comparison of the change in the probabilities of *estimate* based on varying clinical evidence with the likelihood ratios calculated in another SICS-I substudy shows that variables further upstream of estimate such as respiratory should be taken more into account [15]. For example, while the positive and negative likelihood ratios of a high respiratory rate are as suggestive as those of a delayed CRT, the query shows that the probability of being estimated to have low cardiac function was considerably lower in those without dCRT-M (0.25) than in those with dCRT-M (0.46) and the probability of a patient with tachypnea being estimated to have low or high cardiac function was virtually the same. This is despite tachypnea having a positive and negative likelihood ratio of 1.16 and 0.68, respectively.

XSL•FO
RenderX

Conclusion and Future Implications

This study confirms that the accuracy of cardiac function estimates remains low for both students and physicians, and it identifies noradrenaline administration and delayed CRT or mottling as seemingly the major factors influencing these estimates. Although it will remain challenging to try to replicate the thought process of the examiner, not only methodologically, but also because different individuals have different levels of knowledge and different examination routines, Bayesian networks seem like a promising tool to help break down and better understand the educated guessing process. The insight gained in studies such as this one, can help teach students think about how they think and, on a clinical level, provide much-needed guidance for prioritization of variables during clinical examination. In fact, our team is currently compiling the knowledge acquired in the SICS-I substudies to build an interactive game for medical students, residents, and specialists. This electronic learning tool will ask the player to estimate cardiac function using the same scale and data from variables such as bedside monitor hemodynamic variables, ventilator and pump settings, and urine output.

Authors' Contributions

TK and JCF performed the data analysis and drafted the manuscript, and all other authors reviewed and provided feedback with each draft. All authors approved of the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Variables included in the Bayesian network and the respective Cramér V similarity measure.
[PDF File (Adobe PDF File), 158 KB - medinform_v7i4e15358_app1.pdf]

References

1. Sevransky J. Clinical assessment of hemodynamically unstable patients. Curr Opin Crit Care 2009 Jun;15(3):234-238 [FREE Full text] [doi: 10.1097/MCC.0b013e32832b70e5] [Medline: 19387339]
2. Cecconi M, De Backer D, Antonelli M, Beale R, Bakker J, Hofer C, et al. Consensus on circulatory shock and hemodynamic monitoring. Task force of the European Society of Intensive Care Medicine. Intensive Care Med 2014 Dec;40(12):1795-1815 [FREE Full text] [doi: 10.1007/s00134-014-3525-z] [Medline: 25392034]
3. Vincent J, De Backer D. Circulatory shock. N Engl J Med 2013 Oct 31;369(18):1726-1734. [doi: 10.1056/NEJMra1208943] [Medline: 24171518]
4. Sakr Y, Reinhart K, Vincent J, Sprung CL, Moreno R, Ranieri VM, et al. Does dopamine administration in shock influence outcome? Results of the Sepsis Occurrence in Acutely Ill Patients (SOAP) Study. Crit Care Med 2006 Mar;34(3):589-597. [doi: 10.1097/01.CCM.0000201896.45809.E3] [Medline: 16505643]
5. Saugel B, Malbrain MLNG, Perel A. Hemodynamic monitoring in the era of evidence-based medicine. Crit Care 2016 Dec 20;20(1):401 [FREE Full text] [doi: 10.1186/s13054-016-1534-8] [Medline: 27993153]
6. Hiemstra B, Eck RJ, Keus F, van der Horst ICC. Clinical examination for diagnosing circulatory shock. Curr Opin Crit Care 2017 Aug;23(4):293-301 [FREE Full text] [doi: 10.1097/MCC.0000000000000420] [Medline: 28570301]
7. Perel A, Saugel B, Teboul J, Malbrain MLNG, Belda FJ, Fernández-Mondéjar E, et al. The effects of advanced monitoring on hemodynamic management in critically ill patients: a pre and post questionnaire study. J Clin Monit Comput 2016 Oct;30(5):511-518. [doi: 10.1007/s10877-015-9811-7] [Medline: 26661527]
8. Suess EM, Pinsky MR. Hemodynamic Monitoring for the Evaluation and Treatment of Shock: What Is the Current State of the Art? Semin Respir Crit Care Med 2015 Dec;36(6):890-898. [doi: 10.1055/s-0035-1564874] [Medline: 26595049]
9. Elder A, Japp A, Verghese A. How valuable is physical examination of the cardiovascular system? BMJ 2016 Jul 27;354:i3309-i3329. [doi: 10.1136/bmj.i3309] [Medline: 27598000]
10. Soto-Ferrari M, Prieto D, Munene G. A Bayesian network and heuristic approach for systematic characterization of radiotherapy receipt after breast-conservation surgery. BMC Med Inform Decis Mak 2017 Jun 28;17(1):93 [FREE Full text] [doi: 10.1186/s12911-017-0479-4] [Medline: 28659177]
11. Stivaros SM, Gledson A, Nenadic G, Zeng X, Keane J, Jackson A. Decision support systems for clinical radiological practice -- towards the next generation. Br J Radiol 2010 Nov;83(995):904-914 [FREE Full text] [doi: 10.1259/bjr/33620087] [Medline: 20965900]
12. Peelen L, de Keizer NF, Jonge ED, Bosman R, Abu-Hanna A, Peek N. Using hierarchical dynamic Bayesian networks to investigate dynamics of organ failure in patients in the Intensive Care Unit. J Biomed Inform 2010 Apr;43(2):273-286 [FREE Full text] [doi: 10.1016/j.jbi.2009.10.002] [Medline: 19874913]
13. Scutari M, Auconi P, Caldarelli G, Franchi L. Bayesian Networks Analysis of Malocclusion Data. Sci Rep 2017 Nov 10;7(1):15236 [FREE Full text] [doi: 10.1038/s41598-017-15293-w] [Medline: 29127377]

XSL•FO
RenderX

14. Mukherjee S, Speed TP. Network inference using informative priors. Proc Natl Acad Sci U S A 2008 Sep 23;105(38):14313-14318 [FREE Full text] [doi: 10.1073/pnas.0802272105] [Medline: 18799736]
15. Hiemstra B, Koster G, Wiersema R, Hummel YM, van der Harst P, Snieder H, SICS Study Group. The diagnostic accuracy of clinical examination for estimating cardiac index in critically ill patients: the Simple Intensive Care Studies-I. Intensive Care Med 2019 Feb;45(2):190-200. [doi: 10.1007/s00134-019-05527-y] [Medline: 30706120]
16. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015 Oct 28;351:h5527 [FREE Full text] [doi: 10.1136/bmj.h5527] [Medline: 26511519]
17. Scutari M. Learning Bayesian Networks with the bnlearn R Package. J. Stat. Soft 2010;35(3) [FREE Full text] [doi: 10.18637/jss.v035.i03]
18. Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn 2006 Mar 28;65(1):31-78. [doi: 10.1007/s10994-006-6889-7]
19. Højsgaard S. Graphical Independence Networks with the gRain Package for R. J. Stat. Soft 2012 Feb 28;46(10). [doi: 10.18637/jss.v046.i10]
20. Annis C. Statistical Engineering. Joint, Marginal, and Conditional Distributions URL: http://www.statisticalengineering.com/joint_marginal_conditional.htm [accessed 2018-12-02]
21. Barber D. Bayesian Reasoning And Machine Learning. Cambridge: Cambridge University Press; 2012.
22. Forrester JS, Diamond GA, Swan HJ. Correlative classification of clinical and hemodynamic function after acute myocardial infarction. The American Journal of Cardiology 1977 Feb;39(2):137-145. [doi: 10.1016/s0002-9149(77)80182-3] [Medline: 835473]
23. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011 Mar 17;12:77 [FREE Full text] [doi: 10.1186/1471-2105-12-77] [Medline: 21414208]
24. Koster G, van der Horst ICC. Critical care ultrasonography in circulatory shock. Curr Opin Crit Care 2017 Aug;23(4):326-333. [doi: 10.1097/MCC.0000000000000428] [Medline: 28590257]
25. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak 2016 Nov 03;16(1):138 [FREE Full text] [doi: 10.1186/s12911-016-0377-1] [Medline: 27809908]
26. Krupat E, Wormwood J, Schwartzstein RM, Richards JB. Avoiding premature closure and reaching diagnostic accuracy: some key predictive factors. Med Educ 2017 Nov;51(11):1127-1137. [doi: 10.1111/medu.13382] [Medline: 28857266]
27. Barnett ML, Boddupalli D, Nundy S, Bates DW. Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple Physicians vs Individual Physicians. JAMA Netw Open 2019 Mar 01;2(3):e190096 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.0096] [Medline: 30821822]

Abbreviations

APACHE: Acute Physiology and Chronic Health Evaluation
CCUS: critical care ultrasonography
DBP: diastolic blood pressure
dCRT-M: delayed capillary refill time or mottling
ICU: intensive care unit
LR-: negative likelihood ratios
LR+: positive likelihood ratios
MAP: mean arterial pressure
NPV: negative predictive values
PPV: positive predictive values
SAPS: Simplified Acute Physiology Score
SBP: systolic blood pressure
SICS-I: Simple Intensive Care Studies-I

XSL•FO
RenderX

Edited by G Eysenbach; submitted 04.07.19; peer-reviewed by P Bergl, T Aslanidis; comments to author 11.09.19; revised version received 17.09.19; accepted 23.09.19; published 30.10.19.

Please cite as:
Kaufmann T, Castela Forte J, Hiemstra B, Wiering MA, Grzegorczyk M, Epema AH, van der Horst ICC, SICS Study Group
A Bayesian Network Analysis of the Diagnostic Process and its Accuracy to Determine How Clinicians Estimate Cardiac Function in Critically Ill Patients: Prospective Observational Cohort Study
JMIR Med Inform 2019;7(4):e15358
URL: http://medinform.jmir.org/2019/4/e15358/
doi:10.2196/15358
PMID:31670697

©Thomas Kaufmann, José Castela Forte, Bart Hiemstra, Marco A Wiering, Marco Grzegorczyk, Anne H Epema, Iwan C C van der Horst, SICS Study Group. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 30.10.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Differential Diagnosis Assessment in Ambulatory Care With an Automated Medical History–Taking Device: Pilot Randomized Controlled Trial

Adrien Jean-Pierre Schwitzguebel[1*], MD; Clarisse Jeckelmann[2*], MA; Roberto Gavinio[3], PhD; Cécile Levallois[3], MD; Charles Benaïm[1*], MD; Hervé Spechbach[3*], MD

[1]Division of Physical Medicine and Rehabilitation, Department of Rheumatology, Lausanne University Hospital, Lausanne, Switzerland

[2]Faculty of Medicine, University of Lausanne, Lausanne, Switzerland

[3]Ambulatory Emergency Care Unit, Department of Primary Care Medicine, Geneva University Hospitals, Geneva, Switzerland

[*]these authors contributed equally

Corresponding Author:
Adrien Jean-Pierre Schwitzguebel, MD
Division of Physical Medicine and Rehabilitation
Department of Rheumatology
Lausanne University Hospital
Avenue Pierre Decker 4
Lausanne
Switzerland
Phone: 41 797620562
Email: adrien.schwitzguebel@gmail.com

Abstract

Background: Automated medical history–taking devices (AMHTDs) are emerging tools with the potential to increase the quality of medical consultations by providing physicians with an exhaustive, high-quality, standardized anamnesis and differential diagnosis.

Objective: This study aimed to assess the effectiveness of an AMHTD to obtain an accurate differential diagnosis in an outpatient service.

Methods: We conducted a pilot randomized controlled trial involving 59 patients presenting to an emergency outpatient unit and suffering from various conditions affecting the limbs, the back, and the chest wall. Resident physicians were randomized into 2 groups, one assisted by the AMHTD and one without access to the device. For each patient, physicians were asked to establish an exhaustive differential diagnosis based on the anamnesis and clinical examination. In the intervention group, residents read the AMHTD report before performing the anamnesis. In both the groups, a senior physician had to establish a differential diagnosis, considered as the gold standard, independent of the resident's opinion and AMHTD report.

Results: A total of 29 patients were included in the intervention group and 30 in the control group. Differential diagnosis accuracy was higher in the intervention group (mean 75%, SD 26%) than in the control group (mean 59%, SD 31%; P=.01). Subgroup analysis showed a between-group difference of 3% (83% [17/21]-80% [14/17]) for low complexity cases (1-2 differential diagnoses possible) in favor of the AMHTD (P=.76), 31% (87% [13/15]-56% [18/33]) for intermediate complexity (3 differential diagnoses; P=.02), and 24% (63% [34/54]-39% [14/35]) for high complexity (4-5 differential diagnoses; P=.08). Physicians in the intervention group (mean 4.3, SD 2) had more years of clinical practice compared with the control group (mean 5.5, SD 2; P=.03). Differential diagnosis accuracy was negatively correlated to case complexity (r=0.41; P=.001) and the residents' years of practice (r=0.04; P=.72). The AMHTD was able to determine 73% (SD 30%) of correct differential diagnoses. Patient satisfaction was good (4.3/5), and 26 of 29 patients (90%) considered that they were able to accurately describe their symptomatology. In 8 of 29 cases (28%), residents considered that the AMHTD helped to establish the differential diagnosis.

Conclusions: The AMHTD allowed physicians to make more accurate differential diagnoses, particularly in complex cases. This could be explained not only by the ability of the AMHTD to make the right diagnoses, but also by the exhaustive anamnesis provided.

(JMIR Med Inform 2019;7(4):e14044) doi:10.2196/14044

XSL•FO
RenderX

KEYWORDS

differential diagnosis; decision making; computer-assisted; hospital outpatient clinics; general practitioners; clinical applications software; patient engagement

Introduction

Background

In studies performed in the United States on medical errors in primary care medicine, diagnostic errors are the most common [1-3] and the most expensive [4,5], as well as the cause of most malpractice claims [1,4,6]. A prevalence of diagnostic errors in outpatient care of at least 5% has been reported [7]. Despite their importance, diagnostic errors are underemphasized and underidentified [6,8], and the development of novel strategies to improve the accuracy of the initial diagnosis should be a priority.

Interactive computerized interviews completed by patients have several advantages and are shown to be as accurate as classic clinician records. Notably, they permit a significant difference in time taken during the consultation [9], thus demonstrating that the initial triage could be performed in less time [10]. Physicians also receive more data than that from conventional history taking [11-15]. In addition, false positive answers to classic interviews may less likely occur as answers could be optional, thus allowing blank responses [16]. In the waiting room, patients have reported high satisfaction by helping their physician through the completion of interactive computerized interviews [17,18]. The interview is better organized and permits the physician to easily consolidate the anamnesis with supplementary questions, depending on the data provided [16]. Patients are also more likely to reveal sensitive data to a computer than to a physician [19-21]. Finally, the process is an effective strategy to empower patients to be active in their own care (patient engagement) [22,23].

At present, 2 types of interactive computerized interviews exist to facilitate the anamnesis and diagnosis before the consultation, that is, symptom checkers and automated medical history–taking devices (AMHTDs). Recently, 23 symptom checkers were evaluated with standardized vignettes. The correct diagnosis was made in 58% of the cases, and a correct triage was performed in 80% [24], which can be considered as insufficient. Another solution includes an AMHTD based on a single symptom or localization [17]. This type of system can be useful and accurate, provided that the clinical presentation is typical, for example, a patient presenting with calf pain after strenuous exercise and a potential sciatica.

Objectives

The primary aim of this pilot study was to investigate whether the DIAANA AMHTD allowed physicians to establish a more accurate DD, with the DD of a senior physician considered as the gold standard. Secondary aims were to assess the accuracy of the DD list established by the AMHTD, identify factors that might influence the usefulness of the AMHTD, and evaluate physician and patient satisfaction with its use.

We tested a novel AMHTD, named *DIAANA* (DIAgnosis & ANAmnesis; created by Logic-based Medicine Sàrl), to help the physician to establish the differential diagnosis (DD) more accurately, based on broad possibilities of disease or trauma localization, triggering factors, and symptoms. The physician can therefore begin his consultation with an exhaustive anamnesis summary including a more precise localization and nature of symptoms as well as a high-sensitivity DD list with corresponding triggering factors for each diagnosis. We consider that this tool could help the physician in his/her diagnostic reasoning and to perform tasks more efficiently, without being substituted by the AMHTD.

Methods

Study Design

We conducted a pilot, single-center, unblinded, 1:1 parallel-group, randomized efficacy trial. No follow-up was necessary. There were no changes in the protocol after trial commencement. The study protocol was optimized and approved by an independent expert methodologist. It was not registered as it was considered to be a pilot phase. Given that recruitment began just after the approval, it would therefore have not been relevant to register the study after the beginning of the recruitment. The protocol was approved by the Medical Ethics Committee of Geneva University Hospitals (Geneva, Switzerland; REQ-2017-00878). No bugs were fixed during the trial. As this was a purely observational study without identifiable side effects or negative consequences for patients, only oral informed consent was obtained, supported by a brief written description of the project. Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and online TeleHealth V 1.6 (see Multimedia Appendix 1) was used to improve and standardize the quality of this paper [25].

Patient Population

From May to September 2018, we prospectively enrolled adult patients presenting to the emergency outpatient unit of our institution and suffering from symptoms covered by the AMHTD. Symptoms were localized to the superior member (apart from the hand, as the device had not yet been programmed to take related conditions into consideration), the trunk, and the inferior member, with the exception of strictly dermatologic concerns and toes and inversion ankle trauma as the diagnosis is generally obvious. We excluded patients with a medical situation considered as urgent and unable to complete the digitalized AMHTD (sight problems, advanced age, and non-French-speaking). Patients were enrolled only when one of the senior physicians in charge of the project (CL, TW, RG, MB, and HS) and one of the coordinators (CJ and BV) were available.

Randomization and Recruitment

At the beginning of the study, 18 residents of the emergency outpatient unit were stratified, and 1:1 matched by their years of clinical experience (orthopedics, rheumatology, and physical medicine counted twice) and then randomized. When a patient

XSL•FO
RenderX

was allocated to a resident physician using the emergency software system, the coordinating researcher evaluated the patient's potential eligibility. The senior physician then confirmed the patient's eligibility and applied the exclusion criteria. Depending on the resident physician's allocation, the patient was included in either the intervention or the control group. In each group, the recruitment was blocked after the inclusion of 30 patients.

DIAANA Tool Presentation

The DIAANA AMHTD functions as follows: On the basis of an interactive questionnaire completed by the patient before the consultation, which includes 269 questions (mainly multiple choice), it performs an exhaustive anamnesis focused on the problem and proposes a panel of DDs with a high sensitivity, selected on a panel of 126 diagnostic entities. The artificial reasoning system of DIAANA mimics how a specialist physician would reason to establish a DD. The information transmitted is in an easy-to-use form for the physician that includes a summary of the anamnesis centered on relevant elements from the questionnaire and a list of possible diagnoses with their emergency level, potential contributing factors, and first-line management proposals. Multimedia Appendix 2 illustrates an example of a patient suffering from deep vein thrombosis that was initially confounded with a tennis leg. More detailed information is available on the AMHTD's website [26].

DIAANA Tool Development

For 3 years, AS was involved in the development of the AMHTD, taking into consideration all aspects of the diagnosis and management of orthopedic, rheumatologic, vascular, neuropathic, and sports-related medical conditions, with the help of a few sources [27-29] as well as peer advice.

The system was built with triggering conditions that are turned on when the patient selects a specific answer. The triggering condition will then call up new questions and diagnostic entities. As an example, if the patient clicks *leg* on the general localization, the trigger *leg* is turned on, and a more specific question about the leg localization appears (see Multimedia Appendix 2). AS built a first draft of DIANNA including the principal questions of a proper musculoskeletal anamnesis. Then, he considered the 126 selected diagnosis entities in more depth and added more specific questions for each diagnosis step by step. The accuracy of DIANNA depends, therefore, on the accuracy of the patient's answer as well as the exhaustivity of the questions and diagnostic entities. As an example, if the correct localization (eg, *ankle*) is not selected, specific questions (eg, trauma in external rotation) and a specific diagnosis (eg, syndesmosis sprain) will not be triggered and thus be missing in the DIANNA summary.

Hundreds of episodes of testing with healthy volunteers, medical students, and patients were performed during the development process, and the formulation of questions, triggering conditions, and the DIANNA summary were adjusted according to feedback from users. A final development phase was conducted with the feedback of 20 patients presenting to the emergency outpatient unit, and the first version of the digital content of the tool was then frozen for the pilot study. This frozen version remains available upon request to the corresponding author.

Intervention

In the intervention group, patients in the AMHTD group were asked to complete a digital form on a touch pad by the coordinator (and without help) before the medical consultation. The AMHTD summary was then printed and given to the resident physician before the consultation. At the end of the consultation, but before consulting the complementary medical examination results (radiographs and blood laboratory results), the resident physician established his/her DD on the diagnosis list (see Multimedia Appendix 3**)** on a touch pad, without the help of the research coordinator**.** In parallel, the senior physician established the gold standard DD on the same list. In the control group, the resident physician established his DD on the diagnosis list at the end of the consultation, but before consulting the complementary medical examinations. The senior physician followed the same procedure. For ethical reasons, the use of the AMHTD had no influence on patient care as the clinical management was fully decided upon by the senior physician who had no access to the summary generated.

Outcomes

The primary outcome was the percentage of correct DDs established by the resident physician compared with the senior physician. Secondary outcomes included (1) the percentage of correct AMHTD DDs and the percentage of correct AMHTD DDs followed/not followed by the resident, as well as the percentage of incorrect AMHTD DDs followed by the resident and the number of incorrect AMHTD DDs; (2) overall patient satisfaction on the understandability of AMHTD questions (1-5 Likert scale), ability to describe symptoms accurately (percentage), and respect of the patient's wish to use the AMHTD at home and to keep the generated summary (percentages); (3) resident's feedback on the wish to obtain the integrality of the AMHTD summary (percentage), whether the AMHTD found DDs that would have been omitted otherwise (percentage), and if the use of the device saved time (1-5 Likert scale); and (4) the percentage of correct DDs depending on case complexity, defined as the number of DDs present in the gold standard DD (1-2 DDs=low complexity; 3 DDs=intermediate complexity; and 4-5 DDs=high complexity). The stratification for the case complexity definition used has never been published. The rationale was to highlight that the AMHTD was built and conceived to help the physician when the diagnosis might be confusing or in the case of a complex situation. Indeed, it would not be relevant to ask the patient to provide a complete anamnesis if the physician can complete it in 2 min for a problem such as benign soft tissue trauma.

Statistical Analyses

A sample size of 30 patients per group was chosen as recommended for pilot studies to achieve an appropriate level of statistical power [30]. It corresponds to the detection of a potential difference of 21% between groups for a power of 80% and an alpha significance level of 5%. Descriptive statistics were used to describe baseline characteristics. Differences between groups in the intention-to-treat analysis were evaluated

XSL•FO
RenderX

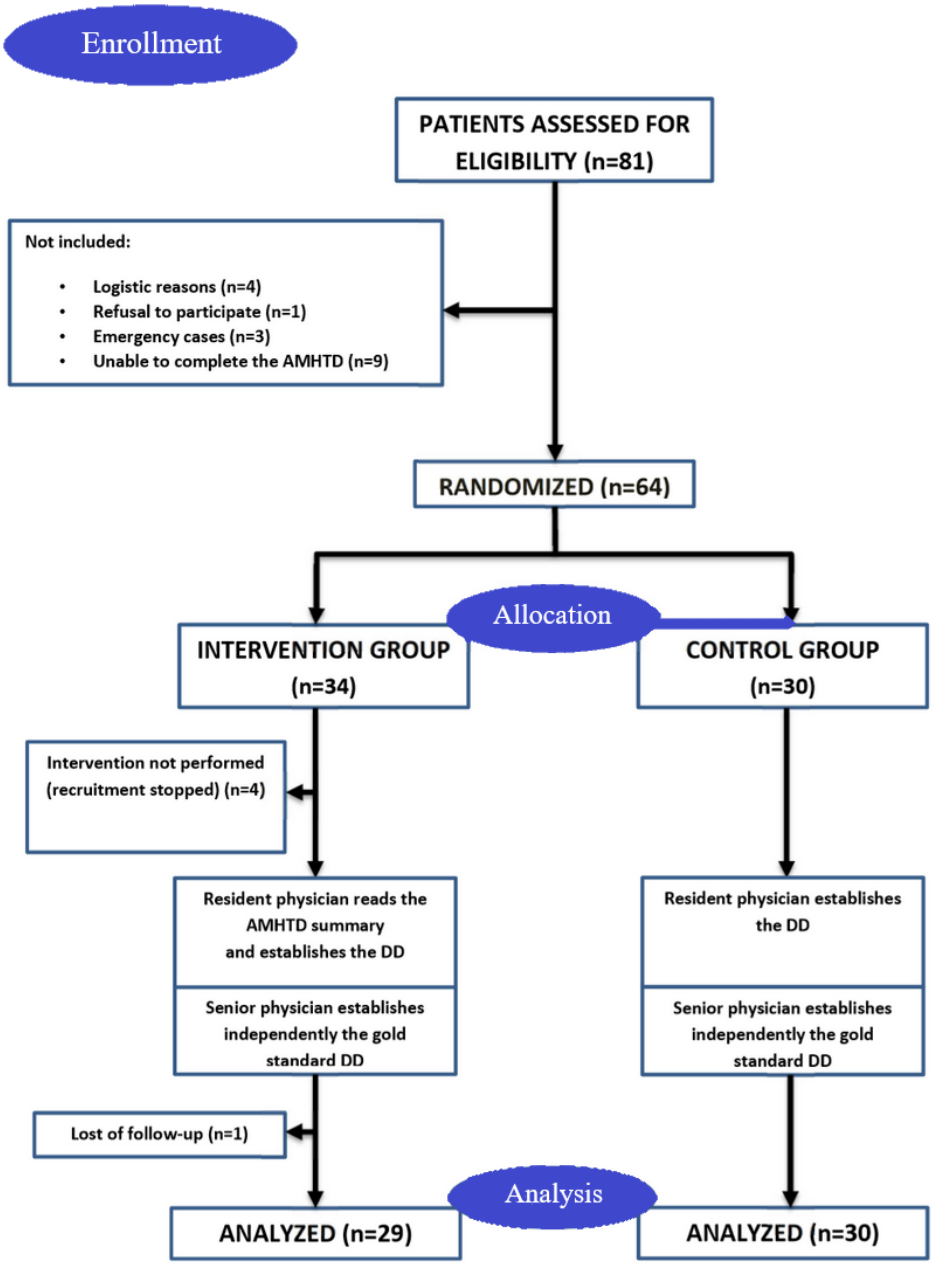
using Student *t* test or the Wilcoxon rank-sum test, when appropriate. Analysis of covariance was performed considering the covariables of interest (primary outcome, case complexity, and resident's years of experience) with a *P* value <.20 considered as significant in univariate analysis. *P* values <.05 were considered as statistically significant. All analyses were performed using R v3.4.2 Portable (Free Software Foundation Inc).

Results

Population

Of the 81 patients screened, 64 were randomized and allocated to residents (Figure 1). Among the randomized patients, 4 allocated to the intervention group were not included as 30 patients were already included in the intervention group; 1 patient was lost to follow-up. In the final analysis, 29 patients were included in the intervention group and 30 in the control group. Preintervention patient demographics, case complexity, and initial complaint/s did not differ between the groups (Table 1). Residents in the control group had more years of practice (*P*=.03).

Figure 1. Study flow chart. AMHTD: automated medical history–taking device; DD: differential diagnosis.



XSL•FO
RenderX

Table 1. Baseline characteristics.

Baseline characteristics	AMHTD[a] (n=29)	Control group (n=30)	*P* value
Age (years)			
Mean (SD)	38 (14)	42.1 (16)	.29
Range	17-66	19-75	.29
Male gender, n (%)	23 (79)	22 (73)	.82
Physician's practice (years)			
Mean (SD)	4.3 (2)	5.5 (2)	.03
Range	3-8	3-8	.03
Case complexity (number of differential diagnoses to find)			
Mean (SD)	3.1 (1)	2.9 (1)	.60
Range	1-5	1-5	.60
Initial complaint, n (%)			
Elbow pain	1 (3)	1 (3)	>.99
Shoulder pain and trauma	3 (10)	2 (7)	.97
Back pain and trauma	5 (17)	7 (23)	.80
Pelvic pain	2 (7)	0 (0)	.46
Knee pain and trauma	8 (28)	6 (20)	.70
Ankle trauma	4 (14)	6 (20)	.77
Foot trauma	2 (7)	2 (7)	>.99
Soft tissue trauma and swelling	4 (14)	5 (17)	>.99

[a]AMHTD: automated medical history–taking device.

Analysis of Accuracy of Differential Diagnosis

In the univariate analysis, the percentage of correct DDs was (1) higher in the intervention group (mean 75% [SD 26%] vs mean 59% [SD 31%], respectively; *P*=.03); (2) negatively correlated to case complexity (*r*=0.41; *P*=.001); and (3) negatively correlated to residents' years of practice (*r*=0.04; *P*=.72). The *P* value of the analysis of covariance model, including the percentage of DDs found and case complexity was .01. Considering case complexity, we observed between-group differences in favor of the AMHTD of 3% (83% [17/21]-80% [14/17]) for low-complexity cases, 31% (87% [13/15]-56% [18/33]) for intermediate-complexity cases, and 24% (63% [34/54]-39% [14/35]) for high-complexity cases (Table 2). The type of DD made by the senior physician, depending on the case complexity, is presented in the Multimedia Appendix 4.

By comparison, the AMHTD was able to find 73% (SD 30%) of correct DDs for the whole cohort: 91% (SD 20%) for low-complexity cases; 67% (SD 24%) for moderate-complexity cases; and 58% (SD 32%) for high-complexity cases (see Multimedia Appendix 5). The AMHTD also proposed 5(SD 4) incorrect diagnostic proposals. Residents did not list 10% (SD 19%) of the correct DDs proposed by the AMHTD and listed 21% (SD 51%) of incorrect DDs.

Table 2. Percentage of correct differential diagnoses per group.

DD[a] studied	AMHTD[b] (n=29)		Control group (n=30)		Univariate analysis *P* value	Multivariate analysis *P* value
	Mean (SD)	Range	Mean (SD)	Range		
DD accuracy	75 (26)	25-100	59 (31)	0-100	.03	<.001
Low complexity (1-2 DDs to find)	83 (25)	50-100	80 (26)	50-100	.76	—[c]
Moderate complexity (3 DDs to find)	87 (18)	67-100	56 (26)	0-100	.02	—
High complexity (4-5 DDs to find)	63 (25)	25-100	39 (29)	0-80	.08	—

[a]DD: differential diagnosis.

[b]AMHTD: automated medical history–taking device.

[c]Not applicable.

XSL•FO
RenderX

Users Satisfaction

Patient satisfaction was good regarding overall satisfaction with questions and their understandability, and 26 of 29 (90%) patients considered that they were able to accurately describe their symptoms. Of note, 14 of 29 (48%) patients wished to use the AMHTD at home, and 20 of 29 (69%) resident physicians wished to obtain the full report of the AMHTD. Although 8 of 29 (28%) residents considered that the device helped to establish the DD, they estimated overall that the AMHTD was neither time-saving nor time-wasting (see Multimedia Appendix 6).

Discussion

Principal Findings

Our results confirmed that the AMHTD significantly allowed the physician to establish a more exhaustive DD (from 59% to 75%). This effect was more important in moderate-complexity (from 56% to 87%) and high-complexity (from 39% to 63%) cases. Of note, the diagnostic list established by the AMHTD was not as accurate as expected (73%, 66/90) and was more precise for low-complexity cases. Overall patient satisfaction (4.3/5) was good, including the ability to accurately describe the presented symptomatology (90%, 26/29). Thus, our results were in agreement with the main factors that guarantee the success of electronic health (eHealth) [31], that is, an improved diagnosis and clinical management, as well as patient-centered care. Our panel of patients presenting to the outpatient unit had common pathologies and was managed by residents at the end of their training. These conditions are common in outpatient services in Switzerland, and our results should be applicable to other hospitals in the country.

Limitations

Our study has some limitations. First, it was an unblinded pilot study with a limited sample size in 1 care center. Therefore, we did not anticipate statistically significant results and did not register our protocol following ethics committee approval. Second, our groups were not balanced as resident physicians in the control group had more years of practice, thus leading to a potential selection bias that could have induced an overestimation of the ability to find a correct DD in the control group. Therefore, the positive effect of 16% (75% [68/90]-59% [50/85]) on the accuracy of the DD might be underestimated. Third, even though our senior physicians were experts in the fields of orthopedics and emergency medicine, the gold standard DD might be flawed, especially in more complex cases. This may be a potential explanation for the observed poorer accuracy of the AMHTD DDs in complex cases. Finally, our AMHTD is still under development, and the reliability of patient responses may be suboptimal, especially because of the absence of images to help in patient symptom localization. This could potentially lead to a degree of uncertainty related to the summary generated. Concerning the DIANNA tool digital content, even if we are fully satisfied with the anamnesis summary, the list of diagnoses might lack accuracy.

Interpretation and Comparison With Prior Research

At present, artificial intelligence systems are still unable to replace physicians for the establishment of a correct DD [31]. Despite this, artificial intelligence allows to complement the work of the physician [32] and even establish an accurate list of problems [33] as shown recently with IBM Watson. The physician's ability to establish a DD can be improved by providing a case summary and a list of possible diagnoses [32,34]. In contrast with other existing digital systems designed to work hand-to-hand with the physician, such as Ada (Ada Health GmbH), K (K Health), and the Mayo Clinic Symptom Checker (Mayo Clinic), DIAANA is focused on the anamnesis rather than the diagnosis, and highly specialized in injury/disease of the musculoskeletal system. To the best of our knowledge, these abovementioned systems have not been challenged in randomized trials. In addition, we were unable to find any relevant literature concerning other similar systems in the field of general medicine or orthopedics. For instance, in the field of psychiatry, a self-report tool allowed the physician to perform a more accurate diagnosis [35]. Similarly, in acute pediatric assessment, it was shown that junior physicians were able to significantly improve the quality of their diagnostic workup and reduce diagnostic omission errors with the use of a Web-based diagnostic reminder system [36]. These observations are concordant with our results as we showed that it was possible to significantly improve the quality of the DD by providing the physician with an exhaustive anamnesis summary and a list of possible DDs. However, in our study, whether the physician was helped by the exhaustive anamnesis summary or by the DD panel remains open. Both may be useful, although we would suggest that the medical history summary may be superior as the DD panel was not as accurate as expected. Indeed, the DD accuracy of the AMHTD alone (73%, 66/90) was slightly superior to the resident physician in the control group (59%, 55/85), but not superior to the resident physician aided by the AMHTD (75%, 68/90). The reliability of the AMHTD DD without the interpretation of the physician is, therefore, not sufficient. On the other hand, the physician may have underestimated the AMHTD DD reliability, as 10% (9/90) of diagnoses were omitted by residents, but suggested by the AMHTD. This means that if the physician had systematically followed the suggestions of the AMHTD, he/she would have found 85% (78/90) of correct DDs instead of 75% (68/90). The physician should be also aware that the correct diagnosis may be absent on the diagnosis list and, in this case, he/she should not waste energy and resources by trying to explore the entire diagnosis list in depth.

The AMHTD presented was conceptualized as a consultation complement for the physician, and not as a substitute. Physician-informatics partnership is the cornerstone of quality of care improvement, not only because it preserves human relationships [31,37], but also because it is the only condition under which diagnostic assistance has been proven to date. In addition to the existing solutions presented above, it has been shown that patients with unresolved medical issues who submitted their cases on the Web to a panel of specialized case-solvers estimated being helped in their diagnosis process in 60% of the cases [38]. We used the DD as a primary outcome rather than the finally retained diagnosis. Even if only the final diagnosis makes clinical sense, it is well known that only an exhaustive DD can lead to a correct diagnosis with any certainty in medical practice. Using the DD as a primary outcome allowed

XSL•FO
RenderX

to increase the effect size because the success rate in establishing a DD is poorer than finding the correct diagnosis. Moreover, to identify situations where a rare but serious diagnosis is missed, thousands of patients should be included if the primary outcome was to be considered as the final diagnosis.

The use of eHealth devices for training purposes is on the rise, as reflected in the increasing use of anamnesis and diagnostic supporting tools used by medical students [39]. We consider that our AMHTD presents ideal characteristics for the training of resident physicians by providing an exhaustive anamnesis and a list of DDs with their degree of emergency and associated factors, as well as initial management guidance. Moreover, the device could be used as a tool for asynchronous teleconsultation.

Workload and workflow disruption are recognized as negative factors influencing the outcome of eHealth interventions [31]. We hypothesized that the exhaustive information collected by the AMHTD would allow the physicians to gain some time. Surprisingly, our physicians estimated that the AMHTD was neither time-saving nor time-wasting. Unfortunately, it was not possible to differentiate the potential time gain for clinical evaluation and reasoning from the time associated with the study itself, for example, contact with the coordinating researcher or waiting for the AMHTD summary to be generated. It is also possible that in low-complexity cases, where the medical history is easily performed, the AMHTD becomes time-consuming. We were unable to measure objectively the consultation time, which may be fragmented when physicians are managing more than one patient at the same time. Completion of the AMHTD form takes some time for patients (20 min in our experience). However, as evidenced by the high satisfaction rate, patients are generally happy to take the necessary time to complete the form. In our study, patients completed the AMHTD form when the waiting time was estimated to be greater than 20 min before the start of the consultation.

Overall, patient satisfaction was good. Of 29 patients, 12 (41%) expressed willingness to keep the AMHTD at home, thus emphasizing the subjective importance for the patients to keep their medical folder and the eHealth tool. We did not provide patients with the AMHTD summary because of the necessity to remain noninterventional in the context of the study for ethical purposes and to avoid causing anxiety to patients when reading highly sensitive DDs. A minority of residents (8/29, 28%) considered the AMHTD as meaningful, and this might reflect the lack of usefulness of the AMHTD for low-complexity cases. Interestingly, 69% (20/29) of physicians wished to obtain the entire AMHTD form, thus potentially highlighting the need to obtain the most accurate and least transformed information as possible, even to the detriment of their time. This contrasts with our initial point of view that the AMHTD summary was sufficient, and the full form would lead to time loss for the physician.

Conclusions

The tested musculoskeletal-focused AMHTD allowed physicians to make a more accurate DD, particularly for complex cases. This could be explained not only by the ability of the AMHTD to propose the right diagnosis but also by the exhaustive anamnesis provided. Patients and physicians expressed overall satisfaction with the process. On the basis of these pilot study results, further research will aim to assess and clarify the following points: confirmation of the findings and a fine-tuned assessment of the accuracy of the established DD, depending on complexity; objective measurement of consultation time; and an evaluation of the physicians' learning curve, both in terms of the accuracy of the DD and duration of the consultation.

Acknowledgments

The authors thank Angèle Gayet-Ageron (Clinical Research Center, University of Geneva, and Geneva University Hospitals) for methodological support and Rosemary Sudan for English revision, as well as Beatriz Villars and Timothée Wuillemin for logistic support (Division of Primary Care Medicine, Department of Community Medicine, Primary Care and Emergency, Department of Medicine, Geneva University Hospitals).

Authors' Contributions

AS, CB, and HS designed the study and provided input throughout the study. CJ, RG, and CL collected the data. HS provided clinical expertise throughout the study and assisted with the finalization of the instrument. CB analyzed the data, assisted by AS and CJ. AS wrote the manuscript together with contributions from all authors. All authors read and approved the final manuscript.

Conflicts of Interest

AS and, to a lesser extent, HS are partners in the limited liability company that owns the DIAANA AMHTD. To decrease any conflicts of interest as much as possible, the choice of the study design and the statistical analyses were the responsibility of CL and CB, with the support of Angèle Gayet-Ageron.

This randomized study was not registered. The editor granted an exception of ICMJE rules for prospective registration of randomized trials because the risk of bias appears low and the study was considered formative. However, readers are advised to carefully assess the validity of any potential explicit or implicit claims related to primary outcomes or effectiveness.

Multimedia Appendix 1
CONSORT - EHEALTH checklist (V 1.6.1).

XSL•FO
RenderX

[PDF File (Adobe PDF File), 27873 KB - medinform_v7i4e14044_app1.pdf]

Multimedia Appendix 2
Example of a patient complaining of muscle cramp.
[PDF File (Adobe PDF File), 912 KB - medinform_v7i4e14044_app2.pdf]

Multimedia Appendix 3
Differential diagnosis list.
[PDF File (Adobe PDF File), 474 KB - medinform_v7i4e14044_app3.pdf]

Multimedia Appendix 4
Types of differential diagnoses selected by the senior physician for each level of complexity.
[PDF File (Adobe PDF File), 19 KB - medinform_v7i4e14044_app4.pdf]

Multimedia Appendix 5
Differential diagnoses found by the automated medical history–taking device.
[PDF File (Adobe PDF File), 83 KB - medinform_v7i4e14044_app5.pdf]

Multimedia Appendix 6
Automated medical history–taking device group: patient and resident physician satisfaction.
[PDF File (Adobe PDF File), 83 KB - medinform_v7i4e14044_app6.pdf]

References

1. Phillips RL, Bartholomew LA, Dovey SM, Fryer GE, Miyoshi TJ, Green LA. Learning from malpractice claims about negligent, adverse events in primary care in the United States. Qual Saf Health Care 2004 Apr;13(2):121-126 [FREE Full text] [doi: 10.1136/qshc.2003.008029] [Medline: 15069219]
2. Holohan TV, Colestro J, Grippi J, Converse J, Hughes M. Analysis of diagnostic error in paid malpractice claims with substandard care in a large healthcare system. South Med J 2005 Nov;98(11):1083-1087. [doi: 10.1097/01.smj.0000170729.51651.f7] [Medline: 16351028]
3. Sandars J, Esmail A. The frequency and nature of medical error in primary care: understanding the diversity across studies. Fam Pract 2003 Jun;20(3):231-236. [doi: 10.1093/fampra/cmg301] [Medline: 12738689]
4. Chandra A, Nundy S, Seabury SA. The growth of physician medical malpractice payments: evidence from the National Practitioner Data Bank. Health Aff (Millwood) 2005(Suppl Web Exclusives):W5-240. [doi: 10.1377/hlthaff.w5.240] [Medline: 15928255]
5. Thomas EJ, Studdert DM, Newhouse JP, Zbar BI, Howard KM, Williams EJ, et al. Costs of medical injuries in Utah and Colorado. Inquiry 1999;36(3):255-264. [Medline: 10570659]
6. Graber M. Diagnostic errors in medicine: a case of neglect. Jt Comm J Qual Patient Saf 2005 Feb;31(2):106-113. [doi: 10.1016/S1553-7250(05)31015-4] [Medline: 15791770]
7. Singh H, Meyer AN, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ Qual Saf 2014 Sep;23(9):727-731 [FREE Full text] [doi: 10.1136/bmjqs-2013-002627] [Medline: 24742777]
8. Schiff G, Kim S, Abrams R, Cosby K, Lambert B, Elstein A, et al. Diagnosing diagnosis errors: lessons from a multi-institutional collaborative project. In: Henriksen K, Battles JB, Marks ES, Lewin DI, editors. Advances in Patient Safety: From Research to Implementation (Volume 2: Concepts and Methodology). Rockville, MD: Agency for Healthcare Research and Quality; 2005:255-278.
9. Scott D, Hallett C, Fettiplace R. Data-to-text summarisation of patient records: using computer-generated summaries to access patient histories. Patient Educ Couns 2013 Aug;92(2):153-159 [FREE Full text] [doi: 10.1016/j.pec.2013.04.019] [Medline: 23746770]
10. Armstrong S. The apps attempting to transfer NHS 111 online. Br Med J 2018 Jan 15;360:k156. [doi: 10.1136/bmj.k156] [Medline: 29335297]
11. Bachman JW. The patient-computer interview: a neglected tool that can aid the clinician. Mayo Clin Proc 2003 Jan;78(1):67-78. [doi: 10.4065/78.1.67] [Medline: 12528879]
12. Bingham P, Lilford RJ, Chard T. Strengths and weaknesses of direct patient interviewing by a microcomputer system in specialist gynaecological practice. Eur J Obstet Gynecol Reprod Biol 1984 Sep;18(1-2):43-56. [doi: 10.1016/0028-2243(84)90032-7] [Medline: 6548716]
13. Simmons Jr EM, Miller OW. Automated patient history-taking. Hospitals 1971 Nov 1;45(21):56-59. [Medline: 5095667]

XSL•FO
RenderX

14. Quaak MJ, Westerman RF, Schouten JA, Hasman A, van Bemmel JH. Computerization of the patient history--patient answers compared with medical records. Methods Inf Med 1986 Oct;25(4):222-228. [Medline: 3773779]
15. Schuman SH, Curry HB, Braunstein ML, Schneeweiss R, Jebaily GC, Glazer HM, et al. A computer-administered interview on life events: improving patient-doctor communication. J Fam Pract 1975 Aug;2(4):263-269. [Medline: 1185132]
16. Bachman J. Improving care with an automated patient history. Fam Pract Manag 2007;14(7):39-43 [FREE Full text] [Medline: 17696057]
17. Arora S, Goldberg AD, Menchine M. Patient impression and satisfaction of a self-administered, automated medical history-taking device in the emergency department. West J Emerg Med 2014 Feb;15(1):35-40 [FREE Full text] [doi: 10.5811/westjem.2013.2.11498] [Medline: 24695871]
18. Slack WV, Leviton A, Bennett SE, Fleischmann KH, Lawrence RS. Relation between age, education, and time to respond to questions in a computer-based medical interview. Comput Biomed Res 1988 Feb;21(1):78-84. [Medline: 3345654]
19. Greist JH, Gustafson DH, Stauss FF, Rowse GL, Laughren TP, Chiles JA. A computer interview for suicide-risk prediction. Am J Psychiatry 1973 Dec;130(12):1327-1332. [doi: 10.1176/ajp.130.12.1327] [Medline: 4585280]
20. Carr AC, Ghosh A, Ancill RJ. Can a computer take a psychiatric history? Psychol Med 1983 Feb;13(1):151-158. [doi: 10.1017/s0033291700050157] [Medline: 6844461]
21. Paperny DM, Aono JY, Lehman RM, Hammar SL, Risser J. Computer-assisted detection and intervention in adolescent high-risk health behaviors. J Pediatr 1990 Mar;116(3):456-462. [doi: 10.1016/s0022-3476(05)82844-6] [Medline: 2308041]
22. Ammenwerth E, Schnell-Inderst P, Hoerbst A. Patient empowerment by electronic health records: first results of a systematic review on the benefit of patient portals. Stud Health Technol Inform 2011;165:63-67. [doi: 10.3233/978-1-60750-735-2-63] [Medline: 21685587]
23. Lancaster K, Abuzour A, Khaira M, Mathers A, Chan A, Bui V, et al. The use and effects of electronic health tools for patient self-monitoring and reporting of outcomes following medication use: systematic review. J Med Internet Res 2018 Dec 18;20(12):e294 [FREE Full text] [doi: 10.2196/jmir.9284] [Medline: 30563822]
24. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. Br Med J 2015 Jul 8;351:h3480 [FREE Full text] [doi: 10.1136/bmj.h3480] [Medline: 26157077]
25. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
26. DIANNA Diagnosis and Anamnesis Assistant. URL: http://www.diaana.fr [accessed 2019-08-09]
27. UpToDate. URL: http://www.utdol.com [accessed 2019-07-22]
28. The National Center for Biotechnology Information. URL: https://www.ncbi.nlm.nih.gov/pubmed [accessed 2019-09-19]
29. Brukner P, Khan K. Brukner & Khan's Clinical Sports Medicine. Fourth Edition. New South Wales, Sydney: McGraw Hill Sports-Medicine; 2012.
30. Viechtbauer W, Smits L, Kotz D, Budé L, Spigt M, Serroyen J, et al. A simple formula for the calculation of sample size in pilot studies. J Clin Epidemiol 2015 Nov;68(11):1375-1379. [doi: 10.1016/j.jclinepi.2015.04.014] [Medline: 26146089]
31. Granja C, Janssen W, Johansen MA. Factors determining the success and failure of eHealth interventions: systematic review of the literature. J Med Internet Res 2018 May 1;20(5):e10235 [FREE Full text] [doi: 10.2196/10235] [Medline: 29716883]
32. Cahan A, Cimino JJ. A learning health care system using computer-aided diagnosis. J Med Internet Res 2017 Mar 8;19(3):e54 [FREE Full text] [doi: 10.2196/jmir.6663] [Medline: 28274905]
33. Devarakonda MV, Mehta N, Tsou C, Liang JJ, Nowacki AS, Jelovsek JE. Automated problem list generation and physicians perspective from a pilot study. Int J Med Inform 2017 Sep;105:121-129 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.05.015] [Medline: 28750905]
34. Kostopoulou O, Rosen A, Round T, Wright E, Douiri A, Delaney B. Early diagnostic suggestions improve accuracy of GPs: a randomised controlled trial using computer-simulated patients. Br J Gen Pract 2015 Jan;65(630):e49-e54 [FREE Full text] [doi: 10.3399/bjgp15X683161] [Medline: 25548316]
35. Brodey B, Purcell SE, Rhea K, Maier P, First M, Zweede L, et al. Rapid and accurate behavioral health diagnostic screening: initial validation study of a web-based, self-report tool (the SAGE-SR). J Med Internet Res 2018 Mar 23;20(3):e108 [FREE Full text] [doi: 10.2196/jmir.9428] [Medline: 29572204]
36. Ramnarayan P, Winrow A, Coren M, Nanduri V, Buchdahl R, Jacobs B, et al. Diagnostic omission errors in acute paediatric practice: impact of a reminder system on decision-making. BMC Med Inform Decis Mak 2006 Nov 6;6:37 [FREE Full text] [doi: 10.1186/1472-6947-6-37] [Medline: 17087835]
37. Moxham C, Chambers N, Girling J, Garg S, Jelfs E, Bremner J. Perspectives on the enablers of e-heath adoption: an international interview study of leading practitioners. Health Serv Manage Res 2012 Aug;25(3):129-137. [doi: 10.1258/hsmr.2012.012018] [Medline: 23135887]
38. Meyer AN, Longhurst CA, Singh H. Crowdsourcing diagnosis for patients with undiagnosed illnesses: an evaluation of CrowdMed. J Med Internet Res 2016 Jan 14;18(1):e12 [FREE Full text] [doi: 10.2196/jmir.4887] [Medline: 26769236]
39. Graber ML, Tompkins D, Holland JJ. Resources medical students use to derive a differential diagnosis. Med Teach 2009 Jun;31(6):522-527. [doi: 10.1080/01421590802167436] [Medline: 19811168]

XSL•FO
RenderX

Abbreviations

AMHTD: automated medical history–taking device
DD: differential diagnosis
DIAANA: DIAgnosis & ANAmnesis
eHealth: electronic health

Edited by G Eysenbach; submitted 17.03.19; peer-reviewed by J Cimino, JA Sánchez-Margallo, K Fuji; comments to author 15.06.19; revised version received 09.08.19; accepted 02.09.19; published 04.11.19.

Please cite as:
Schwitzguebel AJP, Jeckelmann C, Gavinio R, Levallois C, Benaïm C, Spechbach H
Differential Diagnosis Assessment in Ambulatory Care With an Automated Medical History–Taking Device: Pilot Randomized Controlled Trial
JMIR Med Inform 2019;7(4):e14044
URL: http://medinform.jmir.org/2019/4/e14044/
doi:10.2196/14044
PMID:31682590

©Adrien Jean-Pierre Schwitzguebel, Clarisse Jeckelmann, Roberto Gavinio, Cécile Levallois, Charles Benaïm, Hervé Spechbach. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 04.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Automatic Detection of Hypoglycemic Events From the Electronic Health Record Notes of Diabetes Patients: Empirical Study

Yonghao Jin[1], BSc; Fei Li[1], PhD; Varsha G Vimalananda[2,3], MPH, MD; Hong Yu[1,2,4,5], PhD

[1]Department of Computer Science, University of Massachusetts Lowell, Lowell, MA, United States

[2]Center for Healthcare Organization and Implementation Research, Bedford, MA, United States

[3]Section of Endocrinology, Diabetes and Metabolism, School of Medicine, Boston University, Boston, MA, United States

[4]Department of Medicine, University of Massachusetts Medical School, Worcester, MA, United States

[5]Department of Computer Science, University of Massachusetts Amherst, Amherst, MA, United States

Corresponding Author:
Hong Yu, PhD
Department of Computer Science
University of Massachusetts Lowell
220 Pawtucket St
Lowell, MA, 01854
United States
Phone: 1 9789343620
Email: Hong_Yu@uml.edu

Abstract

Background: Hypoglycemic events are common and potentially dangerous conditions among patients being treated for diabetes. Automatic detection of such events could improve patient care and is valuable in population studies. Electronic health records (EHRs) are valuable resources for the detection of such events.

Objective: In this study, we aim to develop a deep-learning–based natural language processing (NLP) system to automatically detect hypoglycemic events from EHR notes. Our model is called the High-Performing System for Automatically Detecting Hypoglycemic Events (HYPE).

Methods: Domain experts reviewed 500 EHR notes of diabetes patients to determine whether each sentence contained a hypoglycemic event or not. We used this annotated corpus to train and evaluate HYPE, the high-performance NLP system for hypoglycemia detection. We built and evaluated both a classical machine learning model (ie, support vector machines [SVMs]) and state-of-the-art neural network models.

Results: We found that neural network models outperformed the SVM model. The convolutional neural network (CNN) model yielded the highest performance in a 10-fold cross-validation setting: mean precision=0.96 (SD 0.03), mean recall=0.86 (SD 0.03), and mean F1=0.91 (SD 0.03).

Conclusions: Despite the challenges posed by small and highly imbalanced data, our CNN-based HYPE system still achieved a high performance for hypoglycemia detection. HYPE can be used for EHR-based hypoglycemia surveillance and population studies in diabetes patients.

(JMIR Med Inform 2019;7(4):e14340) doi:10.2196/14340

KEYWORDS

natural language processing; convolutional neural networks; hypoglycemia; adverse events

Introduction

An estimated 29.1 million Americans aged 20 years or older have diabetes mellitus [1]. Current standards of care call for stringent glycemic control to prevent the complications of diabetes. Intensive drug therapy, particularly in older adults, increases the frequency of hypoglycemia, defined as blood glucose less than 70 mg/dL [2]. Treatment-associated hypoglycemia is the third-most common adverse drug event in patients with diabetes mellitus. Severe hypoglycemia, requiring third-party help or with blood glucose below 54 mg/dL, is associated with seizures, coma, and death and results in about 25,000 emergency department visits and 11,000 hospitalizations annually among Medicare patients in the United States [3]. In

XSL•FO
RenderX

addition, mild hypoglycemia causes troublesome symptoms, such as anxiety, palpitations, and confusion, and is associated with increased mortality. A cross-sectional study of Veterans Health Administration patients with diabetes indicated that 50% of those aged 75 years or older taking insulin and/or sulfonylureas were at risk of hypoglycemia [2].

Electronic health records (EHRs) are important resources for documenting hypoglycemia [3]. However, studies have shown that many hypoglycemic events are not represented within the structured EHR information but are described in EHR notes [4]. Manual chart review could be prohibitively expensive compared to automatic methods [5,6]. Automatically extracting hypoglycemia-related information from EHR notes can be a valuable complement to structured EHR data for guiding the management of diabetes, developing high-risk alerts, monitoring the impact of quality-improvement work, and informing research on hypoglycemia prevention [3]. In clinical settings, similar systems could be used to prefill structured EHR information from patient notes.

However, reliably detecting hypoglycemic events in EHR notes is very challenging. First, the descriptions of hypoglycemia vary broadly across clinical notes (eg, "patient with hypoglycemia," "she has low bs [blood sugar] level," and "bs is in low 20") and it is difficult to manually specify rules to accurately detect all the variations. Second, hypoglycemia, as with most adverse events, is relatively rare. Therefore, it is difficult to collect enough patient data to train a high-performing machine learning model.

In this paper, we are aiming to develop a machine learning–based natural language processing (NLP) system that is able to reliably detect hypoglycemic events from EHR notes. As we are the first group to develop such a system, there are no publicly available reference datasets and baseline models for this task. We assembled an annotated dataset from 500 EHR notes, with sentences labeled as hypoglycemia related or not by experts. We trained and evaluated different sentence classification models on this dataset to find the best model architecture and hyperparameter settings for this task.

Methods

Dataset

With approval from the Institutional Review Board at the University of Massachusetts Medical School, we randomly selected 500 deidentified EHR notes from among all diabetic patients who had been treated at the UMass Memorial Medical Center in 2015. Since hypoglycemia is a relatively rare event in the general population [2,3], we only selected notes containing hypoglycemia code 251 from the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM): *Other disorders of pancreatic internal secretion*. We selected only these notes to increase the frequency of hypoglycemia occurrence and still cover most of the patterns in descriptions of hypoglycemic symptoms.

For annotation, we divided each note into sentences with the natural language toolkit [7]. Two domain experts annotated each sentence as containing a hypoglycemic event (*Positive*) or not (*Negative*). A sentence was annotated as *Positive* if it described any hypoglycemia-related diagnosis or symptoms (eg, "patient has low blood sugar level"). To measure the accuracy of the annotation, we randomly selected 50 annotated EHR notes and asked a third domain expert to review the annotations in those notes. The third domain expert agreed with all existing annotations, which reflects the high quality of our annotation.

Problem Formalization

We formalized the detection of hypoglycemic events as a sentence classification problem: given sentence x, our models will classify its category y as either *Positive* or *Negative*. We proposed three deep learning models to tackle the classification task, the details of which are described in the following section.

Model Designs

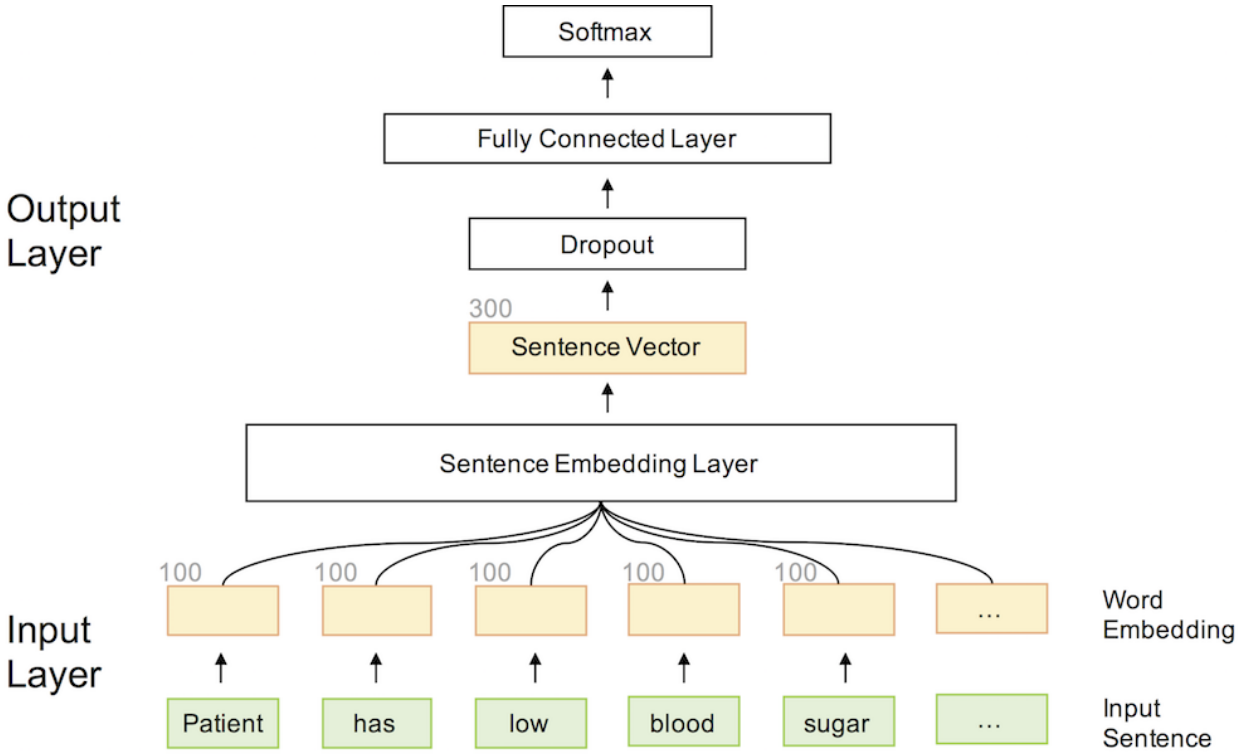
Deep Learning Model

Overview

Deep learning models have been widely adopted in various machine learning tasks, including computer vision [8,9], speech recognition [10], and NLP [11-13]. These models typically take raw data as input and apply one or more hidden layers of transformation to automatically learn the mapping between input and output. Deep learning models have already been investigated in sentence classification problems [14]. In this paper, we followed Kim's work [14] by adopting a feed-forward neural network architecture (see Figure 1). Our model, High-Performing System for Automatically Detecting Hypoglycemic Events (HYPE), is composed of three layers: an input layer, a hidden layer, and an output layer. We investigated three kinds of hidden layers: recurrent neural network (RNN) [15], convolutional neural network (CNN) [16], and temporal convolutional neural network (TCN) [17]. We describe the details of our system in the following sections.

XSL•FO
RenderX

Figure 1. Model architecture of our High-Performing System for Automatically Detecting Hypoglycemic Events (HYPE). The architecture can be divided into three parts: (1) an input layer computing word embeddings for each word, (2) a sentence embedding layer always generating sentence vectors of a fixed dimension regardless of the input sentence length, and (3) an output layer projecting the sentence vector onto a probability score for each class.



Input Layer

Given a sentence, we first tokenized it into l words. We then represented each word by a distributed vector using an embedding resource that was pretrained using Word2Vec on a combined text corpus of PubMed and PubMed Central Open Access [18,19]. In this work, we used 100-dimensional pretrained embeddings. For the words that were not in the pretrained embeddings, we randomly initialized them. Specifically, the input layer takes a tokenized sentence containing l words as input and outputs an ln matrix W, where the i-th row of W is the n-dimensional embedding of the i-th word in the sentence.

Hidden Layer

The dimension of the matrix W we get from the input layer is ln, where l is the sentence length. Therefore, W cannot be directly processed by a standard feed-forward neural network. To handle this problem, we used a hidden layer to transform W to a fixed-length vector C. In this work, we experimented with three variations: RNN, CNN, and TCN.

For RNN, we used long short-term memory (LSTM) [20], which is a common type of neural network for processing sequential data [21,22] (see Figure 2). Given a matrix W, we sequentially fed each row vector into the LSTM unit, along with the hidden vector generated at the previous step. We then used the hidden vector at the previous step, h_l, as the representation of this sentence. At the same time, we could process the sentences in both forward and reverse orders using a bidirectional version of the RNN. The final sentence vector H is the concatenation of the last vectors from both directions h_l and h_l. A formalized description and details of the RNN are provided in Multimedia Appendix 1.

For the CNN, we utilized a widely used architecture [14] (see Figure 3). Specifically, we applied several filters with fixed-length windows to slide on the sentence. For the i-th filter, it generated multiple value $c_i=[c_{i,1}, c_{i,2}, ..., c_{i,l-m+1}]$, where m is the length of the window. Next, a max-over-time pooling was applied to c to produce the output value of this filter. Finally, the outputs of these filters were concatenated to form the sentence representation H. A formalized description and details of the CNN are provided in Multimedia Appendix 1.

For the TCN, we employed a recently proposed architecture [17]. It utilized a one-dimensional fully convolutional network and a causal convolution network at the same time. In a fully convolutional network, the output layer is the same length as the input layer after the convolution operation. The causal convolution ensures that there is no leakage of information from the future to the past (ie, the output at time t is convolved only with elements from time t and earlier in the input layer). Dilated convolution and residual connections were used in each layer to help maintain a long history size and train a deep network [23]. A formalized description and details of the TCN are provided in Multimedia Appendix 1.

XSL•FO
RenderX

Figure 2. Recurrent neural network layer with forward and backward connections. In a unidirectional setting, the backward connections (dashed lines) are absent.

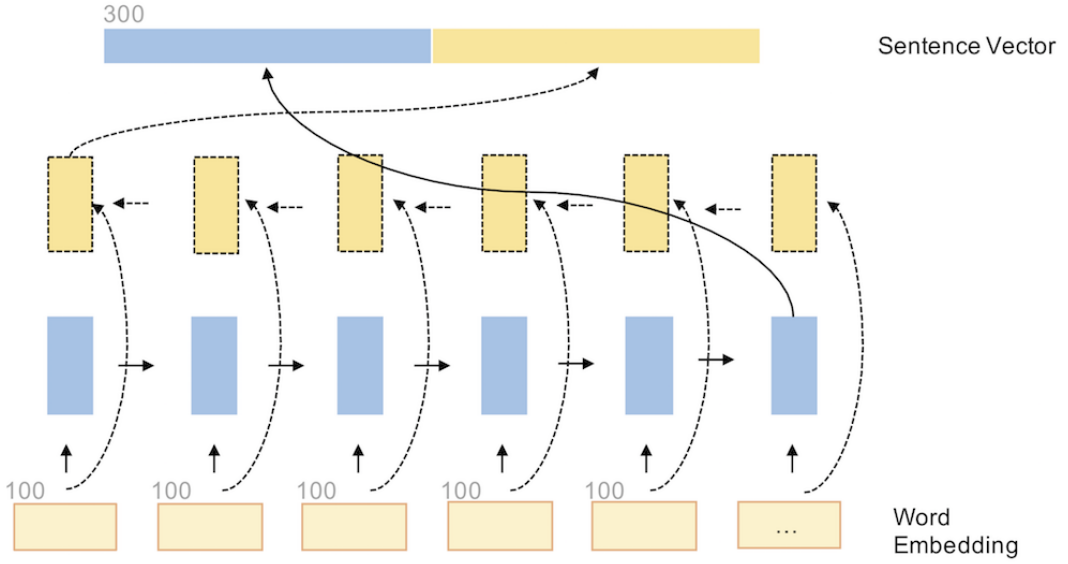
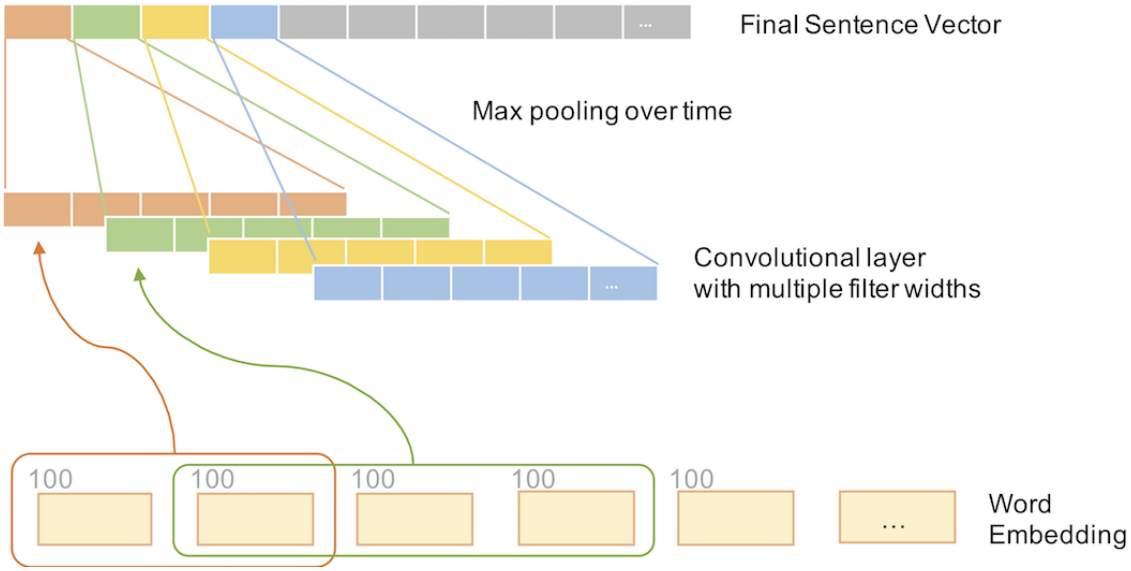


Figure 3. Convolutional neural network layer. Each color represents a different filter with possibly different window size. The max pooling operation produces a single signal value for each filter and the sentence vector is constructed by concatenating signal values from all filters.



Output Layer

The output layer predicts whether the current sentence contains a hypoglycemic event (*Positive*) or not (*Negative*), based on the hidden representation *H* from the previous layer. The output layer squashes the hidden representation to a two-dimensional vector (ie, matrix multiplication) and transforms it to probability scores of *Positive* and *Negative* classes (ie, computing softmax). To train our model, we used the cross-entropy loss and standard backpropagation algorithm. The models were trained for 50 epochs with early stopping (ie, the parameter settings with the best performance on the development set were chosen for evaluation on the testing set).

Baseline Model

We applied support vector machines (SVMs) [24], commonly used learning algorithms for classification problems, as our baseline model. SVMs have been shown to outperform neural network models in some clinical applications [25]. SVMs use kernels to separate data points belonging to different classes in a nonlinearly transformed space. We used the scikit-learn package, version 0.19.0 [26], in Python, version 2.7 (Python Software Foundation), to implement the SVM model and performed grid search for the best hyperparameter settings, such as different kernel functions, down-sampling rate, class weights, penalty parameters, and various n-grams. Training was repeated until convergence of the cost function. We experimented with two kinds of feature vectors: word embedding and *term frequency-inverse document frequency* (TFIDF) matrix. With word embedding vectorization, each sentence is vectorized by the mean of its word embeddings. With TFIDF vectorization, each sentence is vectorized by a long sparse vector with the dimension equal to the vocabulary size. Each dimension of the vector is the TFIDF of the corresponding word in the sentence with respect to the training set corpus; common stop words are removed.

XSL•FO
RenderX

Hyperparameter Settings of Deep Learning Models

We performed a grid search for the optimal hyperparameter settings for the deep learning models using the development set (see Table 1). Overall, the final performance was not very sensitive to the hyperparameter settings. However, we observed that different choices of the learning rate could greatly affect the convergence time. Our best-performing model was trained using the Adam algorithm [27] with an optimum batch size of 64 and learning rate of 5×10^{-5}. To prevent overfitting, we added a dropout layer [28] with an optimum dropout rate of 0.5 in the output layer. The dimension of the word embeddings was set to 100 and the optimum sentence vector setting was 300.

Table 1. Hyperparameter settings in our model.

Hyperparameter	Optimum value	Search range
Learning rate	5×10^{-5}	$\{1\times10^{-3}, 1\times10^{-4}, ..., 1\times10^{-6}\}$
Batch size	64	{16, 32, 64, 128, 256}
Sentence vector size	300	{100, 200, 300, 400, 500}
Dropout rate	0.5	{0.1, 0.2, 0.3, ..., 0.8}
Down-sampling rate	0[a]	{0, 0.1, ..., 1}

[a]The optimum setting had no down-sampling.

Evaluation Metrics

We performed 10-fold cross-validation. The dataset was randomly split into 10 groups of 50 notes. For each fold, we used one group as the testing set and the rest made up the training set. The development set was constructed by randomly selecting 10% of the notes from the training set.

We report recall, precision, and F1 scores for the performance of our models. They are all quantities between 0 and 1. Let P denote the set of the positive instances in the testing dataset and A denote the set of instances that are predicted to be positive by the model. Obviously, the set $P\cap A$ represents the set of positive instances that get correctly classified. Recall is the number of true positive instances divided by the number of positive instances in the dataset (ie, $|P\cap A|/|P|$). Precision is the number of true positive instances divided by the number of predicted positive instances (ie,$|P\cap A|/|A|$). However, either precision or recall is a good measure for model performance. For example, a simple model could consistently predict every instance to be positive and therefore achieve the maximum recall. On the other hand, it could reject every instance and achieve the maximum precision. The F1 score, which is defined by the harmonic mean of the recall and precision (ie, $2\times[precision\times recall]/[precision+recall]$), is a much more objective measure and is common for comparing model performance. In our 10-fold cross-validation scheme, precision, recall, and F1 scores were calculated for each fold, and we report the means and standard deviations for all the folds.

We also report the receiver operating characteristic (ROC) curve, which is created by plotting the true positive rate and false positive rate with different thresholds. However, in a highly imbalanced dataset as in this case, where only 3% of sentences are *Positive*, the ROC curve is not sufficient to reflect the true performances of different models because a classifier could achieve a high-performing ROC curve via bias toward the majority class [29]. Thus, the precision-recall (PR) curve is used to remedy this problem. Because we used 10-fold cross-validation, every sentence in the dataset was assigned to the testing set once and thus received a decision score. The ROC and PR curves were constructed by pooling all the decision scores. We performed two-sample *t* tests for measuring statistical differences between different models.

Results

Dataset

After removing identical sentences from the dataset, the 500 EHR notes contained a total of 41,034 sentences (mean 82, SD 50) with 1316 (3.21%) (mean 2.6, SD 3) annotated as *Positive*. The average number of words per sentence was 11.2 (SD 11), with a minimum of 2 and a maximum of 318. The distribution of positive instances among notes was not particularly even, as is common in the case of adverse events. A total of 387 out of 500 notes (77.4%) contained positive instances and the maximum number of positive sentences from one note was 17. A total of 46.73% (615/1316) of positive sentences mentioned the word *hypoglycemia* directly and 22.11% (291/1316) mentioned keywords concerning blood sugar level; this includes quantitative lab results (eg, “BS [blood sugar] is 68”) or qualitative descriptions (eg, “blood sugar is high”). The rest of the sentences were mostly concerned with various hypoglycemic symptoms (eg, “feeling dizzy”).

Comparisons Between the HYPE and the Baseline Model

As shown in Table 2, all deep learning models outperformed the best baseline SVM model—with TFIDF vectorization and radial basis function kernel—in precision, recall, and F1 scores. For the RNN-based HYPE, LSTM and bidirectional long short-term memory (bi-LSTM) had similar performances. The TCN-based HYPE slightly outperformed the RNN-based HYPE and achieved a balanced precision and recall. The CNN-based HYPE performed the best and was the most time-efficient model due to the simplicity and parallelism of its architecture.

XSL•FO
RenderX

Table 2. Performance of the SVM (support vector machine) baseline and HYPE (High-Performing System for Automatically Detecting Hypoglycemic Events) based on different kinds of neural networks.

Performance measures	SVM	*P* value[a]	LSTM[b]	*P* value	Bi-LSTM[c]	*P* value	TCN[d]	*P* value	CNN[e]	*P* value
Precision, mean (SD)	0.74 (0.07)	<.001	0.91 (0.02)	<.001	0.91 (0.02)	<.001	0.92 (0.03)	.05	0.96 (0.03)	N/A[f]
Recall, mean (SD)	0.57 (0.05)	<.001	0.86 (0.02)	.02	0.87 (0.04)	.10	0.89 (0.04)	N/A	0.86 (0.03)	.10
F1, mean (SD)	0.64 (0.03)	<.001	0.88 (0.02)	<.001	0.88 (0.02)	.001	0.90 (0.02)	.30	0.91 (0.02)	N/A
PR-AUC[g]	0.745	N/A	0.934	N/A	0.942	N/A	0.964	N/A	0.966	N/A
ROC-AUC[h]	0.970	N/A	0.996	N/A	0.997	N/A	0.998	N/A	0.998	N/A

[a]*P* values are based on two-sample *t* tests between the performance of the system and the best-performing system; values <.05 are significant.

[b]LSTM: long short-term memory.

[c]bi-LSTM: bidirectional long short-term memory.

[d]TCN: temporal convolutional neural network.

[e]CNN: convolutional neural network.

[f]N/A: not applicable.

[g]PR-AUC: precision-recall area under the curve.

[h]ROC-AUC: receiver operating characteristic area under the curve.

In terms of the receiver operating characteristic area under the curve (ROC-AUC), all of our models achieved good scores (>0.95) because of the highly imbalanced nature of our dataset. We also reported the precision-recall area under the curve (PR-AUC) value of each model, which is more suitable for skewed datasets [29], as in our case. The ROC and PR curves show that the CNN model has the best PR curve and PR-AUC value (see Figure 4).

Down-Sampling for Data Imbalance

To address data imbalance, we experimented with down-sampling by randomly selecting a subset of the negative training examples at the start of each epoch. We used the best-performing CNN-based HYPE in the down-sampling experiments. As shown in Table 3, down-sampling increased the weight of the minority class, thus increasing the recall. However, the precision dropped because of the lack of the negative examples during training. Therefore, the overall performance decreased when using down-sampling.

Influence of the Training Data Size

To investigate the influence of the training data size on the model performance, we varied the number of examples in the training set. A certain percentage of training examples were randomly selected, while the development and test sets remained the same. We again used the CNN-based HYPE for these experiments. As shown in Table 4, the precision of our model was only sensitive to the training size at the very smallest level.

Figure 4. Precision-recall (PR) and receiver operating characteristic (ROC) curves of each model. Bi-LSTM: bidirectional long short-term memory; CNN: convolutional neural network; LSTM: long short-term memory; SVM: support vector machine; TCN: temporal convolutional neural network.

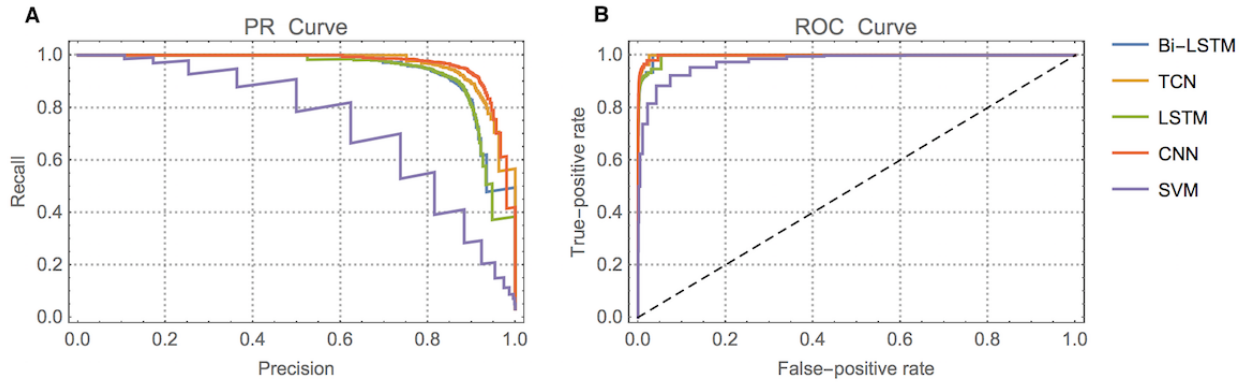


Table 3. Effect of down-sampling on convolutional neural network (CNN) model performance.

Performance measures	Ratio of positive to negative training examples, mean (SD)		
	1:1	1:4	1:9
Precision	0.46 (0.03)	0.86 (0.04)	0.93 (0.03)
Recall	0.92 (0.02)	0.89 (0.03)	0.88 (0.02)
F1	0.62 (0.03)	0.87 (0.03)	0.91 (0.02)

XSL•FO
RenderX

Table 4. Convolutional neural network (CNN) model performance with percentage reduction in training examples.

Performance measures	Percentage reduction in training examples, mean (SD)				
	5%	10%	20%	40%	80%
Precision	0.81 (0.38)	0.97 (0.02)	0.96 (0.02)	0.96 (0.03)	0.95 (0.02)
Recall	0.03 (0.03)	0.43 (0.05)	0.67 (0.03)	0.77 (0.04)	0.85 (0.03)
F1	0.05 (0.05)	0.60 (0.04)	0.79 (0.03)	0.86 (0.03)	0.90 (0.02)

However, recall progressively deteriorated as the training size decreased. As the example size becomes smaller, the model tends to be more conservative about making positive predictions. The overall performance (F1) increases as the number of training examples increases, which is expected.

Discussion

Principal Findings

Our results show that HYPE outperformed SVMs by a large margin in every evaluation metric. One major difference between HYPE and SVMs is how they represent an input sentence. SVMs use bag-of-words and n-grams to represent the input sentence as a sparse vector. In contrast, HYPE uses neural networks to convert the input sentence into a dense vector, which is able to improve the representation ability while avoiding sparsity [18]. Our results also show that neural network models can successfully be trained using a relatively small and imbalanced dataset: a total of 41,034 sentences, of which 1316 sentences were positive instances. The implication is significant as the "knowledge-bottleneck" challenge has made it unrealistic to annotate a large amount of clinical data for supervised machine learning applications.

Comparisons Between Different Hidden Layers of HYPE

In our results, HYPE achieved good performance for detecting sentence-level hypoglycemia, even though the data were imbalanced. We also found that the commonly used approach of down-sampling did not improve performance. While CNN-based HYPE achieved the best precision (mean 0.96, SD 0.03), TCN-based HYPE achieved the best recall (mean 0.89, SD 0.04). One possible explanation for the difference in recall is that CNN is able to capture only the local contextual expressions of hypoglycemic events. TCN is a version of CNN that is equipped with residual connections and diluted convolutions; as such, TCN has the advantage of capturing information in a long context. However, CNN outperformed TCN for the overall performance. CNN also outperformed the two RNN-based models (ie, LSTM and bi-LSTM). This suggests that RNN is less effective than CNN in capturing the contextual patterns of hypoglycemic events. The performance of CNN might be further improved by adding an attention mechanism but we leave this investigation for future work. As for time efficiency, RNN-based HYPE was 10 times slower than the CNN in training. This is because we need to perform many expensive computations in the LSTM units and RNN is hard to parallelize due to its recurrent nature. Thus, CNN is more suitable for our task than RNN.

Effects of Tuning Word Embeddings

A common practice for NLP tasks when working with a small dataset is to fix the pretrained word embeddings during training. The rationale is that when the model encounters a word in the testing set that is not presented in the training set, the model is still able to make correct predictions because its embedding is close to a similar word presented in the training set. However, in our experiments if the embeddings were fixed, we observed a 3%-4% performance loss in F1 score. The best-performing approach was to update word embeddings through backpropagation. The reason for the performance loss of fixed pretrained embeddings might be that the vocabulary size used for describing hypoglycemic events is both small and domain specific. Pretrained embeddings allow a model to attain useful information on general words in the open domain, but fine-tuning word embeddings allows the model to learn domain-specific knowledge. An interesting example is that, if word embeddings were fixed, the model would not be able to discriminate "blood sugar is low" from "blood sugar is high." This may be because the words "high" and "low" have similar distributions in the open domain and because their embeddings are very close to each other. If we tuned their embeddings, the model could learn that "low" and "high" have very different semantics.

Error Analysis

We manually examined the error cases and identified two types of common errors. First, HYPE often failed in cases where hypoglycemic events were indicated by numerical measurements of blood sugar levels. Our model could easily identify sentences such as "BS is low" as hypoglycemic events but it often made mistakes when it encountered sentences such as "BS is 68" or "fsbs [finger stick blood sugar] noted to be 9." Such sentences are difficult to identify for many reasons. One reason is that the word embedding we used in this work transformed numbers to zero during training in order to avoid sparsity [18]. Therefore, the number value was lost in the embedding space and it was impossible for the model to learn a *less than* operation to identify low blood sugar value. Also, the units of the numeric value were often absent and, therefore, needed to be inferred from the context. In the above examples, "68" should be "68 mg/dL" and "9" should be "9 nmol/L." Since such information may not be obtained from the sentence, external human knowledge along with clear definitions for hypoglycemic blood glucose values must be incorporated. In the future, we will explore effective approaches to cope with this issue.

The second type of error was negated events, such as "The patient had no seizures, headaches, abdominal pain, sweating, or other adrenergic symptoms of hypoglycemia." In this

XSL•FO
RenderX

example, HYPE failed to understand the negated word "no" and identified this sentence as a hypoglycemic event. Because the number of such sentences was small, it would be difficult to solve this problem by adding additional features to capture the negation expression. Therefore, we need to incorporate additional approaches for negation identification [30].

Limitations and Future Work

The main limitation of our study is that we selected EHR notes using only diabetes-related ICD-9-CM codes, so the scale of our dataset was relatively small and may not have reflected the true distribution of hypoglycemia sentences in real-world applications. Moreover, because HYPE focuses on sentence-level event detection, it will miss hypoglycemic events that are expressed across multiple sentences. In future work, we will explore document-level hypoglycemic event detection.

Conclusions

In this study, we developed and evaluated state-of-the-art machine learning models to detect hypoglycemia events from EHR notes. We explored three different deep learning models—RNN, CNN, and TCN—and found that the CNN model performed the best, achieving 96% precision and 89% recall. Our work is an important step toward automated surveillance of hypoglycemic events in EHRs and helping clinicians, health care system leaders, and researchers in their efforts to prevent hypoglycemia and to safely manage diabetes mellitus.

Acknowledgments

This work was supported by a grant from the National Heart, Lung, and Blood Institute of the National Institutes of Health (grant number: R01HL125089). The content is solely the responsibility of the authors and does not represent the views of the National Institutes of Health. We would like to acknowledge the annotation team: Heather Keating, Raelene Goodwin, Brian Corner, Nadya Frid, and Feifan Liu.

Authors' Contributions

YJ, FL, and HY conceptualized and designed this study. YJ implemented the tools. YJ and FL processed the data. YJ wrote the manuscript. FL, HY, and VGV reviewed and contributed to editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
A formalized description and details of the recurrent neural network (RNN), the convolutional neural network (CNN), and the temporal convolutional neural network (TCN).
[PDF File (Adobe PDF File), 73 KB - medinform_v7i4e14340_app1.pdf]

References

1. Centers for Disease Control and Prevention. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention; 2011. National diabetes fact sheet: National estimates and general information on diabetes and prediabetes in the United States, 2011 URL: https://www.cdc.gov/diabetes/pubs/pdf/ndfs_2011.pdf [accessed 2019-11-03]
2. Bell DS, Yumuk V. Frequency of severe hypoglycemia in patients with non-insulin-dependent diabetes mellitus treated with sulfonylureas or insulin. Endocr Pract 1997;3(5):281-283. [doi: 10.4158/EP.3.5.281] [Medline: 15251781]
3. Lipska KJ, Ross JS, Wang Y, Inzucchi SE, Minges K, Karter AJ, et al. National trends in US hospital admissions for hyperglycemia and hypoglycemia among Medicare beneficiaries, 1999 to 2011. JAMA Intern Med 2014 Jul;174(7):1116-1124 [FREE Full text] [doi: 10.1001/jamainternmed.2014.1824] [Medline: 24838229]
4. Workgroup on Hypoglycemia, American Diabetes Association. Defining and reporting hypoglycemia in diabetes: A report from the American Diabetes Association Workgroup on Hypoglycemia. Diabetes Care 2005 May;28(5):1245-1249. [doi: 10.2337/diacare.28.5.1245] [Medline: 15855602]
5. Hu Z, Melton GB, Moeller ND, Arsoniadis EG, Wang Y, Kwaan MR, et al. Accelerating chart review using automated methods on electronic health record data for postoperative complications. AMIA Annu Symp Proc 2016;2016:1822-1831 [FREE Full text] [Medline: 28269941]
6. Martinez D, Ananda-Rajah MR, Suominen H, Slavin MA, Thursky KA, Cavedon L. Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans. J Biomed Inform 2015 Feb;53:251-260 [FREE Full text] [doi: 10.1016/j.jbi.2014.11.009] [Medline: 25460203]
7. Bird S, Klein E, Loper E. Natural Language Processing with Python: Analyzing Text With the Natural Language Toolkit. Sebastopol, CA: O'Reilly Media; 2009.
8. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May 24;60(6):84-90 [FREE Full text] [doi: 10.1145/3065386]

XSL•FO
RenderX

9. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large-scale visual recognition challenge. Int J Comput Vis 2015 Apr 11;115(3):211-252 [FREE Full text] [doi: 10.1007/s11263-015-0816-y]
10. Hinton G, Deng L, Yu D, Dahl G, Mohamed A, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process Mag 2012 Nov;29(6):82-97. [doi: 10.1109/msp.2012.2205597]
11. Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning. 2008 Presented at: 25th International Conference on Machine Learning; July 5-9, 2008; Helsinki, Finland URL: http://dl.acm.org/citation.cfm?id=1390177 [doi: 10.1145/1390156.1390177]
12. Socher R, Pennington J, Huang E, Ng A, Manning C. Semi-supervised recursive autoencoders for predicting sentiment distributions. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011 Presented at: Conference on Empirical Methods in Natural Language Processing; July 27–31, 2011; Edinburgh, Scotland p. 151-161 URL: https://www.aclweb.org/anthology/D11-1014.pdf
13. Socher R, Lin C, Manning C. Parsing natural scenes and natural language with recursive neural networks. In: Proceedings of the 28th International Conference on Machine Learning. 2011 Jun 28 Presented at: 28th International Conference on Machine Learning; June 28-July 2, 2011; Bellevue, WA URL: http://machinelearning.wustl.edu/mlpapers/paper_files/ICML2011Socher_125.pdf
14. Yoon K. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 26-28, 2014; Doha, Qatar. [doi: 10.3115/v1/D14-1181]
15. Schuster M, Paliwal K. Bidirectional recurrent neural networks. IEEE Trans Signal Process 1997 Nov 01;45(11):2673-2681. [doi: 10.1109/78.650093]
16. Collins M, Duffy N. Convolution kernels for natural language. In: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. 2001 Presented at: 14th International Conference on Neural Information Processing Systems: Natural and Synthetic; December 3-8, 2001; Vancouver, BC p. 625-632. [doi: 10.7551/mitpress/1120.003.0085]
17. Bai S, Kolter J, Koltun V. arXiv. 2018 Mar 04. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling URL: http://arxiv.org/abs/1803.01271 [accessed 2018-08-24]
18. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13). Volume 2. 2013 Presented at: 26th International Conference on Neural Information Processing Systems (NIPS'13). Volume 2; December 5-10, 2013; Lake Tahoe, NV p. 3111-3119.
19. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional semantics resources for biomedical text processing. In: Proceedings of the 5th International Symposium on Languages in Biology and Medicine. 2013 Presented at: 5th International Symposium on Languages in Biology and Medicine; December 12-13, 2013; Tokyo, Japan.
20. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735]
21. Jagannatha AN, Yu H. Bidirectional RNN for medical event detection in electronic health records. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 13-15, 2016; San Diego, CA p. 473-482 URL: https://www.aclweb.org/anthology/N16-1056.pdf [doi: 10.18653/v1/n16-1056]
22. Jagannatha AN, Yu H. Structured prediction models for RNN-based sequence labeling in clinical text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2016 Presented at: Conference on Empirical Methods in Natural Language Processing; November 1-5, 2016; Austin, TX p. 856-865 URL: https://www.aclweb.org/anthology/D16-1082.pdf [doi: 10.18653/v1/d16-1082]
23. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition; June 26-July 1, 2016; Las Vegas, NV URL: https://arxiv.org/pdf/1512.03385.pdf [doi: 10.1109/CVPR.2016.90]
24. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/BF00994018]
25. Munkhdalai T, Liu F, Yu H. Clinical relation extraction toward drug safety surveillance using electronic health record narratives: Classical learning versus deep learning. JMIR Public Health Surveill 2018 Apr 25;4(2):e29 [FREE Full text] [doi: 10.2196/publichealth.9361] [Medline: 29695376]
26. Pedregosa G, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
27. Kingma D, Ba J. Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations. 2015 Presented at: 3rd International Conference on Learning Representations; May 7-9, 2015; San Diego, CA URL: https://arxiv.org/pdf/1412.6980.pdf

XSL•FO
RenderX

28. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 2014 Jun 14;15:1929-1958 [FREE Full text]
29. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. 2006 Presented at: 23rd International Conference on Machine Learning; June 25-29, 2006; Pittsburgh, PA. [doi: 10.1145/1143844.1143874]
30. Agarwal S, Yu H. Biomedical negation scope detection with conditional random fields. J Am Med Inform Assoc 2010;17(6):696-701 [FREE Full text] [doi: 10.1136/jamia.2010.003228] [Medline: 20962133]

Abbreviations

bi-LSTM: bidirectional long short-term memory
bs/BS: blood sugar
CNN: convolutional neural network
EHR: electronic health record
fsbs: finger stick blood sugar
HYPE: High-Performing Systems for Automatically Detecting Hypoglycemic Events
ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification
LSTM: long short-term memory
NLP: natural language processing
PR: precision-recall
PR-AUC: precision-recall area under the curve
RNN: recurrent neural network
ROC: receiver operating characteristic
ROC-AUC: receiver operating characteristic area under the curve
SVM: support vector machine
TCN: temporal convolutional neural network
TFIDF: term frequency-inverse document frequency

Edited by G Eysenbach; submitted 12.04.19; peer-reviewed by S Musy, M Torii; comments to author 14.05.19; revised version received 08.07.19; accepted 19.10.19; published 08.11.19.

Please cite as:
Jin Y, Li F, Vimalananda VG, Yu H
Automatic Detection of Hypoglycemic Events From the Electronic Health Record Notes of Diabetes Patients: Empirical Study
JMIR Med Inform 2019;7(4):e14340
URL: http://medinform.jmir.org/2019/4/e14340/
doi:10.2196/14340
PMID:31702562

©Yonghao Jin, Fei Li, Varsha G Vimalananda, Hong Yu. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 08.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Fast Healthcare Interoperability Resources (FHIR) as a Meta Model to Integrate Common Data Models: Development of a Tool and Quantitative Validation Study

Emily Rose Pfaff[1], MS; James Champion[1], BS; Robert Louis Bradford[1], BS; Marshall Clark[1], BS; Hao Xu[2], PhD; Karamarie Fecho[2], PhD; Ashok Krishnamurthy[2], PhD; Steven Cox[2], BS; Christopher G Chute[3], MD, DrPH; Casey Overby Taylor[3], PhD; Stan Ahalt[2], PhD

[1]North Carolina Translational and Clinical Sciences Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[2]Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[3]Johns Hopkins University, Baltimore, MD, United States

Corresponding Author:
Emily Rose Pfaff, MS
North Carolina Translational and Clinical Sciences Institute
University of North Carolina at Chapel Hill
160 N Medical Drive
Chapel Hill, NC, 27599
United States
Phone: 1 919 843 4712
Email: epfaff@email.unc.edu

Abstract

Background: In a multisite clinical research collaboration, institutions may or may not use the same common data model (CDM) to store clinical data. To overcome this challenge, we proposed to use Health Level 7's Fast Healthcare Interoperability Resources (FHIR) as a meta-CDM—a single standard to represent clinical data.

Objective: In this study, we aimed to create an open-source application termed the Clinical Asset Mapping Program for FHIR (CAMP FHIR) to efficiently transform clinical data to FHIR for supporting source-agnostic CDM-to-FHIR mapping.

Methods: Mapping with CAMP FHIR involves (1) mapping each source variable to its corresponding FHIR element and (2) mapping each item in the source data's value sets to the corresponding FHIR value set item for variables with strict value sets. To date, CAMP FHIR has been used to transform 108 variables from the Informatics for Integrating Biology & the Bedside (i2b2) and Patient-Centered Outcomes Research Network data models to fields across 7 FHIR resources. It is designed to allow input from any source data model and will support additional FHIR resources in the future.

Results: We have used CAMP FHIR to transform data on approximately 23,000 patients with asthma from our institution's i2b2 database. Data quality and integrity were validated against the origin point of the data, our enterprise clinical data warehouse.

Conclusions: We believe that CAMP FHIR can serve as an alternative to implementing new CDMs on a project-by-project basis. Moreover, the use of FHIR as a CDM could support rare data sharing opportunities, such as collaborations between academic medical centers and community hospitals. We anticipate adoption and use of CAMP FHIR to foster sharing of clinical data across institutions for downstream applications in translational research.

(JMIR Med Inform 2019;7(4):e15199) doi:10.2196/15199

KEYWORDS

health information interoperability; electronic health records; data sharing; controlled vocabularies

XSL•FO
RenderX

Introduction

Background

The proliferation of common data models (CDMs) for electronic health record (EHR) data has had a positive impact on cross-institutional data sharing and large-scale participant recruitment [1-3]. At present, the 3 major clinical CDMs in use by the academic community are Informatics for Integrating Biology & the Bedside (i2b2) [4], Patient-Centered Outcomes Research Network (PCORnet) [5], and Observational Medical Outcomes Partnership (OMOP) [6], each of which uses a slightly different architecture to achieve the same result: to represent and store EHR data in a relational database. The ability to query common data structures and provision data to collaborators in a shared format reduces the burden on data analysts and enforces common definitions that allow clinical data to be appropriately merged and compared across institutions. However, despite these affordances, there is no guarantee that all institutions involved in a multisite collaboration are using the same CDM, potentially negating the advantage.

The challenge of cross-institutional sharing of clinical data has risen in the context of the Biomedical Data Translator program [7-9], funded by the National Center for Advancing Translational Sciences. The Translator program aims "to design and prototype a 'Translator' system capable of integrating existing biomedical data sets...and 'translating' those data into insights that can accelerate translational research, support clinical care, and leverage clinical expertise to drive research innovations" [8]. Clinical data are central to the program and critical for its success. Yet, despite the importance of clinical data, Translator teams have not adopted a uniform CDM to enable the sharing of clinical data across the consortium. Moreover, even if current Translator teams were to adopt a uniform CDM, (1) future Translator collaborators and users may not be positioned to support the agreed-upon model and (2) the dynamic and complex nature of clinical data could render an agreed-upon model quickly obsolete.

Common Data Models: Current State

The challenge of cross-institutional EHR data sharing is by no means limited to the Translator program [10]. Institutions wishing to engage in data sharing may not support the same CDM—perhaps one uses i2b2, whereas another uses OMOP. In such cases, it is likely not possible for one institution to simply agree to stand up a new CDM to accommodate the other institution. Mapping institutional EHR data to any one of the major CDMs is resource- and personnel-intensive and requires an ongoing commitment to maintain and refresh infrastructure and data over time. The North Carolina Translational and Clinical Sciences Institute (NC TraCS), home of University of North Carolina at Chapel Hill's National Institutes of Health-funded Clinical and Translational Science Award (CTSA), participates in 2 i2b2-powered networks (CTSA Accrual to Clinical Trials [ACT] and the Carolinas Collaborative) and one PCORnet-powered network (Stakeholders, Technology, and Research Clinical Research Network [STAR CRN]). In a poll of STAR collaborators, we discovered that maintaining CDM infrastructure consumes, on average, just over 1 full-time equivalent (FTE) per CDM per year. Initial implementation effort varies by CDM, but ranged from 0.8 to 3.8 FTE in our poll (see Table 1). As institutions are asked to adopt more CDMs (and the number of available CDMs multiply), this level of effort increases and can quickly become untenable, even with existing expertise, education, and documentation. Moreover, as can be seen in Table 1, the effort expended can differ greatly between sites, with some sites needing to expend far more resources than others to achieve the same goals.

The core function of a CDM is to enable clinical data harmonization and interoperability. CDMs thus share a goal with Health Level 7's Fast Healthcare Interoperability Resources (HL7's FHIR), a health care data representation standard increasingly supported by major EHR vendors. In FHIR, clinical data are split into *resources* or data domains. As of version 4.0.0, FHIR provides 22 nondraft-status Base resources, such as Patient, Practitioner, and Encounter, and 33 nondraft-status Clinical resources, such as Procedure, Observation, and MedicationRequest. Other types of FHIR resources include Foundation, Financial, and Specialized [11]. Each resource comprises structured fields that describe the resource—for example, the Encounter resource contains fields to capture the type of encounter, length of stay, and discharge disposition. In addition to defining specific resources and fields, FHIR enforces the use of established code sets (eg, LOINC, SNOMED CT, and ICD-9/-10) or FHIR-specific value sets in many of its fields to maximize standardization. Where provided fields and code sets are not sufficient, FHIR offers the ability for individual users and organizations to build *extensions* to the standard to capture data that are not explicitly defined by HL7 [12]. Considering these characteristics, FHIR can also be considered a CDM [13].

A testament to the connection between FHIR and CDMs is the fact that several efforts have already been initiated to map either i2b2, PCORnet, or OMOP to FHIR [14-19]. The software applications described in these previous studies each map a single-source data model to FHIR; to be able to map additional data models, a user must use multiple applications.

XSL•FO
RenderX

Table 1. Effort expended by Stakeholders, Technology, and Research Clinical Research Network sites to stand up and maintain common data models.

Site[a] and number of common data models (CDMs) currently maintained	Number of full-time equivalent (FTE) to stand up *one* new CDM	Number of FTE to maintain all CDMs
University of North Carolina Chapel Hill		
3 (PCORnet[b], i2b2[c] [2 separate ontologies])	Informatics: 1.0	Informatics: 2.0
	Project Management: 0.5	Project Management: 1.0
Site 1		
3 (PCORnet, i2b2, OMOP[d])	Informatics : 2.3	Informatics: 5.0
	Project Management: 1.5	Project Management: 2.0
Site 2		
3 (PCORnet, i2b2 [2 separate ontologies])	Total: 2.5	Informatics: 3.0
		Project Management: 2.0
Site 3		
6 (PCORnet, i2b2, OMOP, 3 regional models)	Informatics: 2.5	Informatics: 3.0
	Project Management: 0.3	Project Management: 0.3
Site 4		
2 (PCORnet, i2b2)	Total: 0.8	Total: 0.6

[a]Non-University of North Carolina STAR sites have been masked. Sites that did not differentiate between project management and informatics FTE have their effort reported as "Total."

[b]Patient-Centered Outcomes Research Network.

[c]Informatics for Integrating Biology & the Bedside.

[d]Observational Medical Outcomes Partnership.

Objective of This Study

Rather than continuing to treat each of these source data models individually, we proposed to improve interoperability, standardization, and semantic harmonization by enabling transformation from any of these (or other) models to FHIR with a single, source-agnostic tool, Clinical Asset Mapping Program for FHIR (CAMP FHIR), that can read from any CDM and map to its straightforward views. This approach will facilitate a multi-institutional collaboration by providing an application that harmonizes across CDMs.

Mapping an individual CDM to FHIR is resource-intensive; collaborating institutions may have *mismatched* CDMs (in terms of model, version, or both); and new CDMs will likely continue to emerge. Although not a replacement for CDMs, on a project-by-project basis, FHIR-formatted data generated by CAMP FHIR can enable easier cross-site data harmonization, supplementing the advances that CDMs have already made in this space. Upon widespread adoption, CAMP FHIR and applications such as these could encourage eventual uptake of FHIR as a *meta-CDM*—a single standard to represent clinical data sourced from any data model.

Methods

Designing the Transformation Pipeline

Simply proposing FHIR as yet another CDM for institutions to map their EHRs to would add to, not alleviate, the institutional burden illustrated in Table 1. To avoid this burden and associated costs, we designed CAMP FHIR (and its mapping process) to (1) leverage as much existing CDM mapping and curation work as possible and (2) allow our team to share our mapping work with others, giving other institutions an opportunity to use CAMP FHIR with minimal site-specific changes and resource expenditure required. Thus far, CAMP FHIR has been used to transform data from the i2b2 and PCORnet data models, though the application is designed to allow input from any source data model.

To use FHIR as a meta-CDM, it is important to recognize that unlike i2b2, OMOP, and PCORnet, FHIR was designed to be a standard for data *exchange*—not data persistence. Thus, any CDM-to-FHIR mapping effort involves translating a relational CDM to a serialized format. CAMP FHIR was developed expressly to address this challenge and to make the conversion process as simple as possible.

CAMP FHIR is designed to apply the FHIR standard to EHR data from any source system, although we focused our initial development on the CDMs used at our institution, specifically i2b2 and PCORnet. We began with i2b2, which can be considered a CDM when used with a standard ontology. (University of North Carolina; UNC's local i2b2 uses a custom ontology and is, thus, not strictly a CDM, but we have also created a version of our i2b2 mapping scripts that uses the i2b2 ACT ontology, which is standardized.) We first profiled the overlap and gaps between the i2b2 schema (version 1.7.10) and the corresponding FHIR resources (version 3.0.1) and then began to map individual variables (see *Mapping Details*, below).

XSL•FO
RenderX

As we mapped, we quickly faced the challenge of accounting for inconsistencies between the out-of-the-box i2b2 schema and our institution's modified local version. The challenge was compounded by our knowledge that other institutions that use i2b2 have their own local modifications to contend with, although this is less of an issue if the use of standard ontologies is enforced. Regardless, the realization that some site-specific flexibility would be necessary led us to develop CAMP FHIR with the assumption that many local data sources have idiosyncrasies (even if they are CDMs). (As an example, laboratory reference ranges can be very difficult to harmonize, with different organizations storing the range either as a single field or 2 fields [low end and high end], and with or without comparators such as < and >.) Considering this, it should therefore be the responsibility of the local database layer to model views to conform to CAMP FHIR's specific input format. Putting the mapping responsibility on the database layer (rather than within the application itself) provides more flexibility and portability, giving the application the capability to interface with any clinical relational database schema. To achieve this architecture, we chose to use the object-relational mapping tool, Hibernate, which is an open-source Java (Oracle) persistence framework for mapping a relational database to an object-oriented domain model.

After completing the mappings for i2b2, we then created a separate set of mappings for the PCORnet CDM (version 4.1). PCORnet is much stricter about its data model than i2b2, not allowing for local variation in structure or code sets. For this reason, the PCORnet CAMP FHIR mappings are especially portable and would require few (if any) changes to run at any site using the PCORnet CDM.

We provide the PCORnet and ACT mappings as *starter scripts* in the CAMP FHIR GitHub repository [20] to acclimate new users to the tool. To use CAMP FHIR, users run these scripts (or their own versions) to create views within their source database that conform to our CAMP FHIR standard, populate a code mapping table for any value sets, and point CAMP FHIR at the database.

Mapping Details

Using CAMP FHIR to map to FHIR from any source data model involves two major tasks: (1) mapping each source variable to its corresponding FHIR element; and (2) for variables with strict value sets (eg, race, smoking status, and discharge disposition), mapping each item in the source model's value set to the corresponding FHIR value set item. We completed these tasks for both the i2b2 and PCORnet data models.

Our i2b2 and PCORnet data marts contain data for 2.9 million patients, with data spanning from July 2004 to the present. Our i2b2 data mart has values populated for 100% of the variables supported by the ACT ontology [21]. Our PCORnet data mart supports all version 4.1 tables [22] other than OBS_GEN, DISPENSING, and DEATH_CAUSE.

Informatics for Integrating Biology & the Bedside (i2b2)

UNC at Chapel Hill's local i2b2 implementation contains data in the following domains: patient demographics, encounter details, diagnoses, procedures, point-of-care location, patient vital signs, laboratory tests, medications, clinical observations, social history, and insurer. To map i2b2 to FHIR, a group of 3 informaticians experienced with the underlying structure and data definitions within UNC's local i2b2 (1) took each unique concept in a given domain (eg, *diagnosis date* from the domain Diagnosis), (2) reviewed the FHIR documentation for the corresponding resource for that domain (eg, FHIR's condition resource), (3) determined which field within that FHIR resource was the best fit for the source concept, and (4) recorded the suggested mapping for inclusion in one of the CAMP FHIR views.

For variables with strict value sets, an additional step was necessary. If the best-fit FHIR field had its own strict value set, then each value set item in the source set was mapped to its nearest equivalent in the FHIR set. All value set mappings were stored in a single table, which was then loaded into the i2b2 database. An excerpt from this mapping table is provided in Table 2.

To adhere to our goal of standardization, we opted not to create custom value sets for use within FHIR, opting instead to use the exact value sets provided in the FHIR specification. The tradeoff for this strict adherence to standardization is potential loss of data or loss of granularity. Not every source value set item had an equivalent in FHIR. For example, when the source value was *Other* or another generic catch-all, there was generally not a match in the FHIR set. At this time, unmappable items in our source value sets are left null in the FHIR version of the data. There were also several instances where the FHIR value set was less granular than the source dataset, resulting in a loss of detail after mapping. An example is discharge disposition, where UNC's local value set contains 44 choices, and FHIR's value set contains 11. The current version of the i2b2 value set mapping table (using the ACT ontology) can be found in the CAMP FHIR GitHub repository [20].

All mapping tasks were divided among the 3 informaticians, with each person's mappings peer-reviewed by the other 2. After the mappings were finalized, the mapping team defined database views for each mapped domain. As the views themselves are completely independent of the i2b2 data model, even though they were designed during our i2b2 mapping exercise, they ultimately became the generic set of *CAMP FHIR views* that are packaged with the tool for use with any data model.

For i2b2, the views served to transform the data from the native star schema (with the majority of the data stored in a central fact table, OBSERVATION_FACT) to a normalized format more easily consumable by CAMP FHIR. The code snippet in Textbox 1 shows the construction of the OBSLABS_2FHIR view from OBSERVATION_FACT.

For views containing variables that needed value set conversions (eg, smoking status descriptors in the Observation [Vitals] view), we joined to the prepopulated mapping table when creating the view and populated the FHIR version of each value set item rather than the *local* option.

Patient-Centered Outcomes Research NetworkPCORnet mapping proceeded in much the same way as i2b2; each table,

XSL•FO
RenderX

variable, and value set was mapped to FHIR following the steps outlined above, and a value set transformation table was loaded into the PCORnet database. The current version of the PCORnet value set mapping table can be found in the CAMP FHIR GitHub repository [20]. The code snippet in Textbox 2 shows the construction of the OBSLABS_2FHIR view from the LAB_RESULT_CM table. (Note the join to our custom PCORNET_FHIR_MAPPING table to transform the value sets for RESULT_MODIFIER and ABN_IND.)

Table 2. Excerpt from University of North Carolina's Informatics for Integrating Biology and the Bedside-Fast Healthcare Interoperability Resources mapping table.

TABLE_CD	COLUMN_CD	LOCAL_IN_CD	FHIR_OUT_CD	FHIR_SYSTEM
VISIT_DIMENSION	INOUT_CD	EMERGENCY	EMER	https://hl7.org/fhir/STU3/v3/ActEncounterCode/vs.html
VISIT_DIMENSION	INOUT_CD	INPATIENT	IMP	https://hl7.org/fhir/STU3/v3/ActEncounterCode/vs.html
VISIT_DIMENSION	INOUT_CD	OUTPATIENT	AMB	https://hl7.org/fhir/STU3/v3/ActEncounterCode/vs.html

Textbox 1. Structured Query Language (SQL) to create the Clinical Asset Mapping Program for Fast Healthcare Interoperability Resources view OBSLABS_2FHIR from Informatics for Integrating Biology & the Bedside 's OBSERVATION_FACT textbox.

```
select distinct
 ofc.patient_num||'-'||ofc.encounter_num||'-'||ofc.provider_id||'-'||
    to_char(ofc.start_date, 'DD-MON-YYYY')||'-'||ofc.concept_cd||'-'||
    ofc.instance_num as OBS_IDENTIFIER,
 'Patient/'||ofc.patient_num as OBS_SUBJECT_REFERENCE,
 'Encounter/'||ofc.encounter_num as OBS_CONTEXT_REFERENCE,
 'http://hl7.org/fhir/ValueSet/observation-category' as OBS_CATEGORY_SYST,
 'laboratory' as OBS_CATEGORY_CODE,
 'Laboratory' as OBS_CATEGORY_DISPLAY,
 'http://loinc.org' as OBS_CODE_CODING_SYST,
  ofc.concept_cd as OBS_CODE_CODING_CODE,
  cd.NAME_CHAR as OBS_CODE_CODING_DISPLAY,
  ofc.nval_num as OBS_VALUEQUANTITY_VALUE,
  case when ofc.VALTYPE_CD = 'N' and ofc.TVAL_CHAR = 'E' then null
     when ofc.VALTYPE_CD = 'N' and ofc.TVAL_CHAR = 'L' then '<'
     when ofc.VALTYPE_CD = 'N' and ofc.TVAL_CHAR = 'G' then '>'
     when ofc.VALTYPE_CD = 'N' and ofc.TVAL_CHAR = 'LE' then '<='
     when ofc.VALTYPE_CD = 'N' and ofc.TVAL_CHAR = 'GE' then '>='
     else null end as OBS_VALUEQUANTITY_COMPARATOR,
  ofc.units_cd as OBS_VALUEQUANTITY_CODE,
  case when ofc.VALTYPE_CD = 'T' then ofc.TVAL_CHAR
     else null end as OBS_VALUESTRING,
  ofc.START_DATE as OBS_ISSUED,
  null as OBS_EFFECTIVEDATETIME
from
  observation_fact ofc
     left join concept_dimension cd on ofc.concept_cd=cd.concept_cd
     inner join visit_dimension vd ON vd.encounter_num = ofc.encounter_num
where
  ofc.concept_cd like 'LOINC%'
```


XSL•FO
RenderX

Textbox 2. Structured Query Language (SQL) to create the Clinical Asset Mapping Program Fast Healthcare Interoperability Resources view OBSLABS_2 Fast Healthcare Interoperability Resources from Patient-Centered Outcomes Research Network's LAB_RESULT_CM table.

```
select distinct
  labs.LAB_RESULT_CM_ID as OBS_IDENTIFIER,
  'Patient/'||labs.PATID as OBS_SUBJECT_REFERENCE,
  'Encounter/'||labs.ENCOUNTERID as OBS_CONTEXT_REFERENCE,
  'http://hl7.org/fhir/ValueSet/observation-category' as OBS_CATEGORY_SYST,
  'laboratory' as OBS_CATEGORY_CODE,
  'Laboratory' as OBS_CATEGORY_DISPLAY,
  'http://loinc.org' as OBS_CODE_CODING_SYST,
  LAB_LOINC as OBS_CODE_CODING_CODE,
  null as OBS_CODE_CODING_DISPLAY,
  labs.RESULT_NUM as OBS_VALUEQUANTITY_VALUE,
  nvl(tcc1.FHIR_OUT_CD,null) as OBS_VALUEQUANTITY_COMPARATOR,
  case
    when labs.RESULT_UNIT = 'NI' then null
    else labs.RESULT_UNIT
    end as OBS_VALUEQUANTITY_CODE,
  case
    when labs.RESULT_QUAL = 'NI' then null
    else nvl(labs.RESULT_QUAL,labs.RAW_RESULT)
    end as OBS_VALUESTRING,
  labs.RESULT_DATE as OBS_ISSUED,
  nvl(labs.SPECIMEN_DATE,labs.LAB_ORDER_DATE) as OBS_EFFECTIVEDATETIME,
  case
    when labs.NORM_MODIFIER_LOW IN ('EQ,''GE,''GT,''NO') then
    labs.NORM_MODIFIER_LOW||' '||labs.NORM_RANGE_LOW
    else labs.NORM_RANGE_LOW
    end as OBS_REFRANGE_LOW,
  case
    when labs.NORM_MODIFIER_HIGH IN ('EQ,''GE,''GT,''NO') then
    labs.NORM_MODIFIER_HIGH||' '||labs.NORM_RANGE_HIGH
    else labs.NORM_RANGE_HIGH
    end as OBS_REFRANGE_HIGH,
  nvl(tcc2.FHIR_OUT_CD,null) as OBS_INTERPRETATION_CODE,
  'http://hl7.org/fhir/ValueSet/observation-interpretation' as
  OBS_INTERPRETATION_SYST
from
  lab_result_cm labs
  left join PCORNET_FHIR_MAPPING tcc1 on tcc1.column_cd='RESULT_MODIFIER'
  and labs.RESULT_MODIFIER=tcc1.local_in_cd
  left join PCORNET_FHIR_MAPPING tcc2 on tcc2.column_cd='ABN_IND' and
  labs.ABN_IND=tcc2.local_in_cd
```

In contrast with the i2b2 mapping exercise, we found that we had more gaps and mismatches to handle between PCORnet and FHIR owing to PCORnet's much stricter data model. Mapping results fell into 3 categories: (1) variable and/or value set was mappable and was mapped; (2) variable and/or value set was mappable and will be mapped in a future CAMP FHIR release; or (3) variable and/or value set has no equivalent (or no exact equivalent) in FHIR and cannot be mapped either partially or fully. A list of variables and value sets in the third category of PCORnet data are provided in Tables 3 and 4.

Regardless of source data model, we operationalize the Hibernate mappings using the open-source HAPI-FHIR API, which is an implementation of the HL7 FHIR specification for Java. HAPI-FHIR supports all versions of FHIR, although CAMP-FHIR currently supports FHIR version 3. Taken together, this setup allows CAMP FHIR to read in the mapped data (via Hibernate), convert to the FHIR standard (via HAPI-FHIR), and output valid FHIR files in XML or JavaScript Object Notation (JSON) format. This process is illustrated with fictitious data in Figure 1.

CAMP FHIR is intended to transform CDM data for a given cohort, rather than an entire warehouse of EHR data. We have found performance to be quite efficient with a predefined cohort, as detailed in Table 5.

XSL•FO
RenderX

Table 3. Patient-Centered Outcomes Research Network 4.1 data with no (noncustom) exact Fast Healthcare Interoperability Resources equivalent.

Table[a,b]	Field(s) with no Fast Healthcare Interoperability Resources (FHIR) equivalent
DEMOGRAPHIC	SEXUAL_ORIENTATION, GENDER_IDENTITY, BIOBANK_FLAG
DIAGNOSIS	DX_ORIGIN, DX_POA
PROCEDURE	PX_SOURCE
VITAL	VITAL_SOURCE, BP_POSITION, TOBACCO[c], TOBACCO_TYPE
LAB_RESULT_CM	RESULT_LOC
PRO_CM	Entire table cannot be mapped
PRESCRIBING	RX_SOURCE
DEATH	Entire table (other than DEATH_DATE) cannot be mapped
DEATH_CAUSE	Entire table cannot be mapped

[a]This table is inclusive of all PCORnet 4.1 fields that did not map to one of the FHIR resources accounted for in the *current version* of CAMP FHIR, which does not include all PCORnet fields. There may be additional unmappable fields uncovered in future versions of CAMP FHIR. Current resources are: Patient, Encounter, Condition, Procedure, Observation, MedicationRequest, and Practitioner.

[b]PCORnet 4.1 tables not intended to hold EHR data are not accounted for here: ENROLLMENT, PCORNET_TRIAL, and HARVEST

[c]Note that this refers specifically to smokeless tobacco. Smoking status is mappable.

Table 4. Patient-Centered Outcomes Research Network 4.1 value sets with no (noncustom) exact Fast Healthcare Interoperability Resources equivalents.

Value set[a]	Comment
DEMOGRAPHIC.RACE	No Fast Healthcare Interoperability Resources (FHIR) value for multiple races
ENCOUNTER.ENC_TYPE	No FHIR equivalent for visits of type EI (emergency department admit to inpatient hospital stay), IC (institutional professional consult)
ENCOUNTER.DISCHARGE_STATUS	Imperfect FHIR equivalents for several discharge statuses; 17 possible values in Patient-Centered Outcomes Research Network (PCORnet) versus 11 in FHIR; values were mapped where possible.
ENCOUNTER.ADMITTING_SOURCE	Imperfect FHIR equivalents for several admitting sources; 16 possible values in PCORnet versus 10 in FHIR; values were mapped where possible.

[a]PCORnet 4.1 values of *No information*, *Unknown*, and *Other* were rarely mappable to FHIR and are not noted each time.

XSL•FO
RenderX

Figure 1. An example of demographic data transformation. CAMP FHIR: Clinical Asset Mapping Program for Fast Healthcare Interoperability Resources; i2b2: Informatics for Integrating Biology & the Bedside.

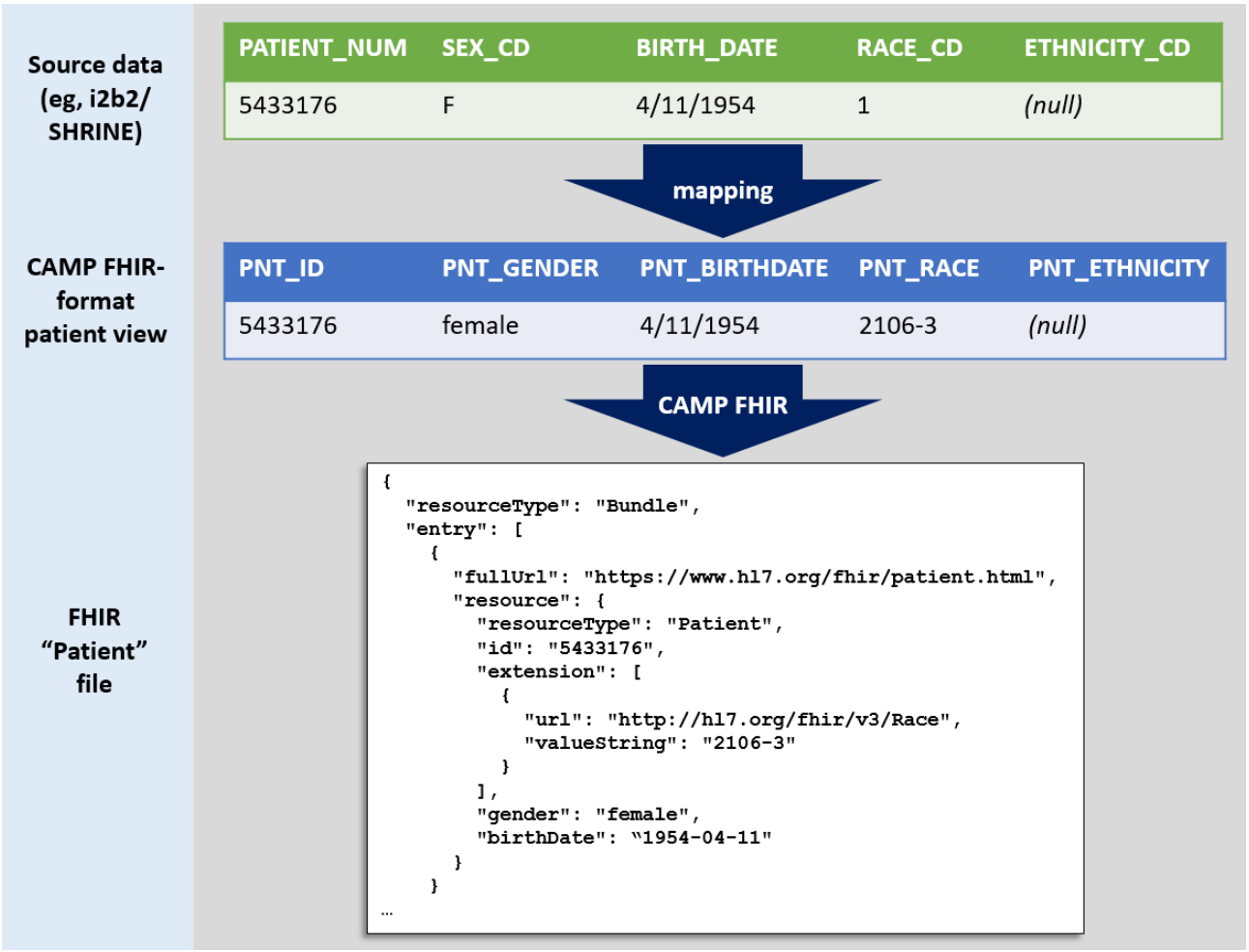


Table 5. Clinical Asset Mapping Program Fast Healthcare Interoperability Resources's (CAMP FHIR) performance extracting data from the Patient-Centered Outcomes Research Network common data model.

Domain	Time to populate database view[a] (seconds)	Time to write JavaScript Object Notation files to disk (seconds)	Number of records
Patient	6	6	15,945
Condition	480	415	2,766,556
Encounter	200	115	1,010,823
Observation (Labs)	390	350	2,081,826
Observation (Vitals)	360	250	1,663,897
Medication Request	450	420	2,435,813
Practitioner	7	7	36,749
Procedure	80	80	442,921

[a]Database server specifications: OS: Red Hat Enterprise Linux Server release 6.10 (Santiago), Processor: Intel(R) Xeon(R) CPU E5-2690 v2 @ 3.00GHz, Database: Oracle 12.1.0.2.0 (Enterprise Edition), Database memory_target: 2 GB, Database size: 464 GB.

Results

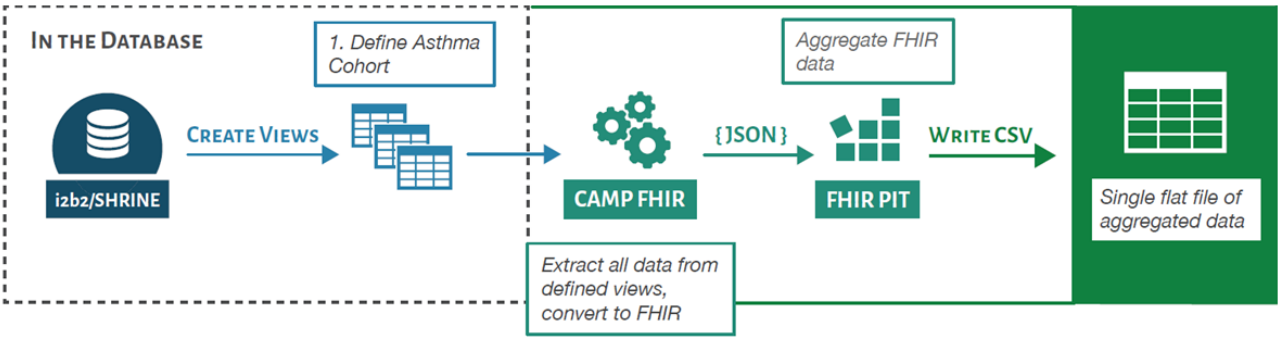
Asthma Use Case

The JSON-formatted FHIR files output by CAMP FHIR would rarely be the end deliverable for any project. Rather, the FHIR files are a launching point for further transformation, as can be seen in the context of our work with the Translator program. For Translator, we have used CAMP FHIR to extract and transform data from our institution's i2b2 database on approximately 23,000 patients with asthma, including their associated encounters, laboratory results, vital signs, diagnoses, procedures, medications, and smoking status. (Although this particular use case is using an asthma cohort, the same processes and level of effort would apply to any defined cohort; nothing in the transformation effort described here is specific to asthma.)

XSL•FO
RenderX

For this use case, the JSON-formatted FHIR files output by CAMP FHIR were then ingested by a second application, termed FHIR PIT (FHIR Patient data Integration Tool) [23]. FHIR PIT is a custom, open-source application that was developed as part of the Translator program to integrate FHIR-formatted clinical data with environmental exposures data (ie, airborne pollutant exposures, roadway exposures, and socioeconomic exposures) for downstream application in translational research. The resulting data then are accessible via an API endpoint, termed Integrated Clinical and Environmental Exposures Service (ICEES) [24]. As we use FHIR as a standard, any Translator institution or non-Translator institution is able to use CAMP FHIR to transform their source clinical data and use FHIR PIT to provision integrated data via ICEES, with very little local variation. The CAMP FHIR to FHIR PIT to ICEES process is illustrated in Figure 2.

Figure 2. The Clinical Asset Mapping Program fast healthcare interoperability resources (CAMP FHIR) pipeline as used for translator. CSV: comma-separated value; JSON: JavaScript Object Notation; PIT: Patient data Integration Tool.



Validation

To validate the output from CAMP FHIR, we compared the ICEES clinical data generated by the CAMP FHIR/FHIR PIT pipeline with equivalent clinical data for the same patient cohort extracted directly from UNC Health Care System's enterprise data warehouse, the Carolina Data Warehouse for Health (CDWH). The validation process included the generation of summary statistics for each variable from the 2 data files, including patient counts, mean values, standard deviations, and quartile values. As we iterated through the validation process, we did encounter issues with the software that needed to be corrected. For example, we initially identified inconsistencies in medication data, which we discovered was because of the fact that our i2b2 instance uses RxNorm to code medications, and our warehouse uses nonstandard medication IDs (generated by our EHR). This issue was resolved by referencing a crosswalk between RxNorm codes and our internal medication IDs to translate between the 2 code sets. Any other similar issues causing inconsistencies between the 2 datasets were ultimately discovered and resolved.

In the end, we were able to successfully demonstrate that the final ICEES output of clinical data from CAMP FHIR/FHIR PIT matched exactly the raw extract of clinical data from the CDWH (ie, our validation test, after code corrections, demonstrated 100% accuracy in the mappings). A total of 53 ICEES fields were mapped to FHIR, including 3 Encounter resource mappings on ED and inpatient visits, 4 patient resource mappings on demographics, 1 Observation resource mapping on BMI, 25 condition resource mappings on diagnoses, and 20 MedicationRequest resource mappings on prescribed medications. Note that the particular set of fields chosen did not include any field that was unmappable or resulted in a loss of granularity, thus ensuring a faithful translation for this particular use case. A list of the ICEES fields is available on the ICEES GitHub repository [25].

Discussion

Principal Findings

We have shown that CAMP FHIR is a sound method for conversion of relational clinical data to the HL7 FHIR format. On the basis of our experience with the Translator program and our participation in various clinical data research networks, we believe that for certain projects, the use of CAMP FHIR to harmonize clinical data across institutions will save resources over the alternative of standing up matching CDMs. Moreover, because we have made our mapping work public, we are hopeful that new users of CAMP FHIR will not always need to undertake this mapping effort themselves, so long as they use one of CAMP FHIR's supported CDMs. If a CAMP FHIR user wanted to build a new set of mappings to support an entirely new data model, 4 weeks of effort split among two informaticians (with appropriate understanding of the CDM in question) would be a reasonable estimate for the amount of effort required to map value sets and variables and perform peer review.

The biggest challenge encountered during the mapping process was limiting subjectivity as much as possible, which we handled with peer review, and resolving mapping decisions where disagreements occurred. Another challenge (with no immediate solution) was determining how to handle valid source system data with no match in FHIR—eg, if a patient's race is *multiple*, without the exact races specified, FHIR has no standard way of storing that information. That means the patient's race becomes null in the FHIR version of their record, unless custom values or extensions are used. Depending on the use case or research question, these losses could be significant. At this time, the way to handle this issue is through documentation, so that the user understands what data can and cannot be represented through the CAMP FHIR process.

Indeed, we did find examples of loss of data (where source data have no good equivalent in FHIR), change of data meaning

XSL•FO
RenderX

(where FHIR equivalents are close, but not an exact match), or loss of granularity (where FHIR value sets have less detail than source value sets). These issues are not uncommon in data transformation in general, and are certainly not limited to transformations to FHIR. In particular, FHIR's current lack of coverage for data on cause of death, patient reported outcomes, genomics, or patient gender identity (to pick a few examples) may disqualify it for use in answering certain research questions. It is important, then, to put data mapped to FHIR (or any transformed data) in its proper context, and acknowledge that at present, FHIR is likely not ready (yet) to be a single source of truth for clinical research data. For a given use case, if highly granular detail from the source system is important to the research question and that detail is lost during transformation to FHIR, then CAMP FHIR may not be sufficient in and of itself for that study. In short, no data model is the right choice for all applications. This should particularly be taken into account where institutions have no choice as to which data model to use, such as studies that must use CDISC ODM/SDTM standards for FDA compliance. However, our hope is that FHIR's breadth of data domains and wide adoption would allow it to serve a large variety of use cases, if not all.

Despite its many potential benefits for data harmonization, output from CAMP FHIR would not serve as a replacement for CDMs (and certainly not enterprise clinical data warehouses). Rather, CAMP FHIR output is better suited to handling data for a defined cohort, particularly in the context of a multi-institutional collaboration involving multiple CDMs. If a participating institution is able to take advantage of the prepackaged mapping scripts included with CAMP FHIR, CAMP FHIR will reduce, though not eliminate, cost and effort barriers to participation in such a collaboration. Although there is no particular size limitation on such a cohort, attempting to store millions of patient records in FHIR files could be unwieldy from a file-size and data-manipulation perspective. However, that limitation alone does not discount FHIR's value as a potential data persistence model, even if the data to be persisted cover individual patient cohorts. The adoption of FHIR as a persistence model is strengthened by the reality that many organizations can export data directly from their EHR using ubiquitous FHIR APIs, thus obviating any translation pathway through other CDMs. This assumes the institutions can agree upon a consistent version of FHIR, which, as is the case with many CDMs, can cause mismatched schemas even within the same data model. Assuming such version agreement is possible, academic medical centers might leverage such a FHIR persistence layer to consolidate data from legacy CDMs, ongoing EHR updates, and accretions from research protocols.

If an institution is capable of natively outputting FHIR files from its EHR, whereas a collaborator prefers to use a CDM as its data source, there is no reason why the native FHIR output could not be combined with the CAMP FHIR output. This provides additional CAMP FHIR use cases—eg, to support rare data sharing opportunities, such as collaborations between academic medical centers and community hospitals. As EHRs increasingly adopt FHIR as a standard for data transmission, it will be far more likely for nonacademic clinical organizations to be able to produce FHIR-formatted data using their EHR than they are to stand up an instance of i2b2, PCORnet, or OMOP, which are more commonly found at academic medical centers. The ability to combine CAMP FHIR output with native FHIR could thus help to democratize the opportunity to participate in data-driven clinical research.

Future Work

Future work will look beyond data harmonization toward the variety of ways in which the output from CAMP FHIR can be used. FHIR output is intended to be used in a variety of downstream applications, as was done as part of the Translator program with ICEES. Other possibilities include consumption and display by a Web application, consumption by an EHR, or conversion to another data format such as Resource Description Framework (RDF). RDF is an example of another interoperability-focused technology that may prove useful in interinstitutional clinical data sharing in the near future. In this context, CAMP FHIR would thus be situated as middleware between raw clinical data and its ultimate use case.

On the basis of the successful implementation and application of CAMP FHIR at our institution, another logical next step is to formally evaluate its performance at another institution for further testing and validation. A critical metric to track will be the amount of local configuration (and effort) necessary to run the application at an outside institution, as users should only need to make minimal changes to implement the pipeline locally. In general, the more *strict* the source CDM (eg, PCORnet), the less we expect local variation to necessitate mapping changes. Less strict CDMs may require more local changes, though the structure of the queries and the *FHIR side* of the mappings should remain constant. In the near future, we plan to (1) add additional views in future releases to cover more FHIR resources, such as Coverage, Location, Medication Dispense, and Medication Administration; (2) build in OMOP mappings; and (3) introduce support for FHIR version 4.0.

Conclusions

The Translator program envisions a future in which the entire range of biomedical data, from clinical data to data derived from chemistry, genomics, anatomy, and beyond, is accessible within a unified framework. Such a framework will allow translational research questions to be formulated and answered via query and computation over federated, interoperable data models. As part of the Translator program, we saw a need for unifying heterogeneous clinical data models from collaborating institutions. CAMP FHIR was motivated by a need to foster the sharing of clinical data across Translator institutions for downstream applications in translational research. As CAMP FHIR's utility ultimately extends beyond the Translator use case, we anticipate its adoption and use across the CTSA consortium and other clinical and translational research collaborations facing a need to harmonize clinical data.

XSL•FO
RenderX

Acknowledgments

The authors wish to thank the partners in the STAR CRN, who kindly contributed data about their expended effort in implementing and maintaining institutional CDMs. The authors also acknowledge Kelsey Urgo's assistance with Figure 2, Nessim Abu-Saif's assistance with validating the PCORnet 4.1 mappings, and Paul Kovach's contributions to the ACT ontology mappings. This study was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, grant numbers OT3TR002019, OT3TR002020, and UL1TR002489, and PCORI grant CDRN-1306-04869.

Authors' Contributions

Initial manuscript draft was prepared by: ERP and JC; programming and data analysis was carried out by: ERP, JC, Bradford, MC, and HX; workflow design was done by: ERP, JC, RLB, MC, HX, KF, AK, SC, CGC, and COT; project leadership was by: SA and AK; manuscript revisions and final approval was done by: ERP, JC, RLB, MC, HX, KF, AK, SC, CGC, COT, and SA.

Conflicts of Interest

None declared.

References

1. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: enabling nationally scalable multi-site disease studies. PLoS One 2013;8(3):e55811 [FREE Full text] [doi: 10.1371/journal.pone.0055811] [Medline: 23533569]
2. Heerman WJ, Jackson N, Roumie CL, Harris PA, Rosenbloom ST, Pulley J, et al. Recruitment methods for survey research: findings from the mid-south clinical data research network. Contemp Clin Trials 2017 Nov;62:50-55. [doi: 10.1016/j.cct.2017.08.006] [Medline: 28823925]
3. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc 2015 May;22(3):553-564 [FREE Full text] [doi: 10.1093/jamia/ocu023] [Medline: 25670757]
4. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130 [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
5. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc 2014;21(4):578-582 [FREE Full text] [doi: 10.1136/amiajnl-2014-002747] [Medline: 24821743]
6. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. Ann Intern Med 2010 Nov 2;153(9):600-606. [doi: 10.7326/0003-4819-153-9-201011020-00010] [Medline: 21041580]
7. Biomedical Data Translator Consortium. The biomedical data translator program: conception, culture, and community. Clin Transl Sci 2019 Mar;12(2):91-94 [FREE Full text] [doi: 10.1111/cts.12592] [Medline: 30412340]
8. Biomedical Data Translator Consortium. Toward a universal biomedical data translator. Clin Transl Sci 2019 Mar;12(2):86-90 [FREE Full text] [doi: 10.1111/cts.12591] [Medline: 30412337]
9. Austin CP, Colvis CM, Southall NT. Deconstructing the translational tower of babel. Clin Transl Sci 2019 Mar;12(2):85 [FREE Full text] [doi: 10.1111/cts.12595] [Medline: 30412342]
10. Nordo AH, Levaux HP, Becnel LB, Galvez J, Rao P, Stem K, et al. Use of EHRs data for clinical research: Historical progress and current applications. Learn Health Syst 2019 Jan;3(1):e10076 [FREE Full text] [doi: 10.1002/lrh2.10076] [Medline: 31245598]
11. HL7 Fundamentals. Resource Index URL: https://www.hl7.org/fhir/resourcelist.html [accessed 2019-08-07]
12. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Sep;23(5):899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
13. Chute CG, Huff SM. The pluripotent rendering of clinical data for precision medicine. Stud Health Technol Inform 2017;245:337-340. [doi: 10.3233/978-1-61499-830-3-337] [Medline: 29295111]
14. Boussadi A, Zapletal E. A fast healthcare interoperability resources (FHIR) layer implemented over i2b2. BMC Med Inform Decis Mak 2017 Aug 14;17(1):120 [FREE Full text] [doi: 10.1186/s12911-017-0513-6] [Medline: 28806953]
15. HL7 Fundamentals. DAF-Research Profile List and Mappings from FHIR to PCORnet CDM and OMOP CDM URL: http://hl7.org/fhir/us/daf-research/2017Jan/daf-research-profile.html [accessed 2019-03-05]
16. The FHIR Project at Georgia Tech. URL: http://omoponfhir.org/ [accessed 2018-11-02]
17. Wagholikar KB, Mandel JC, Klann JG, Wattanasin N, Mendis M, Chute CG, et al. SMART-on-FHIR implemented over i2b2. J Am Med Inform Assoc 2017 Mar 1;24(2):398-402 [FREE Full text] [doi: 10.1093/jamia/ocw079] [Medline: 27274012]

XSL•FO
RenderX

18. Jiang G, Kiefer RC, Sharma DK, Prud'hommeaux E, Solbrig HR. A consensus-based approach for harmonizing the OHDSI common data model with HL7 FHIR. Stud Health Technol Inform 2017;245:887-891 [FREE Full text] [doi: 10.3233/978-1-61499-830-3-887] [Medline: 29295227]
19. Paris N, Mendis M, Daniel C, Murphy S, Tannier X, Zweigenbaum P. i2b2 implemented over SMART-on-FHIR. AMIA Jt Summits Transl Sci Proc 2018;2017:369-378 [FREE Full text] [doi: 10.1093/jamia/ocw079] [Medline: 29888095]
20. GitHub. CAMP FHIR URL: https://github.com/NCTraCSIDSci/camp-fhir [accessed 2019-06-18]
21. ACT Network Wiki. ACT Ontology Version 2.0.1 URL: https://ncatswiki.dbmi.pitt.edu/acts/wiki/ACT%20Ontology%20Version%202.0.1 [accessed 2019-08-07]
22. PCORnet. Common Data Model (CDM) Specification, Version 4.1 URL: https://pcornet.org/download/pcornet-common-data-model-v4-1-specification-5-may-2018/?wpdmdl=1919&refresh=5d43390adacf81564686602 [accessed 2019-08-07]
23. GitHub. FHIR PIT URL: https://github.com/NCATS-Tangerine/FHIR-PIT [accessed 2019-08-07]
24. Fecho K, Pfaff E, Xu H, Champion J, Cox S, Stillwell L, et al. A novel approach for exposing and sharing clinical data: the translator integrated clinical and environmental exposures service. J Am Med Inform Assoc 2019 Apr 26 (forthcoming). [doi: 10.1093/jamia/ocz042] [Medline: 31077269]
25. GitHub. NCATS-Tangerine/icees-api URL: https://github.com/NCATS-Tangerine/icees-api [accessed 2019-08-07]

Abbreviations

ACT: Accrual to Clinical Trials Network
CAMP FHIR: Clinical Asset Mapping Program for FHIR
CDM: common data model
CTSA: Clinical and Translational Science Award
EHR: electronic health record
FHIR PIT: Fast Healthcare Interoperability Resources Patient data Integration Tool
FTE: full-time equivalent
HL7 FHIR: Health Level 7 Fast Healthcare Interoperability Resources
i2b2: Informatics for Integrating Biology & the Bedside
ICEES: Integrated Clinical and Environmental Exposures Service
JSON: JavaScript Object Notation
LOINC: logical observation identifiers names and codes
NC TraCS: North Carolina Translational and Clinical Sciences Institute
OMOP: Observational Medical Outcomes Partnership
PCORnet: Patient-Centered Outcomes Research Network
RDF: Resource Description Framework
SNOMED CT: Systematized Nomenclature of Medicine—Clinical Terms
STAR: Stakeholders, Technology, and Research Clinical Research Network
UNC: University of North Carolina

Edited by C Lovis; submitted 26.06.19; peer-reviewed by M Abdelhamid, M Dugas; comments to author 30.07.19; revised version received 12.08.19; accepted 16.08.19; published 16.10.19.

Please cite as:
Pfaff ER, Champion J, Bradford RL, Clark M, Xu H, Fecho K, Krishnamurthy A, Cox S, Chute CG, Overby Taylor C, Ahalt S
Fast Healthcare Interoperability Resources (FHIR) as a Meta Model to Integrate Common Data Models: Development of a Tool and Quantitative Validation Study
JMIR Med Inform 2019;7(4):e15199
URL: https://medinform.jmir.org/2019/4/e15199
doi:10.2196/15199
PMID:31621639

©Emily Rose Pfaff, James Champion, Robert Louis Bradford, Marshall Clark, Hao Xu, Karamarie Fecho, Ashok Krishnamurthy, Steven Cox, Christopher G Chute, Casey Overby Taylor, Stan Ahalt. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 16.10.2019. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete

XSL•FO
RenderX

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

XSL•FO
RenderX

Viewpoint

Digital Health and the State of Interoperable Electronic Health Records

Jessica Germaine Shull[1,2], MSc, MA, PhD

[1]PhD Program, Department of Biomedicine, University of Barcelona, Barcelona, Spain

[2]Institute for Biomedical Research Bellvitge, Hospital Bellvitge, Barcelona, Spain

Corresponding Author:
Jessica Germaine Shull, MSc, MA, PhD
PhD Program, Department of Biomedicine
University of Barcelona
Oficina F15 A Florensa, 8
Barcelona, 08028
Spain
Phone: 34 627122319
Email: jess.shull@gmail.com

Abstract

Digital health systems and innovative care delivery within these systems have great potential to improve national health care and positively impact the health outcomes of patients. However, currently, very few countries have systems that can implement digital interventions at scale. This is partly because of the lack of interoperable electronic health records (EHRs). It is difficult to make decisions for an individual or population when the data on that person or population are dispersed over multiple incompatible systems. This viewpoint paper has highlighted some key obstacles of current EHRs and some promising successes, with the goal of promoting EHR evolution and advocating for frameworks that develop digital health systems that serve populations—a critical goal as we move further into this data-rich century with an ever-increasing number of patients who live longer and depend on health care services where resources may already be strained. This paper aimed to analyze the evolution, obstacles, and current landscape of EHRs and identify fundamental areas of hindrance for interoperability. It also aimed to highlight countries where advances have been made and extract best practices from these examples. The obstacles to EHR interoperability are not easily solved, but improving the current situation in countries where a national policy is not in place will require a focused inquiry into solutions from various sources in the public and private sector. Effort must be made on a national scale to seek solutions for optimally interoperable EHRs beyond status quo solutions. A list of considerations for best practices is suggested.

(JMIR Med Inform 2019;7(4):e12712) doi:10.2196/12712

KEYWORDS

EHRs; health information technology; machine learning in health

Introduction

Digital health systems and innovative care delivery within these systems have great potential to improve national health care and positively impact the health outcomes of patients. However, currently, very few countries have systems that can implement digital interventions at scale. This is, in part, because of the well-known lack of interoperable electronic health records (EHRs). EHRs are the connective tissue of a health system; yet, most countries have systems that cannot unite the information of all their citizens because even within 1 city, the software used in 1 hospital is incompatible with that used in another, although it may have been procured from the same company.

Once digital health records are interoperable, digital health systems may function on a national scale rather than piecemeal or in isolation. Treatment of noncommunicable diseases could benefit from digital health coaching, personalized delivery of care, and quality of care improvements [1-3]. In addition, there may be thousands of patients who share the symptoms or side effects or respond in revealing patterns, leading to new treatments and personalized medicine, and even new disease surveillance tools could be developed [4]. However, we cannot explore or create these tools without aggregating and sharing patient data under a common set of (secure) standards. Without interoperability, using these tools or implementing remote patient monitoring products is extremely complex or simply happens in small samples. This potential use of health information technology has been discussed in papers dating

XSL•FO
RenderX

back to 2005 [5] and analyzed for impact once EHRs were well established [6]; however, because of legacy systems and the design of those systems, the choices of EHRs are limited and incapable of algorithmic analysis across their entire client base. Epic (Epic Systems Corporation), for example, is an EHR software first designed in 1979 [7], and although the code has evolved since then, the company commands a 10-fold price difference compared with more agile solutions with no evident improvement in quality [8]. Epic is simply the most prevalent EHR available in the United States; the company now archives at least part of the health records of approximately half the US population [8].

Partners HealthCare is a sprawling 10-hospital system in Boston, which, in 2015 to 2016, spent US $1.2 billion [9] implementing an upgraded EHR system from Epic. The intention was to decrease errors and unify a previously disparate network that made it difficult for physicians in 1 hospital within the system to cross-reference information from a patient who had been at another [10]. When the system launched, 1000 Epic employees were required on hand to troubleshoot, which was perhaps foreseen, but the technical issues that persist today may not have been calculated. As recently as February 2019, the system was down for several hours [11], and it does not perform integration that patients are beginning to expect, such as integrating data from a connected glucose monitor. The upgrade was simply not designed with this feature. This integration is possible with a consolidation of the Apple HealthKit and the Dexcom Share2 app, as piloted on an Epic system in 2016 [12], but this extra patching work perhaps should not be required with a billion-dollar price tag. Epic is by no means alone on the list of EHR companies that are faced with a more technically savvy client base who expect better interfaces and services, and they are trying to become more agile; in 2018, the company announced One Virtual System Worldwide, which allows providers to access patient information from other institutions [13], which is a notable step but perhaps not on par with the data integration and analysis capabilities of companies such as Google and Apple.

Digital health investment is at an all-time high, with nearly US $7 billion invested in 2018 in the United States alone [14]. However, most of the solutions being funded are not addressing the administrative foundations of the health care system such as EHRs. Investment in artificial intelligence (AI) for health care, a perhaps more tantalizing venture, is estimated to reach US $6.6 billion by 2021 [15]. Meanwhile, the current state of real workflows at hospitals is not sustainable.

Physicians in the United States may spend half their day filling out patient histories. A 2018 study in Family Medicine found that of 982 patient visits that each lasted on average for 35.8 min, 19.3 min were spent on the EHR [16]. It is imperative that health systems gain time (and quality) where they can. In Spain, there is a notable delay to health care access; the Spanish Ministry of Health reported that in 2018, wait times for surgery had improved but were still 137 days in certain regions [17]. Systems subject to a lack of efficiency cannot afford more time, and valuable analysis is lost because of poor information systems.

Machine Learning With Electronic Health Records

AI is often presented as a solution to ease some of this burden. In a recent paper from Cold Spring Harbor Laboratory, numerous AI and machine learning (ML) opportunities in medicine were outlined and discussed. Among those of note were methods to classify patients according to the tests that doctors ordered for them: “Perhaps deep neural networks, by reevaluating data without the context of our assumptions, can reveal novel classes of treatable conditions [18].” To do this, however, the AI must be carefully taught with data, much or all of it coming from EHRs, that are accurate and standardized. The term *garbage in, garbage out*, attributed most often to George Fuechsel [19], encapsulates the biggest issue with AI and ML. The outputs of the system will inevitably reflect the quality and biases of the data fed into it. At a recent AI hackathon for health outside Barcelona in 2018, the Medical Information Mart for Intensive Care (MIMIC) database was used. It is a freely accessible database that has associated more than 53,423 admissions at a large hospital in Boston from the years 2001 to 2012 [20]. The MIMIC, MIMIC II, and MIMIC III datasets have been used numerous times to demonstrate health analytics, and explorations of ML predicted patient outcomes. However, it was stated at the event that the data took nearly 2 years to *clean* and this length of time is not realistic if we have to use historical and real-time data to treat patients in the present. For MIMIC, creating an interoperable database was complex; standards on how to indicate fluid intake, data from critical care information systems, and data from the Social Security Administration all had to be developed, and this was within a single hospital system.

There is no formula for the exact sample size needed in ML, although the more complex the problem, the more data needed [21]; more complex questions such as disease treatment decisions carry a high level of risk; therefore, 100,000 could be considered a reasonable starting point. If the target is to analyze a specific rare disease, such as idiopathic pulmonary fibrosis, EHRs from multiple sources would have to be accumulated to reach that number, as the disease affects 18 of every 100,000 adults [22]. Extrapolating from these statistics, if we had complete EHRs from, for example, the entire state of Texas with a population of 28.7 million, we may get information on 5000 patients—which is not nearly enough. To effectively analyze health records from patients with rare diseases and to identify indicators within those populations, the ML would need to be able to *read* EHRs from across the United States (population 328,929,623 as on May 23, 2019 [23]) and ideally aggregate data from other countries as well.

We are at the beginning of a data-rich and connected century. To deliver optimal care to the millions of patients who are living longer with more complex and chronic diseases, we need to harness the fundamental technology of interoperable EHRs.

XSL•FO
RenderX

Obstacles

Cost

Health care expenditure for the entire country of Spain was more than €65 billion in 2015, according to a European Observatory on Health Systems and Policy report [24]. A US $1.2 billion expenditure (as occurred at Partners HealthCare) to integrate 1 hospital system is not feasible in countries with fewer resources, as is the case in many single-payer systems. Solutions that do not financially cripple a health care system need to be identified. There are now companies that are harnessing the immense value of health data and are willing to implement integration and analysis systems at no cost in return for access to data [25]. This is a new frontier, akin to the new marketplace of genomic data, which likely has its own set of benefits and repercussions, and it must be analyzed as to what the implications will be.

The cost, however, should be calculated by subtracting the added value of the benefits. With properly integrated EHRs, administrative costs can be lowered, adherence rates to care protocols have shown to improve [26,27], and many in-patient visits could be achieved with remote monitoring or telehealth services. In addition, ideally, the software could scan for coding errors, which are also costly in the United States (see the section Coding and Semantics).

Coding and Semantics

The main technical issue with arriving at interoperability is the huge variation in semantics and coding standards. Hospitals code their patients differently; in our hospital, we use a case number 7 digits long, but a neighboring hospital uses their own system with 6 digits. There is a unique identifier for each patient in the regional system, but many hospitals do not enter that data at all, and that regional identifier is not used at the national level.

The Logical Observation Identifiers Names and Codes (LOINC) is an international standard used by more than 78,000 agencies and health care institutions to code for health measurements and observations [28]. However, in everyday practice, these standards are not used in EHRs. Blood pressure, for instance, in the United States, will be documented as 120/80 mm Hg. LOINC states, “They should be reported as 2 separate variables, systolic (LOINC 8480-6) and diastolic (LOINC 8462-4) [29].” That same reading will be noted as 12/8 mm Hg in some European countries. In addition, blood sugar is annotated differently across borders. In the United Kingdom, blood sugar is annotated in millimoles per liter. A normal reading would be *under 7.8* [30]. In the United States, blood sugar is usually written using milligrams per deciliter; therefore, a normal reading is 70 to 130 mg/dL. These are simplistic examples to illustrate a fundamental concept.

There is also the conundrum of free text. Health care professionals annotate data and events in different ways. In our hospital, arterial hypertension may be listed in more than 4 ways: ht, HTA, hypertension, or hipertensió. In addition, text is used for describing symptom and disease evolution as well as test results. All these would have to be standardized or interpreted to make logical comparisons between charts. There has been an increased use of natural language processing to read free text in an EHR for disease phenotyping [31] and even detecting associations that led to adverse events [32], and it is likely this technology will be applied on a broader scale as it improves and becomes automated.

Human error is also a consideration. Within 1 country, there may be discrepancy among codes for diagnosis. When International Classification of Diseases (ICD), Tenth Revision, Clinical Modification (ICD-10-CM) replaced ICD-9 in the United States in 2015, the coding options increased 10-fold, from 14,400 to 144,000 [33]. The ICD-10-CM codes were linked to reimbursement for health care services, which made it all the more critical that codes be correct because mistakes could be taken as fraud. The Centers for Medicare and Medicaid Services released data indicating preventable billing errors had cost US $31.6 billion in 2018 [34]. However, it has been found that 1 ICD-9 code could be interpreted as 100 different ICD-10-CM codes, and not all of these codes seem logical: Y92.241, hurt at the library; W56.22, struck by Orca, initial encounter [33]. The United States is the only country that uses ICD-10-CM, creating yet another layer of incompatibility. If we had global compatibility of ICD coding, the statistics for global health would be far more accurate, which could, in theory, shift treatment protocols by allocating resources more precisely or seeing new trends in both communicable and noncommunicable diseases.

Privacy Issues

Privacy and security for health care data are of utmost importance. Effort must be made to educate health care administrations on how EHRs work, why they can be considered as safe as banking data, and what cybersecurity checks are in place and emphasize the importance of a continually updated security plan. Often, it is not a real security risk that needs to be addressed but the perception of risk [35]. Blockchain or other technologies should be analyzed for use, and more importantly, personnel who are equipped to detect and patch issues as well as develop solutions should be on staff.

Analysis of Progress

Over the years, there have been substantial efforts in the advocacy for EHR systems to integrate their internal sources of data as well as myriad external sources of patient information. Exemplary work from Mandl and Kohane in 2009 petitioned for EHRs and personally controlled health records to be *built on open standards*, *accommodating both open-source and closed-source software*, including data generated by a patient’s iPhone [36]. The authors advocated as well for federal support to clear the financial and taxonomic barriers to achieve this asking, “Can we produce a medication list for every American that can be obtained through standards-based, interoperable, substitutable applications?” The answer then was no, but open standards efforts are currently being developed and used internationally.

For instance, there is now RxNorm from the US National Library of Medicine, which can *mediate messages between systems not using the same software and vocabulary*, linking names of clinical drugs and drug interaction software [37].

XSL•FO
RenderX

Fast Health Interoperability Resources (FHIR) from HL7 is now widely recognized as the standard for EHR integration; it is used by Google for its Cloud Healthcare API stating that "FHIR specifies a robust, extensible data model for interacting with clinical resources" [38]. Analysis of data from Centers for Medicare & Medicaid Services and the Office of the National Coordinator for Health Information Technology in 2018 revealed that 32% of health information technology developers in the United States are using 2015 FHIR-certified standards, and the biggest EHR companies (including Epic and Cerner) are to some extent using FHIR standards [39]. Microsoft announced their Azure API for FHIR in February 2019 [40], and FHIR standards are also being used for the integration of wearables data and personalized devices [41]. SMART for FHIR is a project, which started in 2010 (FHIR was defined during the project) at Harvard Medical School and Boston Children's Hospital, aimed for medical applications to run without modifications across disparate health information systems. Mandel et al. demonstrated that within 2 months, a couple of software engineers could implement SMART on FHIR for 4 different EHR vendors [42].

In January 2018, Apple announced their version of a personalized EHR called HealthKit [43], which patients can access on their iPhone; it would appear Apple understands the value of providing a service that is user-friendly and that patients can monitor themselves and integrate data from fitness devices that connect to Apple.

Examples of Innovation in National Electronic Health Record Systems

Most countries in the world now use digital health records to some extent. The author of this paper (JS) collaborated with a team from the World Health Organization that was implementing a digital health information system in Sierra Leone in 2007, chosen precisely because it was a nearly entirely paper-based system and therefore a blank canvas. Since then much more infrastructure has been installed, and Sierra Leone has implemented district health information software from HISP in large hospital centers; in 2016, during the Ebola outbreak, a specialized EHR based on OpenMRS was developed for the Ebola treatment centers [44].

Rwanda began implementing OpenClinic electronic medical record in 2007, and it is now used throughout the country, in 20 hospitals and clinics [45]. In 2016, the Rwandan Ministry of Health partnered with Babylon Health, a company that now offers electronic prescriptions and telephone consultations to the now more than 2 million subscribers [46]. In addition, users can access their clinical records anytime via their phone, including images and audio and video of the consultations.

Estonia is another country where significant advances in health information technology innovation have been deployed at scale. The government launched an effort in 2016 to implement blockchain validation into the national EHR [47], the first country in the world to do so. The technology ensures data integrity and substantially reduces the risk of malicious intent or hacking because of blockchain's immutable data logs. This addresses the aforementioned issue of security, often cited by health care administrations when the question of electronic health data sharing is discussed.

In 2016, the Thai Health Information Standards Development Center published a plan for adopting national standards for patient health care summary, laboratory terminology (LOINC), syntax (HL7), and security (MICT) [48]. Thailand has already been notably forward-thinking by creating a unique national identifier system and achieving universal health care coverage in 2002 [49]; hence, the country is familiar with the effort it takes to align all the stakeholders involved in this type of initiative.

Israel has an integrated health monitoring system covering 4.2 million patients [50]. Since the implementation, studies have shown that patients are more adherent to medications [51].

In January 2019, Abu Dhabi launched a *unified health information exchange platform* called Malaffi, which allows approximately 2000 public and private health care providers across the Emirates to access and share information for approximately 3 million people [52]. This top-down approach is very effective when there are adequate funds to enact the process, but not possible in a country similar to the United States, where there is no single authority for a very disparate private health care system.

Belgium has coordinated an interoperable health record for all citizens, which came to full implementation in 2019, called MijnGezondheid [53]. Patient records can now be seen by any physician in any hospital in the country, not an easy feat when considering it includes all periphery hospitals, mental health institutions, pharmacies, and laboratory systems in 2 languages across 3 regions.

Viewpoint on Best Practices

All stakeholders within a health system can participate in shaping EHRs to be useful and evolved. The following are considerations for establishing best practices for effective and interoperable EHRs.

Standards

Adopt international standards such as FHIR, LOINC, and SNOMED CT and introduce these standards starting in medical school and university informatics classes. There should be International Standard Organizations standards required of any wearable that is integrated into an EHR so that physicians can be assured the data are reliable. For instance, a 6-m walking test may be performed by a patient at home and recorded for reference, but the results must be obtained by a device that has been proven to have accurate readings in a clinical setting. This is integral to the policy work on digital health regulation.

Education and Awareness

It is the responsibility of health care administrations to understand interoperability obstacles, the benefits of achieving this, and how it may be done. Investigation is required. In addition, a top-down approach is not the only effective means for adoption of interoperable EHRs. Citizen scientists are constantly developing their own hacks for integration of digital

XSL•FO
RenderX

health data, and indeed everyone, from patients to surgeons and from physiotherapists, nurses, to the billing office, should be involved or at least aware of the design process as it affects them all. Use the principle of user experience design and the way that all digital health platforms should be developed: know your user. A software developer may not intuit a cardiologist's needs (for instance, fast access to images and laboratory results) as opposed to a general practitioner (perhaps most important is an immediate view of history and medications); therefore, physicians, nurses, and administrators must be there to advise and do testing. Physicians can bring solutions that work to hospital administration, highlighting the benefits. On the other side, information technology professionals should be aware of how reimbursement works (among a myriad of other processes) and who needs to see what information when, including the entire arc of care from home caretakers to statisticians.

Ensure awareness of wearables and other sensor data and the fact that eventually patients will likely want this information to be incorporated into their EHR. The new companies developing EHR integration software must also be discerning of the quality and clinical validity of data being integrated.

Privacy

Hospital administration should request education on privacy and cybersecurity issues, and perhaps, ministries of health should offer short courses to strengthen their knowledge base. Ideally, hospital administration will feel comfortable in considering innovative solutions such as blockchain or in hiring the appropriate people who can, to ensure security and integrity of all patients' health data.

Hospital administration, ministries of health, and the general public should know how to access their data, how data are protected, and what the data can do for them. Perhaps there can be public service announcements on television, radio, and social media.

Alternative Solutions

Although the importance of interoperability seems to be a concept now recognized by the large EHR vendors, alternative and economically feasible solutions should be considered by health care administrations. There are solutions that do not require an entire retrofit of a hospital system to deliver data, which avoids the issue of interoperability altogether: Redox, which states it is Health Insurance Portability and Accountability Act compliant and secure, can intake HL7, FHIR, CDA, or X12 data, combine the data, and deliver an output [54].

Seqster is a company that officially entered the marketplace in 2018 and claims to be, "the only technology capable of enabling the majority of 350 million Americans to instantly connect to their EHR(s) along with major fitness/wearable devices, and consumer genetic labs" [55]. They have managed to aggregate and unify health information coming from Epic, Cerner, Strava, and even Fitbit.

There are likely many more companies that will appear in the marketplace as the value is increasingly recognized for having interoperable, clean, and accurate health records that can be data mined for life-saving decision making, research, and public health policy.

Author's Note

Since the writing of this article, Smart on FHIR has been implemented in over 100 Epic sites, and the trend is continuing.

Acknowledgments

The author would like to thank Bellvitge Hospital in Barcelona, Dr Maria Molina, Coordinator of the Interstitial Pulmonary Unit at Bellvitge Hospital and professor at the University of Barcelona, and the digital health community for all the work in this sphere.

Conflicts of Interest

The author consulted for the Digital Therapeutics Alliance, a nonprofit trade association for the digital therapeutics industry.

References

1. Sharma A, Harrington RA, McClellan MB, Turakhia MP, Eapen ZJ, Steinhubl S, et al. Using digital health technology to better generate evidence and deliver evidence-based care. J Am Coll Cardiol 2018 Jun 12;71(23):2680-2690 [FREE Full text] [doi: 10.1016/j.jacc.2018.03.523] [Medline: 29880129]
2. Kruse C, Pesek B, Anderson M, Brennan K, Comfort H. Telemonitoring to manage chronic obstructive pulmonary disease: systematic literature review. JMIR Med Inform 2019 Mar 20;7(1):e11496 [FREE Full text] [doi: 10.2196/11496] [Medline: 30892276]
3. Prahalad P, Tanenbaum M, Hood K, Maahs D. Diabetes technology: improving care, improving patient-reported outcomes and preventing complications in young people with type 1 diabetes. Diabet Med 2018 Apr;35(4):419-429. [doi: 10.1111/dme.13588] [Medline: 29356074]
4. Yang CY, Chen RJ, Chou WL, Lee YJ, Lo YS. An integrated influenza surveillance framework based on national influenza-like illness incidence and multiple hospital electronic medical records for early prediction of influenza epidemics: design and evaluation. J Med Internet Res 2019 Feb 1;21(2):e12341 [FREE Full text] [doi: 10.2196/12341] [Medline: 30707099]

XSL•FO
RenderX

5. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, et al. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health Aff (Millwood) 2005;24(5):1103-1117. [doi: 10.1377/hlthaff.24.5.1103] [Medline: 16162551]
6. Miriovsky BJ, Shulman LN, Abernethy AP. Importance of health information technology, electronic health records, and continuously aggregating data to comparative effectiveness research and learning health care. J Clin Oncol 2012 Dec 1;30(34):4243-4248. [doi: 10.1200/JCO.2012.42.8011] [Medline: 23071233]
7. Epic Systems. 2019. Epic: About URL: https://www.epic.com/about [accessed 2019-05-26]
8. Koppel R, Lehmann C. Implications of an emerging EHR monoculture for hospitals and healthcare systems. J Am Med Inform Assoc 2015 Mar;22(2):465-471. [doi: 10.1136/amiajnl-2014-003023] [Medline: 25342181]
9. McCluskey PD. The Boston Globe. 2016. Mass. General Launches Epic Health Records Upgrade URL: https://www.bostonglobe.com/business/2016/04/05/epic-upgrade-mass-general/9NIikFtLwWS8rysvZOxxyH/story.html [accessed 2018-10-30]
10. McCluskey PD. The Boston Globe. 2015. Partners' $1.2b Patient Data System Seen as Key to Future URL: https://www.bostonglobe.com/business/2015/05/31/partners-launches-billion-electronic-health-records-system/oo4nJJW2rQyfWUWQlvydkK/story.html [accessed 2019-05-27]
11. Vaidya A. Becker's Hospital Review. 2019. Partners Experiences Technical Issues, Including EHR Downtime URL: https://www.beckershospitalreview.com/ehrs/partners-experiences-technical-issues-including-ehr-downtime.html [accessed 2019-05-27]
12. Kumar RB, Goren ND, Stark DE, Wall DP, Longhurst CA. Automated integration of continuous glucose monitor data in the electronic health record using consumer technology. J Am Med Inform Assoc 2016 May;23(3):532-537 [FREE Full text] [doi: 10.1093/jamia/ocv206] [Medline: 27018263]
13. Arndt RZ. Modern Healthcare. 2018. For Epic, Interoperability Comes From Within URL: https://www.modernhealthcare.com/article/20180130/NEWS/180139993/for-epic-interoperability-comes-from-within [accessed 2018-10-01]
14. Zweig M. Rock Health. 2018. Q3 2018: An Entrepreneurs' Market Leads to Digital Health's Biggest Quarter Yet URL: https://rockhealth.com/reports/q3-2018-an-entrepreneurs-market-leads-to-digital-healths-biggest-quarter-yet/ [accessed 2018-10-30]
15. Forbes Magazine. 2019. AI And Healthcare: A Giant Opportunity URL: https://www.forbes.com/sites/insights-intelai/2019/02/11/ai-and-healthcare-a-giant-opportunity/ [accessed 2019-05-27]
16. Young RA, Burge SK, Kumar KA, Wilson JM, Ortiz DF. A time-motion study of primary care physicians' work in the electronic health record era. Fam Med 2018 Feb;50(2):91-99 [FREE Full text] [doi: 10.22454/FamMed.2018.184803] [Medline: 29432623]
17. Guell O. El País. 2018. Surgery Waiting Times in Spain: 93 Days, Slightly Down From 2017 URL: https://elpais.com/elpais/2018/11/29/inenglish/1543509744_110649.html [accessed 2019-05-23]
18. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018 Apr;15(141):20170387 [FREE Full text] [doi: 10.1098/rsif.2017.0387] [Medline: 29618526]
19. Lidwell W, Holden K, Butler J. Universal Principles of Design: 125 Ways to Enhance Usability, Influence Perception, Increase Appeal, Make Better Design Decisions, and Teach through Design. Beverly, MA: Rockport Publishers; 2010.
20. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
21. Brownlee J. Machine Learning Mastery. 2017. How Much Training Data is Required for Machine Learning? URL: https://machinelearningmastery.com/much-training-data-required-machine-learning/ [accessed 2019-05-29]
22. Raghu G, Chen SY, Hou Q, Yeh WS, Collard HR. Incidence and prevalence of idiopathic pulmonary fibrosis in US adults 18-64 years old. Eur Respir J 2016 Jul;48(1):179-186 [FREE Full text] [doi: 10.1183/13993003.01653-2015] [Medline: 27126689]
23. United States Census Bureau. 2019. US and World Population Clock URL: https://www.census.gov/popclock/ [accessed 2019-05-23]
24. Bernal-Delgado E, García-Armesto S, Oliva J, Sánchez-Martinez FI, Repullo JR, Peña-Longobardo LM, et al. WHO/Europe. 2018. Spain Health System Review 2018 URL: http://www.euro.who.int/__data/assets/pdf_file/0008/378620/hit-spain-eng.pdf?ua=1 [accessed 2019-05-24]
25. Stacey J. ACRP: Association of Clinical Research Professionals. 2017. Using EHR Data Extraction to Streamline the Clinical Trial Process URL: https://acrpnet.org/2017/04/01/using-ehr-data-extraction-streamline-clinical-trial-process/ [accessed 2019-05-28] [WebCite Cache ID 78hbcawTv]
26. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform 2016 May 20(Suppl 1):S48-S61 [FREE Full text] [doi: 10.15265/IYS-2016-s006] [Medline: 27199197]
27. Kazley AS, Simpson A, Simpson K, Teufel R. Association of electronic health records with cost savings in a national sample. Am J Manag Care 2014 Jun 1;20(6):e183-e190 [FREE Full text] [Medline: 25180501]
28. LOINC. URL: https://loinc.org/ [accessed 2019-05-27]

XSL•FO
RenderX

29. 55284-4 - LOINC Details. 2018. 55284-4 - Blood Pressure Systolic and Diastolic URL: https://r.details.loinc.org/LOINC/55284-4.html?sections=Comprehensive [accessed 2019-05-27]
30. Diabetes UK. 2019. Blood Sugar Level Ranges URL: https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html [accessed 2019-05-27]
31. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
32. Wong A, Plasek J, Montecalvo S, Zhou L. Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges. Pharmacotherapy 2018 Aug;38(8):822-841. [doi: 10.1002/phar.2151] [Medline: 29884988]
33. Manchikanti M, Kaye AD, Singh V, Boswell MV. The tragedy of the implementation of ICD-10-CM as ICD-10: is the cart before the horse or is there a tragic paradox of misinformation and ignorance? Pain Physician 2015;18(4):E485-E495 [FREE Full text] [Medline: 26218946]
34. Council for Medicare Integrity. 2018. Error Rate Drops, but Medicare Still Lost $31.6 Billion to Preventable Billing Errors in FY2018 URL: http://medicareintegrity.org/error-rate-drops-but-medicare-still-lost-31-6-billion-to-preventable-billing-errors-in-fy2018/ [accessed 2019-05-28]
35. Shahbaz M, Gao C, Zhai L, Shahzad F, Hu Y. Investigating the adoption of big data analytics in healthcare: the moderating role of resistance to change. J Big Data 2019 Jan 31;6(1):6 [FREE Full text] [doi: 10.1186/s40537-019-0170-y]
36. Mandl KD, Kohane IS. No small change for the health information economy. N Engl J Med 2009 Mar 26;360(13):1278-1281. [doi: 10.1056/NEJMp0900411] [Medline: 19321867]
37. National Library of Medicine - National Institutes of Health. 2019. Unified Medical Language System® (UMLS®): RxNorm URL: https://www.nlm.nih.gov/research/umls/rxnorm/ [accessed 2019-09-14]
38. Google Cloud. 2019. Cloud Healthcare API URL: https://cloud.google.com/healthcare/ [accessed 2019-09-14]
39. Posnack S, Barker W. HealthIT. 2018. Heat Wave: The US is Poised to Catch FHIR in 2019 URL: https://www.healthit.gov/buzz-blog/interoperability/heat-wave-the-u-s-is-poised-to-catch-fhir-in-2019 [accessed 2019-05-28]
40. Cartwright HJ. Microsoft Azure Cloud Computing Platform & Services. 2019. Lighting Up Healthcare Data With FHIR®: Announcing the Azure API for FHIR URL: https://azure.microsoft.com/en-us/blog/lighting-up-healthcare-data-with-fhir-announcing-the-azure-api-for-fhir/ [accessed 2019-05-28]
41. Saripalle RK. Leveraging FHIR to integrate activity data with electronic health record. Health Technol 2019 Apr 27:- [FREE Full text] [doi: 10.1007/s12553-019-00316-5]
42. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Sep;23(5):899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
43. Apple Newsroom. 2018. Apple Announces Effortless Solution Bringing Health Records to iPhone URL: https://www.apple.com/newsroom/2018/01/apple-announces-effortless-solution-bringing-health-records-to-iPhone/ [accessed 2019-05-27]
44. Oza S, Jazayeri D, Teich JM, Ball E, Nankubuge PA, Rwebembera J, et al. Development and deployment of the OpenMRS-Ebola electronic health record system for an Ebola treatment center in Sierra Leone. J Med Internet Res 2017 Aug 21;19(8):e294 [FREE Full text] [doi: 10.2196/jmir.7881] [Medline: 28827211]
45. Uwambaye P, Njunwa K, Nuhu A, Kumurenzi A, Isyagi M, Murererehe J, et al. Health care consumer's perception of the electronic medical record (EMR) system within a referral hospital in Kigali, Rwanda. RWJour 2017 May 24;4(1):48. [doi: 10.4314/rj.v4i1.7f]
46. Babyl Health Rwanda. 2018. URL: http://www.babyl.rw/ [accessed 2019-05-28]
47. Angraal S, Krumholz HM, Schulz WL. Blockchain technology: applications in health care. Circ Cardiovasc Qual Outcomes 2017 Sep;10(9):e003800. [doi: 10.1161/CIRCOUTCOMES.117.003800] [Medline: 28912202]
48. Kijsanayotin B, Benchakittipakorn P. Bureau of Thai Health Information System Standards Development. 2016. eHealth in Thailand: Interoperability and Health Information Standards URL: http://www.this.or.th/files/interopbook.pdf [accessed 2019-04-27]
49. Jongudomsuk P, Srithamrongsawat S, Patcharanarumol W, Limwattananon S, Pannarunothai S, Vapatanavong P, et al. Newborn and Birth Defects Database. 2015. The Kingdom of Thailand Health System Review URL: http://apps.searo.who.int/PDS_DOCS/B5410.pdf [accessed 2019-05-27]
50. Thorne M. Allscripts. 2017. Israel: A Case Study in National Connectivity for Better Health URL: https://www.allscripts.com/File%20Library/Case%20Studies/Clalit-Health-Services.pdf [accessed 2019-05-28]
51. Srulovici E, Garg V, Ghilai A, Feldman B, Hoshen M, Balicer RD, et al. Is patient support program participation associated with longer persistence and improved adherence among new users of adalimumab? A retrospective cohort study. Adv Ther 2018 May;35(5):655-665. [doi: 10.1007/s12325-018-0706-0] [Medline: 29748914]
52. Department of Health - Abu Dhabi. 2019. DoH Launches 'Abu Dhabi Health Information Exchange' to Integrate Technological Transformations in Healthcare URL: https://www.haad.ae/haad/tabid/58/ctl/Details/Mid/417/ItemID/808/Default.aspx [accessed 2019-05-29]
53. Mijngezondheid. URL: https://home.mijngezondheid.net/ [accessed 2019-09-17]

XSL•FO
RenderX

54. Redox. 2019. Revamp Legacy Infrastructure for a Digital World URL: https://www.redoxengine.com/provider-organizations/ [accessed 2019-05-28] [WebCite Cache ID 78hhXumSU]
55. Seqster | We Seek Clarity For Health Data. 2019. URL: https://seqster.com/ [accessed 2019-05-28]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
FHIR: Fast Health Interoperability Resources
ICD: International Classification of Diseases
LOINC: Logical Observation Identifiers Names and Codes
MIMIC: Medical Information Mart for Intensive Care
ML: machine learning

Edited by G Eysenbach; submitted 07.11.18; peer-reviewed by KH Yu, S Zheng, A Davoudi; comments to author 31.03.19; revised version received 29.05.19; accepted 02.09.19; published 01.11.19.

Please cite as:
Shull JG
Digital Health and the State of Interoperable Electronic Health Records
JMIR Med Inform 2019;7(4):e12712
URL: http://medinform.jmir.org/2019/4/e12712/
doi:10.2196/12712
PMID:31682583

©Jessica Germaine Germaine Shull. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 01.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Navigating Through Electronic Health Records: Survey Study on Medical Students' Perspectives in General and With Regard to a Specific Training

Anne Herrmann-Werner[1], MD; Martin Holderried[2], MD; Teresa Loda[1], MSc; Nisar Malek[3], MD; Stephan Zipfel[1], MD; Friederike Holderried[3], MD

[1]Department of Psychosomatic Medicine and Psychotherapy, Internal Medicine, University Hospital Tübingen, Tübingen, Germany

[2]Process and Quality Management, Department of Medical Structure, University Hospital Tübingen, Tübingen, Germany

[3]Department of Gastroenterology, Hepatology and Infectious Diseases, Internal Medicine, University Hospital Tübingen, Tübingen, Germany

Corresponding Author:
Teresa Loda, MSc
Department of Psychosomatic Medicine and Psychotherapy
Internal Medicine
University Hospital Tübingen
Osianderstr 5
Tübingen, 72076
Germany
Phone: 49 070712980129
Email: teresa.loda@med.uni-tuebingen.de

Abstract

Background: An electronic health record (EHR) is the state-of-the-art method for ensuring all data concerning a given patient are up to date for use by multidisciplinary hospital teams. Therefore, medical students need to be trained to use health information technologies within this environment from the early stages of their education.

Objective: As little is known about the effects of specific training within the medical curriculum, this study aimed to develop a course module and evaluate it to offer best practice teaching for today's students. Moreover, we looked at the acceptance of new technologies such as EHRs.

Methods: Fifth-year medical students (N=104) at the University of Tübingen took part in a standardized two-day training procedure about the advantages and risks of EHR use. After the training, students performed their own EHR entries on hypothetical patient cases in a safe practice environment. In addition, questionnaires—standardized and with open-ended questions—were administered to assess students' experiences with a new teaching module, a newly developed EHR simulator, the acceptance of the health technology, and their attitudes toward it before and after training.

Results: After the teaching, students rated the benefit of EHR training for medical knowledge significantly higher than before the session (mean 3.74, SD 1.05). However, they also had doubts about the long-term benefit of EHRs for multidisciplinary coworking after training (mean 1.96, SD 0.65). The special training with simulation software was rated as helpful for preparing students (88/102, 86.2%), but they still did not feel safe in all aspects of EHR.

Conclusions: A specific simulated training on using EHRs helped students improve their knowledge and become more aware of the risks and challenges of such a system. Overall, students welcomed the new training module and supported the integration of EHR teaching into the medical curriculum. Further studies are needed to optimize training modules and make use of long-term feedback opportunities a simulated system offers.

(JMIR Med Inform 2019;7(4):e12648) doi:10.2196/12648

KEYWORDS

medical students; electronic health records; eHealth; simulation

XSL•FO
RenderX

Introduction

Electronic health records (EHRs) comprise health information of a patient showing clinical data collected from all professionals involved in the patient's care, including nurses, doctors, therapists, laboratories, and external specialists [1].

Besides the immediate integration of a wealth of clinical data and examination results, implementing the usage of EHR provides numerous benefits, including increased adherence to guidelines in preventive care, decreased paperwork for providers, improvement in overall quality, efficiency of patient care [2], reduction of errors [3,4], enhanced monitoring of drug therapy [4], better daily workflow management [5], easy access of clinical data, legibility of notes, improved problem and medication lists, and better preventive care documentation. Challenges and risks, however, have been reported regarding heightened susceptibility to automation bias, decreased quality of notes because of copying and pasting [6,7], alert fatigue (desensitization) [8], disruption of the patient-physician relationship [9], mismatch of human and machine workflow models, and productivity loss potentially caused by EHR usability issues [2].

However, despite all the current knowledge of the benefits and risks of use of EHR and other technology, there has not been much research on the acceptance of new health technology systems such as EHR particularly among students. Students seem to be generally positive and more receptive to new technologies than more experienced health care providers [4,10]. The acceptance of health technology is mainly influenced by 2 underlying factors: the devices' *perceived ease of use* and *perceived usefulness* [11,12]. More perceived ease of use and higher usefulness might also underlie the findings of Tierney et al [10], with medical students as *digital natives* being closer to technology systems. However, medical students—despite their high exposure to and experience with electronic media—still need specific training in electronic health care systems as they rate their ability to use such clinical information systems as rather low [3,13-16]. The need for training is also mirrored by the fact that accreditation bodies and national catalogs of learning objectives expect medical graduates to be able to communicate clearly orally and in writing, including the documentation process in medical charts coining it a core competency [15,17-19]. However, so far, not enough clarity has emerged as to how and when such training in EHR usage should be integrated into the medical curriculum and which specific competencies should be reached [15,20-22]. In addition, Berndt and Fischer in their recent review [20] concluded that the growing use of EHR "for medical education, [...] poses many new challenges for teaching and learning (e.g., teaching of new data management skills; new roles and responsibilities for students and teachers) which have hardly been addressed." Previous studies have shown that training in the implementation process of EHR in general is useful [2], training in EHR has specifically improved communication when using the EHR [23], and training in the usage of EHR should already be in the focus of medical education fairly early on [9,15]. This also takes into account that most errors in EHR usage come down to issues concerning adequate training, well-prepared implementation, and the possibility of getting accustomed to the system [24,25].

Also requiring attention in line with these considerations are the technical, ethical, and legal points accompanying such training. Before digitalization, students could simply walk into the nurses' station and pick up the paper chart [15]. With EHR, the procedure is quite different as students now need individual login data, and unfortunately, a lot of medical schools deny their students permission to document EHR live, which lessens the potential benefit EHR can have in medical education and might lead to information loss within health care teams [26-28]. Despite the widespread usage of EHR in clinical practice in Germany and elsewhere, surveys show that medical students are often not allowed to make use of its full potential [3,26,29,30]. Simulated training environments offer a safe solution to this issue and are well accepted by students but are so far only rarely used [24,31-33]. However, clear rules of responsibility have to be defined when students are working with live EHR, particularly when considering the complicated general legal regulations in the European Union and the German system [34,35].

In summary, students need access to EHR to become knowledgeable and skilled in its use and to improve their understanding of system-based practice, because future medical practice environments will likely include the use of EHR. As students use EHR regardless of prior preparation, the need for training guidelines definitely exists [3,15,20]. Just as medical schools currently teach proper documentation as part of good clinical care in a paper-based world, they should be similarly obligated to teach students proper use of an EHR in an increasingly electronic world [26]. Atwater et al [9] concluded that "Best practices and strategies for teaching medical trainees in the setting of EHR have not been identified or widely shared with the medical education community." Thus, in this study, we aimed to develop a course module and evaluate it to offer a best practice teaching example.

Methods

Study Design and Participants

This longitudinal study took place at the Medical Faculty of the University of Tübingen in summer term 2018. A paper-based questionnaire was administered before and after the teaching session on EHR. Fifth-year medical students were recruited within their regular seminar in internal medicine. Participation in the EHR training was mandatory. However, participation in this study was on a voluntary basis. Out of 171 students, 116 (response rate 68.8%) participated in the study.

Test System

Teaching was conducted using a specially designed test system that exactly mirrored the EHR software program *Meona* (Meona GmbH, Freiburg, Germany) used in the clinical service at the University Hospital of Tübingen. It was created with 2 imaginary wards (internal medicine and surgery), allowing the virtual accommodation of up to 28 patients. Patient cases were developed by clinical experts in internal medicine before the teaching began. The cases were either simply created as plain

XSL•FO
RenderX

characters or entered with a full medical history and doctor's orders depending on the respective purpose. As there was no link to the actual EHR (Meona) version in clinical use, it provided a safe training environment without any implications for real patients. At the same time, however, the students were able to practice with a perfectly realistic copy of the original EHR system. The system was created and supported by the Information Technology (IT) Department of the University Hospital, which also maintains the actual clinical version. Once every 24 hours, it had to be updated after which one could either use the new blank version or upload the screenshot from before the update.

Teaching

The teaching course on EHR was held as a full-day intensive training over 2 consecutive days (6 hours per day). Before the actual teaching on day 1, students had to fill in the first questionnaire (T_0). Afterward, teaching started with a lecture on the general advantages, disadvantages, and pitfalls of EHR as well as a specific training on how to use the Meona system. As EHR count as a medical device in Germany demanding formalized training, part of the teaching was a standardized video on how to use the EHR system. This was followed by an interactive class including a lecture on how to perform a chart review and common medical errors to avoid. After lunchbreak, students were shown specific procedures within the EHR system (eg, tasks when admitting or discharging a patient) and had ample time to practice with a fictive patient, who was created as a new admission, with the student being asked to enter all the necessary information into the system and make orders accordingly. Day 1 ended with a wrap-up discussion exchanging experience using the EHR. On day 2, teaching started with a short refresher course on main points from the day before. Afterward, students were given specifically designed patient cases to perform a chart review. The cases covered typical patients seen in internal medicine (eg, complicated diabetes and gastrointestinal bleeding). Students first had to work on their own; this was followed by an interactive discussion including medical and technical issues. At the end of day 2, students filled in the second questionnaire (T_1).

Teaching was conducted by 2 experienced clinicians who each held a certificate as an official Meona instructor as well as a Master's Degree in Medical Education.

Questionnaire

We developed a questionnaire based on literature-derived common themes in EHR use and adapted from prior questionnaires in use [9,15,36]. The questionnaire had undergone cognitive pretesting using the method of *think aloud*, where the subject concurrently verbalizes thoughts when answering a questionnaire [37,38]. Consequently, minor adaptions to the questionnaire were made, and it was administered pre teaching (T_0) and post teaching (T_1) to allow for comparisons. The questionnaire can be obtained upon request. Students provided basic sociodemographic data (eg, age, gender, and semester), former training data, IT/electronic health–related data (eg, possession of devices and usage of the internet for health topics), and information regarding their prior experience with traditional chart reviews as well as EHR. In addition, they rated the general potential of EHR as well as the specific benefit for different professional groups (students, physicians, nurses, patients, and other professional groups) and their collaboration. Students also rated the teaching and the test system used. Table 1 provides an overview on the items used.

Table 1. Overview of outcomes and their corresponding measurement of the questionnaire.

Outcome	Item	Number of items
Sociodemographics	Gender, age, and response rate	3
Previous experience with electronic devices (eg, mobile phones, personal computer, and laptop)	Yes/no	6
Previous experience with EHR[a] (participation, contribution, and contact)	Yes/no	5
Benefit for different professions	Likert scale from 0 to 5 ("not at all" to "completely")	6
Concerns and inhibitions	Likert scale from 0 to 3 ("not at all" to "completely")	3
Evaluation of the test system	Likert scale from 0 to 3 ("not at all" to "completely")	6
Evaluation of the teaching module	Likert scale from 0 to 3 ("not at all" to "completely")	5
Students' experiences with EHR	Likert scale from 0 to 3 ("not at all" to "completely")	6

[a]EHR: electronic health record.

Data Analysis

Data analysis was performed using SPSS version 24. For statistical analysis, frequencies, means, and associated SDs were calculated for different items of the questionnaire. Data were normally distributed as tested by the Kolmogorov-Smirnov test. *T* tests for 2-paired samples were conducted to allow comparisons of pre teaching and post teaching. For further comparison, analyses of variance were conducted. Here, the level of significance was $P<.05$. For the comparison of pre teaching and post teaching, data were included only when the students filled in both questionnaires. Furthermore, we considered the cumulative frequencies in percentages for several items such as prior usage of EHR. Here, questionnaires of all 116 students taking part in the study were included, and frequencies were calculated proportionately for each item. At the end of the study, 104 out of 116 students had returned the

XSL•FO
RenderX

complete pre- and postquestionnaires and could be included in the analyses of comparisons. Again, on the singe-item level, frequencies were calculated proportionately. The absolute numbers might differ slightly from 116 or 104 students because of missing data.

Ethics

The Ethics Committee of Tübingen Medical Faculty (#260/2016BO2) approved this study.

Results

Sociodemographics

A total of 116 students participated in the study, and 104 students returned the completed pre- and postquestionnaires and showed up for both appointments of the study. Moreover, 59 (56.7) students were female and 45 (43.3) were male. Their mean age was 25.6 (SD 3.0) years.

Previous Experience With Electronic Devices

Nearly all the students (103/113, 91.1%) had a mobile phone, 111 (98.2%) had a personal computer with internet connection, and 79 (69.9%) owned a tablet. Out of 112 students, 76 (67.8%) stated owning all 3 devices, and 108 (96.4%) students rated the internet as *rather important* or *important* for their daily lives. Students checked their private emails every day, which was significantly more often than their professional ones ($F_{1,4}$=38.04; P<.001).

Previous Experience With Electronic Health Records

Out of 104 students, 67 (64.4%) had already participated in a chart review in general (paper or EHR). However, out of these, only 18 (27%) students had actively contributed to one. Mostly, the chart review was part of their mandatory clinical placements. In addition, 66 out of 101 (65.3%) students already had contact with an EHR system, with proportionally the largest group (36/47, 77% students) having watched someone else using it. Finally, 99 out of 103 students (96.1%) had thus far no formal training in EHR.

Benefit for Different Professions

The students' judgment of the relative benefit of EHR for medical professionals did not vary significantly between T_0 and T_1 regardless of the group (see Table 2). In addition, students rated the benefit of EHR significantly higher for doctors and nurses than for any other professions both before and after training (pre teaching—benefit doctors, mean 4.11; nurses, mean 3.90; therapists, mean 3.67; patients, mean 3.10; medical students, mean 3.55; P<.001 for doctors and nurses compared with all other professions—and post teaching—benefit doctors, mean 3.96; nurses, mean 3.82; therapists, mean 3.68; patients, mean 3.29; medical students, mean 3.68; P<.001 for doctors and P=.03 for nurses compared with all other professions). Analyzed in detail, students rated the benefit of EHR for their medical knowledge significantly higher after the teaching session (Table 2).

Table 2. Ratings of benefits, concerns, and inhibitions of electronic health record.

Item	Pre teaching[a], mean (SD)	Post teaching[a], mean (SD)	T_0–T_1 comparison	
			t test (*df*)	*P* value
Benefit for doctors	4.11 (0.88)	3.96 (0.93)	−0.84 (100)	.40
Benefit for nursing stuff	3.90 (0.98)	3.82 (0.99)	0.92 (99)	.36
Benefit for physiotherapist or speech therapist	3.67 (1.05)	3.68 (0.95)	−0.12 (97)	.90
Benefit for patients	3.10 (1.31)	3.29 (1.32)	−1.52 (97)	.13
Benefit for students	3.55 (1.18)	3.68 (1.08)	−1.16 (97)	.25
Benefit for students' medical knowledge	2.96 (1.27)	3.29 (1.11)	−2.86 (98)	.005
General concerns	0.63 (0.89)	0.69 (0.86)	−0.58 (101)	.56
Inhibitions	0.40 (0.76)	0.38 (0.77)	0.26 (100)	.80
Potential as a collaboration tool	3.18 (0.87)	3.01 (0.88)	1.99 (99)	.049

[a]Agreement ("0" = "not at all" to "5" = "completely").

Concerns and Inhibitions

There was no significant difference before and after training regarding concerns and inhibitions related to EHR use. However, students evaluated EHR's potential long-term benefit as a collaboration tool in the multiprofessional health care team to be significantly lower at T_1 compared with T_0. Table 2 provides further details.

Evaluation of the Test System

The most frequently mentioned positive aspects were the protected and safe environment (29/58 students, 50%) in which to practice as well as the general benefits of an EHR system such as drug interaction warnings. EHR needs a substantial amount of training with proper facilities (16/70 students, 23%) and the fear that other hospitals might have different systems (2/70, 3%) for which they would then not be prepared were some of the critical issues mentioned by the students. In addition, students pointed out that the training system still had some technical difficulties (23/70, 33%; eg, no immediate connection to current treatment guidelines and inappropriate date of birth of the created patients). When presented with a list of areas where support in the future would be needed most, issues concerning active processes such as *change the patient's*

XSL•FO
RenderX

medication (17/104 students, 16%) and *confident navigation through the system* (28/104 students, 27%) were among the most frequent answers.

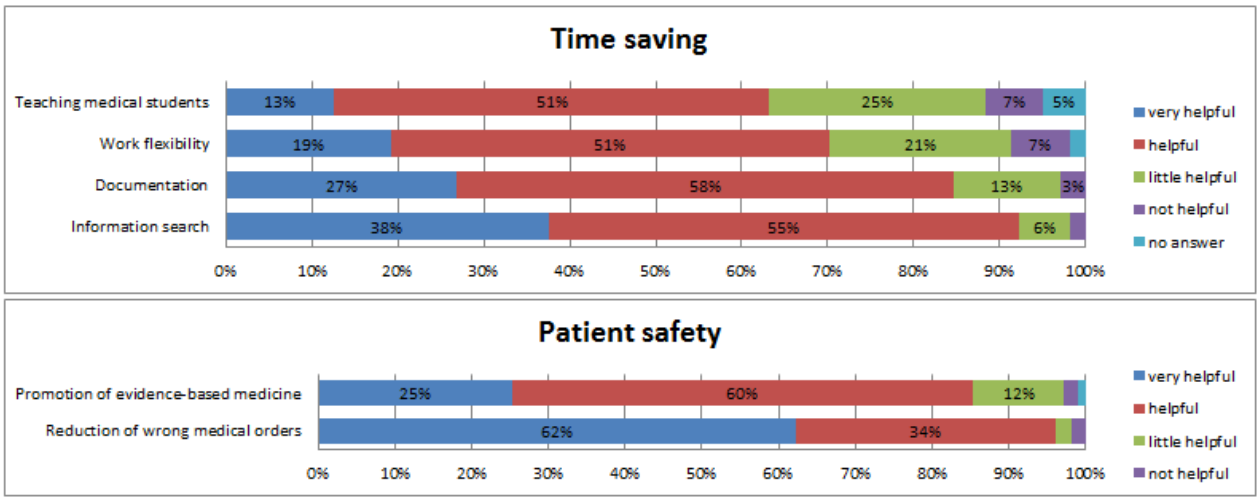
Evaluation of the Teaching Module

Out of 102 students, 88 (86.3%) stated that the teaching prepared them in a *rather good* or *good* way for later usage of EHR. When asked for which area they felt best prepared specifically (eg, navigation, patient admission, placing orders, and changing medication), there was a significant difference among the subthemes ($F_{6,64}$=3.59; P=.002), with students feeling best prepared for reading and understanding the current medication scheme (mean 2.25, SD 0.66) and worst prepared for navigation through the EHR (mean 1.96, SD 0.65). Out of 103 students, 44 (42.7%) would have liked to have the teaching video in a web-based version as well, with another 24 students (23.3%) agreeing that it might be helpful to have such an additional option. However, they also generally appreciated the presence of a real teacher, as 74 out of 103 students (71.8%) stated that a web-based teaching program alone would *not* or *possibly not* be sufficient to reach the desired competencies and understanding.

After the teaching, the students felt rather motivated to work with EHR in the future (mean 3.74, SD 1.05) and considered EHR as a useful tool in clinical practice (mean 3.7, SD 0.04). Looking closer at different aspects of time saving and patient safety when using EHR, EHRs were mostly considered as a *very helpful* or *helpful* tool in later work in hospitals by our students. For the advantages and details they identified, see Figure 1.

In addition, 82 out of 104 students (78.8%) considered the system's offer of templates (eg, normal findings on physical examination and electrocardiography) *helpful* or *very helpful*. Reasons for finding it only *somewhat helpful* or not helpful were *producing data waste*, *limitation of expressions*, and *standard formulations are known by heart anyway*. Accordingly, 98 out of 104 students (94.2%) considered the integrated support system of EHR (eg, immediate warnings about drug interactions) *helpful* or *very helpful*. The 6 students considering the system *not helpful* or *only somewhat helpful* were most critical on the following points: *giving a false sense of security* (n=3), *danger of not thinking critically on one's own* (n=1), and *limited flexibility through forced adherence to guidelines* (n=1). However, the problem of alert fatigue, as mentioned in the Introduction, was not reported by our students. Our medical students seemed unsure about how to judge the potential problem of *copy and paste*—also one of the main risks and pitfalls of EHR—rating it mean 2.3 (SD 1.5) on the abovementioned scale. Of the 48 students being more reserved toward the copy and paste possibility, the majority mentioned worries along the issue of *blind take-over of information/no cross-checking/no reflection on (potentially wrong) diagnoses* (n=41). Only 3 students were concerned about the potential issue of *loss of quality/reduced doctor-patient interaction*.

Figure 1. Percentage of students (N=104) who found the electronic health record very helpful, helpful, little helpful, and not helpful, respectively, with regard to time saving and patient safety aspects. Values less than 3% are not marked on the graph to improve readability.



Discussion

Summary

This study looked at undergraduate medical students' perspectives toward EHR in general and with a specific focus on a particularly designed teaching module. Teaching included a formalized introduction to EHR accounting for legal demands as well as plenty of contextual training applying medical knowledge in the electronic system.

Perspectives on Electronic Health Records

In general, the students in this study had a positive attitude toward EHR usage, which is generally in line with previous findings [4,10]. Interestingly, participants rated the benefit of EHR usage significantly higher for doctors and nurses than for other health care professionals or their patients at both measurements. This might be because of the fact that medical students still mostly see those 2 professions in action of patient care. Interestingly, students in this study rated the potential benefits for coworking in a multiprofessional team significantly lower after their teaching sessions. This seems counterintuitive

XSL•FO
RenderX

as common access to medical charts should foster team collaboration, although caveats have been described in literature with current systems yet lagging the full potential [39-41]. The students' reservations might be explained by the fact that through specific training, they become aware not only of the advantages but also of the shortcomings of EHR usage, as already mentioned in the Introduction, enabling them to evaluate clinical information systems more critically. In addition, they have not been using them in real clinical practice. Therefore, they can only imagine and anticipate or remember complex interactions they have been observing in clinical internships where frequent difficulties and problems with EHR are discussed more prominently than effectively working examples in team interactions. The possible influence of EHR on the patient-physician relationship was not an issue for our students. This is in line with literature showing effective patient-centered interactions despite usage of EHR in the encounters [42,43].

Although clinical decision aids integrated into EHR offer great learning opportunities, there is a danger of *alert fatigue* in users [8,44]. Students in this study did not show such concerns. We assume that this phenomenon might not be prominent in students, who have not been using the system frequently so far but is more of an issue for experienced system users working with EHR on a daily basis. However, it seems crucial to create an awareness of this issue early on. This also accounts for possible negative implications of the *copy&paste* phenomenon. Our medical students were unsure how to rate this issue, but those concerned named well-known reasons in the clinical context [6,7]. So far, literature has not shown any negative educational consequences of copy and paste (eg, impaired critical thinking and reduced self-directed learning). However, there is certainly the need for more standardized examinations on that matter [44-46].

Teaching

In this study, students accepted the new technology well and felt highly motivated to use EHR. They all represented the generation of digital natives—as reflected in their possession and usage of technical devices and the internet and students thus might be especially receptive to the use of new technologies [10]. However, this might not be transferable to an adequate professional use, and despite being digital natives, students do need specific training in technical devices in the health care context [3,13,14,16]. As documentation is a core competency that graduates should show from day 1 of their clinical work, the need for specific training in the usage of EHR is thus undisputable [15,17,19]. There are even demands of whole longitudinal curricula on this issue [15]. This enables several levels of reality: starting with theoretical input in the early years and proceeding to simulated scenarios as well as a structured integration of live EHR use in clinical placements. However, reality does look different: students usually are not officially allowed to document in EHR or sometimes do so without proper training in the systems beforehand [3,26,29,30]. The students in this study also reported mostly just having watched someone using an EHR. However, some students had documented on their own but without training, which poses a legal problem in Germany as EHR count as a medical device is not allowed to be used without a formalized introductory teaching beforehand. Directors of medical schools should be aware of this potentially dangerous issue.

Although the students evaluated their training course positively, it does not seem to have been thorough enough, as students still did not feel safe when navigating through the EHR afterward. This uncertainty might have resulted from students focusing on the medical information and casework, prioritizing this part of the task over organizational and structural learning objectives of this class. This was reflected in the in-class discussions where students' questions mainly concerned medical issues, and EHR seemed to be merely a means for that purpose. A lack of EHR navigating skills is also what Morrow et al [23] discuss when finding that after training, students had significantly better communication skills within the EHR tool but did not show satisfying navigation skills such as finding previous data or creating trend graphs. It may be necessary to separate medical content from technical information or at least specifically stress the importance of structural skills [20].

Simulated Electronic Health Record System

When looking at students' experiences with the new EHR software program (Meona) in general, the feedback was positive. Overall, the students appreciated the features they would also encounter in the live version, although they were also aware of the technical difficulties still present in the newly developed copy of the actual Meona. We want to draw attention to some of these, as we consider this as helpful for other medical schools planning to develop an EHR simulator. When creating a teaching version of an in-use EHR, it is important to keep in mind that the system needs constant updating. In our case, this meant reinstalling the initial version to delete the entries of one student group before the next one works with the program. However, this means that when you have admitted the patient in April and constantly back up to this initial version, the students who have their training class in June are supposed to work with patients who have been on the ward for 2 months with nothing having been done up to this point. In the whole process, it is also crucial to involve IT [47]. This accounts for making them familiar with the content of your teaching before they start to program the virtual patients. It does not foster the degree of felt reality when students work on 19-year-old patients who have been in and out of hospital for the past 20 years because of their poorly controlled diabetes. When creating such a system, it is also important to predefine who will be responsible for tasks, who has the administrator's rights, and when the program is to be updated or reloaded so that you have a secured environment [47]. During our first term of teaching EHR, not having clarified all these issues, we more than once had to manually reenter all patient data because IT made an update without a screenshot first. In addition, in the beginning, we were unable to change minor issues ourselves as we did not have the rights to do so. Thus, it might be helpful to have key users with limited administrative rights who can customize the system accordingly, such as only updating it during semester break.

Simulated systems are created to prepare for reality. There is ample literature regarding the rights of medical students in live EHR systems [3,48]. When using real systems, one has to find the balance between allowing students to be part of the team

XSL•FO
RenderX

with the same duties and ownership as other team members on the one hand [3], whereas, on the other hand, taking into account legal issues of responsibility that might exceed a student's capability level and will need to be reviewed [49,50]. By choosing a mirrored version of the actual EHR system used in our hospital, all students in the class automatically got the training necessary to be allowed to work with the live electronic chart. As a consequence of this teaching, the Medical Faculty of the University of Tübingen together with the Quality Management Department of the University Hospital defined and implemented those rights for all students in their final year to ensure quality of care and reproducibility of the clinical documentation within the EHR system. One key element of this process was to show students' entries color coded as *preliminary documentation* that has to be checked and confirmed by a fully trained physician before release, although this has been shown to be a source of concern among deans of medical schools [30,45]. Thanks to such provisions, students could start using EHR immediately the day they entered their final practical year without endangering patient safety.

Limitations

The study has several limitations. First, we only looked at medical students at one semester and one faculty, which limits generalizability. In addition, the training class was relatively short, being an intensive course over 2 days; thus, some of the results might be not representative enough. Finally, we did not look at transfer into the clinical environment, thus not being able to say if the students' self-ratings would hold up in the actual context of use.

Despite these limitations, we strongly believe that our study delivers valuable insight into aspects of consideration when planning and implementing a teaching class on EHR into the medical curriculum.

Conclusions

Overall, the class showed several advantages, and the training was regarded as helpful. However, it might have been more helpful to separate medical content from the technical aspects to reduce cognitive overload or have at least more teaching time longitudinally, as already practiced in some medical schools [20,51]. Future development could include assigning person-specific logins to track individual progress. In addition, the potential of interprofessional as well as nationwide or even worldwide web-based learning opportunities should be considered [52].

Conflicts of Interest

None declared.

References

1. Garrett NY, Mishra N, Nichols B, Staes CJ, Akin C, Safran C. Characterization of public health alerts and their suitability for alerting in electronic health record systems. J Public Health Manag Pract 2011;17(1):77-83. [doi: 10.1097/PHH.0b013e3181ddcbc0] [Medline: 21135665]
2. Clarke MA, Belden JL, Kim MS. How does learnability of primary care resident physicians increase after seven months of using an electronic health record? A longitudinal study. JMIR Hum Factors 2016 Feb 15;3(1):e9 [FREE Full text] [doi: 10.2196/humanfactors.4601] [Medline: 27025237]
3. Hammoud MM, Margo K, Christner JG, Fisher J, Fischer SH, Pangaro LN. Opportunities and challenges in integrating electronic health records into undergraduate medical education: a national survey of clerkship directors. Teach Learn Med 2012;24(3):219-224. [doi: 10.1080/10401334.2012.692267] [Medline: 22775785]
4. Rouf E, Chumley HS, Dobbie AE. Electronic health records in outpatient clinics: perspectives of third year medical students. BMC Med Educ 2008 Mar 31;8:13 [FREE Full text] [doi: 10.1186/1472-6920-8-13] [Medline: 18373880]
5. Keenan CR, Nguyen HH, Srinivasan M. Electronic medical records and their impact on resident and medical student education. Acad Psychiatry 2006;30(6):522-527. [doi: 10.1176/appi.ap.30.6.522] [Medline: 17139024]
6. O'Donnell HC, Kaushal R, Barrón Y, Callahan MA, Adelman RD, Siegler EL. Physicians' attitudes towards copy and pasting in electronic note writing. J Gen Intern Med 2009 Jan;24(1):63-68 [FREE Full text] [doi: 10.1007/s11606-008-0843-2] [Medline: 18998191]
7. Thornton JD, Schold JD, Venkateshaiah L, Lander B. Prevalence of copied information by attendings and residents in critical care progress notes. Crit Care Med 2013 Feb;41(2):382-388 [FREE Full text] [doi: 10.1097/CCM.0b013e3182711a1c] [Medline: 23263617]
8. Campbell EM, Sittig DF, Ash JS, Guappone KP, Dykstra RH. Types of unintended consequences related to computerized provider order entry. J Am Med Inform Assoc 2006;13(5):547-556 [FREE Full text] [doi: 10.1197/jamia.M2042] [Medline: 16799128]
9. Atwater AR, Rudd M, Brown A, Wiener JS, Benjamin R, Lee WR, et al. Developing teaching strategies in the EHR era: a survey of GME experts. J Grad Med Educ 2016 Oct;8(4):581-586 [FREE Full text] [doi: 10.4300/JGME-D-15-00788.1] [Medline: 27777671]
10. Tierney WM, Overhage JM, McDonald CJ, Wolinsky FD. Medical students' and housestaff's opinions of computerized order-writing. Acad Med 1994 May;69(5):386-389. [doi: 10.1097/00001888-199405000-00013] [Medline: 8166922]

XSL•FO
RenderX

11. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. Manag Inf Syst Q 1989;13(3):319-340. [doi: 10.2307/249008]
12. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manag Sci 2000;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
13. Krause N, Roulette GD, Papp KK, Kaelber D. Assessing medical informatics confidence among 1st and 2nd year medical students. AMIA Annu Symp Proc 2006:989 [FREE Full text] [Medline: 17238608]
14. Borycki E, Griffith J, Reid P, Kushniruk AW, Kuo MH. Do electronic health records help undergraduate health informatics students to develop health informatics competencies? Stud Health Technol Inform 2013;192:1106. [doi: 10.3233/978-1-61499-432-9-838] [Medline: 23920880]
15. Hammoud MM, Dalymple JL, Christner JG, Stewart RA, Fisher J, Margo K, et al. Medical student documentation in electronic health records: a collaborative statement from the Alliance for Clinical Education. Teach Learn Med 2012;24(3):257-266. [doi: 10.1080/10401334.2012.692284] [Medline: 22775791]
16. Kirschner PA, De Bruyckere P. The myths of the digital native and the multitasker. Teach Teach Educ 2017;67:135-142. [doi: 10.1016/j.tate.2017.06.001]
17. Association of American Medical Colleges. Learning objectives for medical student education--guidelines for medical schools: report I of the Medical School Objectives Project. Acad Med 1999;74(1):461-462. [doi: 10.1097/00001888-199901000-00010] [Medline: 9934288]
18. Fischer MR, Bauer D, Mohn K, NKLM-Projektgruppe. Finally finished! National Competence Based Catalogues of Learning Objectives for Undergraduate Medical Education (NKLM) and Dental Education (NKLZ) ready for trial. GMS Z Med Ausbild 2015;32(3):Doc35 [FREE Full text] [doi: 10.3205/zma000977] [Medline: 26677513]
19. Association of American Medical Colleges. The Core Entrustable Professional Activities (EPAs) for Entering Residency URL: https://www.aamc.org/initiatives/coreepas/publicationsandpresentations/426410/publicationshometsr.html [accessed 2019-06-01]
20. Berndt M, Fischer MR. The role of electronic health records in clinical reasoning. Ann N Y Acad Sci 2018 Dec;1434(1):109-114. [doi: 10.1111/nyas.13849] [Medline: 29766520]
21. Pageler NM, Friedman CP, Longhurst CA. Refocusing medical education in the EMR era. J Am Med Assoc 2013 Dec 4;310(21):2249-2250. [doi: 10.1001/jama.2013.282326] [Medline: 24302083]
22. Wald HS, George P, Reis SP, Taylor JS. Electronic health record training in undergraduate medical education: bridging theory to practice with curricula for empowering patient- and relationship-centered care in the computerized setting. Acad Med 2014 Mar;89(3):380-386 [FREE Full text] [doi: 10.1097/ACM.0000000000000131] [Medline: 24448045]
23. Morrow JB, Dobbie AE, Jenkins C, Long R, Mihalic A, Wagner J. First-year medical students can demonstrate EHR-specific communication skills: a control-group study. Fam Med 2009 Jan;41(1):28-33 [FREE Full text] [Medline: 19132569]
24. Ludwick DA, Doucette J. Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries. Int J Med Inform 2009 Jan;78(1):22-31. [doi: 10.1016/j.ijmedinf.2008.06.005] [Medline: 18644745]
25. Dornan T, Boshuizen H, King N, Scherpbier A. Experience-based learning: a model linking the processes and outcomes of medical students' workplace learning. Med Educ 2007 Jan;41(1):84-91. [doi: 10.1111/j.1365-2929.2006.02652.x] [Medline: 17209896]
26. Mintz M, Narvarte HJ, O'Brien KE, Papp KK, Thomas M, Durning SJ. Use of electronic medical records by physicians and students in academic internal medicine settings. Acad Med 2009 Dec;84(12):1698-1704. [doi: 10.1097/ACM.0b013e3181bf9d45] [Medline: 19940575]
27. Schenarts PJ, Schenarts KD. Educational impact of the electronic medical record. J Surg Educ 2012;69(1):105-112. [doi: 10.1016/j.jsurg.2011.10.008] [Medline: 22208841]
28. Solarte I, Könings KD. Discrepancies between perceptions of students and deans regarding the consequences of restricting students' use of electronic medical records on quality of medical education. BMC Med Educ 2017 Mar 13;17(1):55 [FREE Full text] [doi: 10.1186/s12909-017-0887-2] [Medline: 28288618]
29. Welcher CM, Hersh W, Takesue B, Stagg Elliott V, Hawkins RE. Barriers to medical students' electronic health record access can impede their preparedness for practice. Acad Med 2018 Jan;93(1):48-53. [doi: 10.1097/ACM.0000000000001829] [Medline: 28746069]
30. Friedman E, Sainte M, Fallar R. Taking note of the perceived value and impact of medical student chart documentation on education and patient care. Acad Med 2010 Sep;85(9):1440-1444. [doi: 10.1097/ACM.0b013e3181eac1e0] [Medline: 20736671]
31. March CA, Steiger D, Scholl G, Mohan V, Hersh WR, Gold JA. Use of simulation to assess electronic health record safety in the intensive care unit: a pilot study. BMJ Open 2013;3(4):pii: e002549 [FREE Full text] [doi: 10.1136/bmjopen-2013-002549] [Medline: 23578685]
32. Joe RS, Otto A, Borycki E. Designing an electronic medical case simulator for health professional education. Know Manag E-Learn 2011;3(1):63-71. [doi: 10.34105/j.kmel.2011.03.007]

XSL•FO
RenderX

33. Issenberg SB, McGaghie WC, Hart IR, Mayer JW, Felner JM, Petrusa ER, et al. Simulation technology for health care professional skills training and assessment. J Am Med Assoc 1999 Sep 1;282(9):861-866. [doi: 10.1001/jama.282.9.861] [Medline: 10478693]
34. Rienhoff OLC, van Ecke P, Wenzlaff P, Piccolo U. A legal framework for security in European health care telematics. Stud Health Technol Inform 2000;74:1-192. [Medline: 11153481]
35. van der Haak M, Wolff A, Brandner R, Drings P, Wannenmacher M, Wetter T. Data security and protection in cross-institutional electronic patient records. Int J Med Inform 2003 Jul;70(2-3):117-130. [doi: 10.1016/s1386-5056(03)00033-9] [Medline: 12909163]
36. Yu P, Qian S. Developing a theoretical model and questionnaire survey instrument to measure the success of electronic health records in residential aged care. PLoS One 2018;13(1):e0190749 [FREE Full text] [doi: 10.1371/journal.pone.0190749] [Medline: 29315323]
37. van Someren MW, Barnard YF, Sanberg JA. The Think Aloud Method: A Practical Guide to Modelling Cognitive Processes. London: Academic Press; 1994.
38. Ericsson KA, Simon HA. Verbal reports as data. Psychol Rev 1980;87(3):215-251. [doi: 10.1037//0033-295x.87.3.215]
39. Lingard L. Paradoxical truths and persistent myths: reframing the team competence conversation. J Contin Educ Health Prof 2016;36(Suppl 1):S19-S21. [doi: 10.1097/CEH.0000000000000078] [Medline: 27584064]
40. Rudin RS, Bates DW. Let the left hand know what the right is doing: a vision for care coordination and electronic health records. J Am Med Inform Assoc 2014;21(1):13-16 [FREE Full text] [doi: 10.1136/amiajnl-2013-001737] [Medline: 23785099]
41. O'Malley AS, Draper K, Gourevitch R, Cross DA, Scholle SH. Electronic health records and support for primary care teamwork. J Am Med Inform Assoc 2015 Mar;22(2):426-434 [FREE Full text] [doi: 10.1093/jamia/ocu029] [Medline: 25627278]
42. Duke P, Frankel RM, Reis S. How to integrate the electronic health record and patient-centered communication into the medical visit: a skills-based approach. Teach Learn Med 2013;25(4):358-365. [doi: 10.1080/10401334.2013.827981] [Medline: 24112206]
43. Silverman H, Ho Y, Kaib S, Ellis WD, Moffitt MP, Chen Q, et al. A novel approach to supporting relationship-centered care through electronic health record ergonomic training in preclerkship medical education. Acad Med 2014 Sep;89(9):1230-1234 [FREE Full text] [doi: 10.1097/ACM.0000000000000297] [Medline: 24826851]
44. Tierney MJ, Pageler NM, Kahana M, Pantaleoni JL, Longhurst CA. Medical education in the electronic medical record (EMR) era: benefits, challenges, and future directions. Acad Med 2013 Jun;88(6):748-752. [doi: 10.1097/ACM.0b013e3182905ceb] [Medline: 23619078]
45. Knight AM, Kravet SJ, Harper GM, Leff B. The effect of computerized provider order entry on medical student clerkship experiences. J Am Med Inform Assoc 2005;12(5):554-560 [FREE Full text] [doi: 10.1197/jamia.M1839] [Medline: 15905479]
46. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012;19(1):121-127 [FREE Full text] [doi: 10.1136/amiajnl-2011-000089] [Medline: 21685142]
47. Milano CE, Hardman JA, Plesiu A, Rdesinski RE, Biagioli FE. Simulated electronic health record (Sim-EHR) curriculum: teaching EHR skills and use of the EHR for disease management and prevention. Acad Med 2014 Mar;89(3):399-403 [FREE Full text] [doi: 10.1097/ACM.0000000000000149] [Medline: 24448035]
48. Foster LM, Cuddy MM, Swanson DB, Holtzman KZ, Hammoud MM, Wallach PM. Medical student use of electronic and paper health records during inpatient clinical clerkships: results of a national longitudinal study. Acad Med 2018 Nov;93(11S Association of American Medical Colleges Learn Serve Lead: Proceedings of the 57th Annual Research in Medical Education Sessions):S14-S20. [doi: 10.1097/ACM.0000000000002376] [Medline: 30365425]
49. Kirch DG. DocPlayer. 2014. Allowing Medical Student Documentation in the Electronic Health Record URL: https://docplayer.net/663664-Allowing-medical-student-documentation-in-the-electronic-health-record-background-and-purpose.html [accessed 2019-10-10]
50. Chi J, Kugler J, Chu IM, Loftus PD, Evans KH, Oskotsky T, et al. Medical students and the electronic health record: 'an epic use of time'. Am J Med 2014 Sep;127(9):891-895. [doi: 10.1016/j.amjmed.2014.05.027] [Medline: 24907594]
51. Daniel M, George PF, Warrier S, Dodd K, Dollase R, Taylor JS. Warren Alpert Medical School's Doctoring program: a comprehensive, integrated clinical curriculum. Med Heal 2012;95(10):313-316 [FREE Full text]
52. Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. N Engl J Med 2011 Nov 10;365(19):1758-1759. [doi: 10.1056/NEJMp1108726] [Medline: 22047518]

Abbreviations

EHR: electronic health record
IT: information technology

XSL•FO
RenderX

Edited by G Eysenbach; submitted 30.10.18; peer-reviewed by J Graf, N Haller, M Kim, R Alkoudmani; comments to author 05.04.19; revised version received 14.06.19; accepted 19.08.19; published 12.11.19.

Please cite as:
Herrmann-Werner A, Holderried M, Loda T, Malek N, Zipfel S, Holderried F
Navigating Through Electronic Health Records: Survey Study on Medical Students' Perspectives in General and With Regard to a Specific Training
JMIR Med Inform 2019;7(4):e12648
URL: http://medinform.jmir.org/2019/4/e12648/
doi:10.2196/12648
PMID:31714247

©Anne Herrmann-Werner, Martin Holderried, Teresa Loda, Nisar Malek, Stephan Zipfel, Friederike Holderried. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 12.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Primary Care Physicians' Experience Using Advanced Electronic Medical Record Features to Support Chronic Disease Prevention and Management: Qualitative Study

Rana Melissa Rahal[1], BSc, MSc; Jay Mercer[2], MD, CCFP, FCFP; Craig Kuziemsky[3], BSc, BCom, PhD; Sanni Yaya[4,5], MSc, PhD

[1]Population Health Program, University of Ottawa, Ottawa, ON, Canada

[2]Bruyère Continuing Care, Ottawa, ON, Canada

[3]Office of Research Services, MacEwan University, Edmonton, AB, Canada

[4]School of International Development and Global Studies, University of Ottawa, Ottawa, ON, Canada

[5]The George Institute for Global Health, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Rana Melissa Rahal, BSc, MSc
Population Health Program
University of Ottawa
25 University Private
Ottawa, ON, K1N 7K4
Canada
Phone: 1 6138902193
Email: rraha039@uottawa.ca

Abstract

Background: Chronic diseases are the leading cause of death worldwide. In Canada, more than half of all health care spending is used for managing chronic diseases. Although studies have shown that the use of advanced features of electronic medical record (EMR) systems improves the quality of chronic disease prevention and management (CDPM), a 2012 international survey found that Canadian physicians were the least likely to use 2 or more EMR system functions. Some studies show that maturity vis-à-vis clinicians' EMR use is an important factor when evaluating the use of advanced features of health information systems. The Clinical Adoption Framework (CAF), a common evaluation framework used to assess the success of EMR adoption, does not incorporate the process of maturing. Nevertheless, the CAF and studies that discuss the barriers to and facilitators of the adoption of EMR systems can be the basis for exploring the use of advanced EMR features.

Objective: This study aimed to explore the factors that primary care physicians in Ontario identified as influencing their use of advanced EMR features to support CDPM and to extend the CAF to include primary care physicians' perceptions of how their use of EMRs for performing clinical tasks has matured.

Methods: Guided by the CAF, directed content analysis was used to explore the barriers and facilitating factors encountered by primary care physicians when using EMR features. Participants were primary care physicians in Ontario, Canada, who use EMRs. Data were coded using categories from the CAF.

Results: A total of 9 face-to-face interviews were conducted from January 2017 to July 2017. Dimensions from the CAF emerged from the data, and one new dimension was derived: physicians' perception of their maturity of EMR use. Primary care physicians identified the following key factors that impacted their use of advanced EMR features: performance of EMR features, information quality of EMR features, training and technical support, user satisfaction, provider's productivity, personal characteristics and roles, cost benefits of EMR features, EMR systems infrastructure, funding, and government leadership.

Conclusions: The CAF was extended to include physicians' perceptions of how their use of EMR systems had matured. Most participants agreed that their use of EMR systems for performing clinical tasks had evolved since their adoption of the system and that certain system features facilitated their care for patients with chronic diseases. However, several barriers were identified and should be addressed to further enhance primary care physicians' use of advanced EMR features to support CDPM.

(JMIR Med Inform 2019;7(4):e13318) doi:10.2196/13318

XSL•FO
RenderX

KEYWORDS

electronic health record; chronic disease; primary health care; medical informatics

Introduction

Background

According to the World Health Organization, by 2020, chronic diseases will account for 73% of all deaths and 60% of the global burden of disease [1]. The World Health Organization recommends that chronic disease prevention must focus on controlling risk factors such as high blood pressure and tobacco use [1].

Electronic medical records (EMRs) are one of many initiatives available in high-income countries to assist in addressing these risk factors. In a systematic review, approximately 67% of studies showed that EMRs have a positive effect on preventive care, and about 57% of studies found that EMRs contribute to a modest improvement in disease management [2].

Electronic reminder features for preventive or follow-up care automate reminders for specific tests (eg, vaccinations and blood tests) based on recommended guidelines [3]. Advanced EMR features, such as electronic reminders, have been shown to support chronic disease prevention and management (CDPM). When EMR reminders were combined with access to EMR information (eg, history of hypertension and cardiovascular disease), 28% of the patient population was found to be at risk for undiagnosed type 2 diabetes [4].

A grounded theory study of EMR usage ranked EMR features from basic to advanced [5]. Advanced features included automated reminders for tests and screening; using decision support tools, such as a cardiovascular risk tool; using a recall system to search for patients with a specific condition; creating customized templates, such as diabetic flow sheets; and using a graph feature to view the trend of a patient's test results over time [5].

Statement of the Problem

Not all physicians use the advanced features of EMR systems to support CDPM. A 2012 study showed that Canadian physicians were the least likely to use at least two EMR functions [6]. Thus, there is a gap in our understanding of the barriers to and facilitating factors of the use of advanced features in EMR systems.

Factors That Impact the Adoption of Electronic Medical Records

Much of the literature has focused on the factors that contribute to successful EMR adoption. Studies have discussed the need for EMR champions and staff participation to encourage adoption [7-9]. Rogers' diffusion of innovations theory suggests that the characteristics of potential adopters are also a key factor for EMR adoption [10].

In addition, studies have identified the importance of providing adequate education and training to support EMR adoption [11,12]. In the Canadian province of Ontario, the Association of Family Health Teams developed a program comprising individuals known as quality improvement decision support specialists (QIDSS) who were available on-site to assist teams to access and better use EMR data to improve care [13].

Furthermore, some studies have highlighted the importance of advancing the level of health information system (HIS) use to obtain improved clinical outcomes and have suggested that benefits grow over time as users gain experience, as improvements are made in systems, and as workflows are adjusted to users' needs [14,15]. A Canadian study in Ontario assessed the progress in the use of advanced EMR features and found a direct correlation between years of EMR use and EMR maturity [14]. Thus, in evaluating the use of advanced features of EMR systems, it is important to consider how the use of EMR systems by clinicians has evolved since EMR adoption.

Conceptual Framework

In this study, the Clinical Adoption Framework (CAF) [16] was used to categorize the study results and to explore the barriers and facilitators that primary care physicians encounter when using EMR features to support CDPM. Although the CAF does not evaluate the maturity of a clinician's HIS use, the framework is appropriate for this study as it identifies microlevel, mesolevel, and macrolevel factors that influence EMR success.

Several frameworks for HIS adoption have been reported in the literature [16-21]. OntarioMD, a cooperative owned by the Ontario Medical Association and funded by the provincial government, is responsible for certifying EMRs in Ontario [22]. OntarioMD developed the EMR Maturity Model [21] to help clinicians optimize their EMR use by evaluating their level of EMR use. The model evaluates maturity in terms of how the product is used, and users can measure their maturity level for a certain function (eg, appointment scheduling and laboratory results) across 6 maturity levels (see Multimedia Appendix 1) [21]. Thus, this study refers to maturity as the maturity of the user's skill set and clinical processes in using the HIS, rather than the maturity of a product (ie, type of features implemented). The EMR Maturity Model is based on existing models such as the CAF.

The CAF (shown in Multimedia Appendix 2) proposes that successful clinical adoption of HISs at the microlevel depends on the following dimensions: the quality of the system's performance, information, and support service provided for the HIS; its use and user satisfaction; and net benefits. At the mesolevel, the people involved, the organization, and the implementation of the HIS have a direct effect on the microlevel HIS adoption by health care professionals. At the macrolevel, successful clinical adoption depends on health care standards; funding and incentives; legislation/policy and governance; and societal, political, and economic trends. A detailed description of the dimensions for each level can be found in previous studies [16,17,23].

Purpose of the Study

This study explored the barriers primary care physicians encounter while using advanced EMR features to facilitate

XSL•FO
RenderX

CDPM and the factors facilitating their use of these features. Furthermore, this study extends the CAF to include primary care physicians' perceptions of how their use of the EMR system had evolved. Thus, the main contribution of this study was looking at the CAF and the maturity of EMR use from the perspective of primary care providers, as they are the ones managing chronic illness.

Methods

Study Setting and Design

On the basis of existing evidence about factors influencing EMR adoption, a qualitative directed content analysis was conducted using the CAF. A directed content analysis is typically used when existing theory or prior research about a phenomenon needs further description to validate or extend a theoretical framework or theory [24]. Thus, we used directed content analysis to extend the CAF.

The study was conducted at primary care clinics located in the Canadian province of Ontario. Although there are various EMR systems available in Ontario, the most common systems used at primary care clinics are PS Suite EMR (produced by Telus Health) [25], Nightingale On Demand (produced by Telus Health) [26], IndiviCare (produced by Indivica) [27], and OSCAR (produced by OSCAR EMR Inc) [28]. Advanced EMR features available in these systems include but are not limited to the following:

- Drug databases that provide dosing information, administration, and medication allergy alerts.
- Hospital Report Manager [29], an Ontario provincial feature used to electronically integrate patient reports (eg, medical records and diagnostic imaging reports) from hospitals and specialty clinics directly into a patient's chart.
- Ontario Laboratories Information System (OLIS) that automatically receives laboratory results from hospitals directly into the patient's chart [30].
- Electronic fax to electronically receive faxed documents into EMRs.

Study Participants, Sampling, and Recruitment

Eligible participants were primary care physicians located in Ontario who had used EMRs for at least one year. Purposeful sampling was used to represent a range of ages (less than 30 years, 30-40 years, 41-50 years, 51-60 years, 61-70 years, and greater than 71 years), sexes (female and male), and individuals from different cities in Ontario. Face-to-face interviews were conducted.

Data saturation determined the sample size. After 7 interviews, no new ideas were being introduced. Nevertheless, 2 more interviews were conducted to validate that saturation had occurred. A similar study exploring primary care physicians' experience with EMRs also had a sample size of 9 participants [31].

OntarioMD assisted in recruiting participants by sharing an advertisement about this study with its peer leaders. Similarly, Ontario academic family practices were contacted to identify participants, resulting in the Ottawa Hospital Family Health Team reaching out to its members. Recruitment emails were also sent to individual family practices.

Data Collection and Research Instruments

Data were actively collected between January 2017 and July 2017 by the primary author (RR). In-person interviews were audio recorded. Interviews were approximately 20 min to 60 min and were conducted by using a semistructured interview guide (Multimedia Appendix 3). The interview guide was pilot-tested in July 2016 with a primary care physician.

Data Analysis

Audio recordings of interviews were transcribed verbatim. The directed content approach using the CAF helped determine the initial coding scheme [16,17]. Each interview transcript was read line by line; any text that appeared to describe a barrier or facilitating factor was highlighted (RR). Next, NVivo software (QSR International) [32] was used to help code all highlighted text using predetermined codes (RR). Data that could not be coded into one of the categories of the CAF were coded with a label that captured the essence of the barrier or facilitating factor. Finally, 2 team members (RR and SY) independently analyzed transcripts, and 3 team members (RR, SY, and CK) audited the data analysis findings.

Ethical Considerations

The University of Ottawa Research Ethics Board (H01-16-02) granted approval for the study. All participants provided written informed consent before their interview; no personal information was recorded.

Results

Participant Characteristics

Table 1 summarizes the sample and participant characteristics. All participants' practices were located in an urban setting in Ontario. Participants' experience in using an EMR system ranged from 3 to 15 years. Overall, 5 of the participants were part of a group practice using the family health organization's capitated payment model, 3 of the participants were from a family health team (FHT) practice model, and 1 participant was from an independent practice. In addition, 5 of the participants identified themselves as the information technology (IT) leader in their clinic. A total of 4 participants used the EMR system PS Suite, 3 used IndiviCare, and 1 worked with Nightingale On Demand.

Patterns from the data were categorized into themes. In this study, themes refer to barriers and facilitating factors that influenced participants' use of advanced EMR features. A total of 10 themes emerged from the data: 9 themes directly mapped to the dimensions of the CAF and one new theme was derived from our analysis. The dimensions from the framework that directly mapped to the 9 themes were system quality; information quality; service quality; user satisfaction; net benefits; people; organization; legislation, policy, and governance; and funding and incentives. Figure 1 shows the dimensions from the CAF that emerged from the data and the one new dimension (maturity of EMR use) that was derived from our analysis.

XSL•FO
RenderX

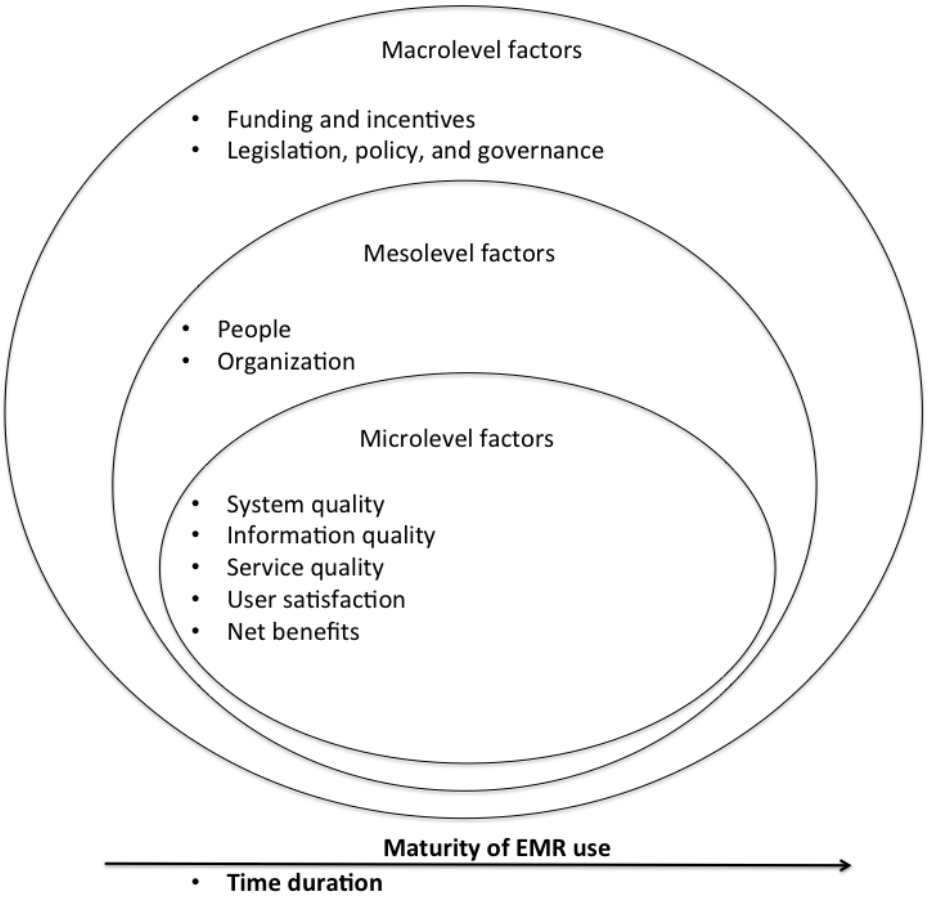
Table 1. Respondents' characteristics.

Participants	Age range (years)	Sex	Primary care model	Experience using electronic medical records (years)	Information technology lead
P1	51-60	Male	FHT[a]	15	Yes
P2	61-70	Female	Independent practice	3	No
P3	61-70	Male	FHT	10	Yes
P4	41-50	Female	FHO[b]	7	Yes
P5	30-40	Male	FHO	7	Yes
P6	51-60	Male	FHO	15	Yes
P7	30-40	Female	FHT	4	No
P8	41-50	Male	FHO	4	No
P9	61-70	Female	FHO	9	No

[a]FHT: family health team.

[b]FHO: family health organization.

Figure 1. Dimensions emerging from the data. EMR: electronic medical record.



Theme 1: System Performance (Microlevel)

The CAF defines the dimension, *system quality*, as the reliability of the system's performance, features, and security and is estimated in terms of performance and reliability, based on the system's response times for standardized tasks, integration with workflow, user-friendliness, and security [16].

Several participants explained that the quick response time for standardized tasks was a system performance factor that facilitated their use of advanced EMR features:

> *When I receive an abnormal test result I get it right away and I don't need to wait for the next day.* [P2, age 61-70 years]

However, 2 participants mentioned that the drug database feature was not user friendly. Owing to the limitations of this feature,

XSL•FO
RenderX

participants used mobile or Web-based drug database applications that were not part of the EMR software as they had an easier interface and quicker response time:

> *It's so confusing…but I can write the same thing in my app…it's just easier to read and it's quicker.* [P4, age 41-50 years]

Participants also described system reliability as a barrier to using advanced EMR features (eg, EMR feature not working).

Theme 2: Completeness of Information (Microlevel)

The dimension *information quality* is defined as the completeness and accuracy of information in addition to the timeliness and relevance of information [16]. Another facilitating factor is the completeness and relevance of information provided by the EMR drug database feature:

> *The system is more sophisticated than the last time…it will show me the various dosage forms…that are available.* [P1, age 51-60 years]

A few participants were concerned about the completeness and relevance of information provided by the EMR graph feature. These limited their ability to plot and view the trend of a patient's test results:

> *It's a terrible graph…because it's not temporally organized…so it's useless as a graph. (P3, age 61-70 years)*

Theme 3: User Training and Technical Support (Microlevel)

The dimension *service quality* is defined as user training and ongoing technical support and availability of support [16]. Participants were asked if they had an IT specialist on-site to support the EMR system. A total of 6 participants raised the issue of vendors' insufficient ongoing technical support to enhance clinic performance, limited ongoing training for advanced EMR use, and ineffective user training. Technical support was not available on-site unless it was paid out of pocket or if a staff member communicated with the vendor.

Theme 4: Perceived Usefulness of Electronic Medical Record Features, Perceived Impact on Productivity, and Perceived Impact on Quality of Care (Microlevel)

The CAF cites *user satisfaction* as one category that measures the dimension *satisfaction*, defined as the subjective opinions of users with regard to their perceived expectations; value; information, system, and service quality; and use of the system. Lau et al [16] assessed the framework's user satisfaction component using indicators of perceived usefulness and value of the system, perceived impact on productivity and integration with workflow, and perceived impact on quality of care [16].

According to several participants, certain EMR features (eg, recall system and diabetic flow sheets) were useful and improved their quality of care, for example:

> *If there's a drug recall, you can find all the patients who are on that drug and call…them to come in. So it's amazing what you could do which you couldn't do on a paper chart.* [P4, age 41-50 years]

Overall, 2 participants stated that using the EMR feature to assess cardiovascular risk was time consuming and inefficient, thus impacting productivity and preventing them from using this advanced ready-made feature. One participant described the use of the cardiovascular risk feature as challenging, in that it was not fully integrated into their EMR system, necessitating the use of other online tools to calculate risk:

> *Anything that's inefficient is dangerous because it creates a barrier for people to do it. It promotes transcription errors. You move the data manually, you're going to type a key wrong.* [P6, age 51-60 years]

Theme 5: Change in Provider Efficiency, Net Cost, and Care Quality (Microlevel)

The CAF portrays *net benefits* as quality, access, and productivity. The framework assesses quality using indicators such as changes in provider effectiveness and appropriateness of care, whereas productivity is measured by indicators of change in provider efficiency, such as the time needed to assess a patient and clinician workflow [16]. The framework also refers to productivity as the change in net costs in terms of cost savings [16].

Participants reported improved workflow efficiency and improved patient efficiency when certain advanced EMR features were used. One participant described how workflow efficiency and patient efficiency were enhanced when they used a customized referral letter template to expedite a specialist referral: "So when I see an abnormal result I can send a referral at that time and its more efficient for me" [P2, age 61-70 years].

Overall, 2 participants suggested that change in productivity was a barrier to their use of advanced EMR features because of the additional cost associated with the EMR system, particularly maintaining, supporting, and upgrading the system to ensure effectiveness and efficiency. Other associated costs included after-sales support from vendors and hiring additional staff to deal with paper documents that were not electronically deposited into the EMR:

> *Since the EMR, we had to hire one person whose job was just to scan stuff in before the e-fax came.…I'm paying someone a full-time job just to scan, which is out of my pocket, which is created because of this technology.* [P4, age 41-50 years]

Furthermore, the quality of provider effectiveness and appropriateness of care were adversely affected when participants could not access patients' test results from hospitals, in the EMR system. Participants mentioned wasting time searching for unavailable laboratory results instead of using that time for other tasks.

Theme 6: Roles and Personal Characteristics (Mesolevel)

The CAF defines the dimension *people* as the individuals or groups involved, their personal characteristics and expectations, and their roles and responsibilities vis-à-vis the HIS [17].

XSL•FO
RenderX

The framework uses an individual's age, gender, experience, and position (eg, being an IT leader) to measure personal characteristics and roles [17]. One participant with over 10 years of EMR experience, who was also the IT leader, described how they exploited the system:

> *I am too far into using EMRs.…I just do what EMR permits.…I really exploit the system.* [P1, age 51-60 years]

On the contrary, another participant (P2) with 3 years of experience using an EMR system revealed that they train their patients to remember when to do blood tests rather than use the reminder feature to prompt the physician for patient preventive services. Clearly, the participants' characteristics and roles impacted their use of advanced EMR features.

Theme 7: Return on Value and Infrastructure (Mesolevel)

The CAF categorizes *organization* as how the HIS fits with the organization's strategy, culture, and structure or processes, as well as information, infrastructure, and return on value [17]. The framework defines return on value of HIS adoption in terms of cost benefit and effectiveness. Infrastructure is measured in terms of technical architectures, level of integration, and the privacy or security in place or planned [17].

Only a few participants stated that the return on value of advanced EMR features was a barrier to the use of these features. One participant said that the electronic fax feature was expensive and not reliable, so their clinic continued to use a paper-based process:

> *And that's a problem with the software. They have an Internet faxing version, but they charge a fortune for it…and it has problems with capacity and reliability.* [P6, age 51-60 years]

Most participants noted that their inability to directly transfer documents among the EMR system and hospitals and pharmacies was a barrier. The majority of participants reported that they received laboratory results directly into their EMR system from private laboratories. However, most hospital results are faxed, scanned, and added to the patient's chart, which was another barrier. The OLIS feature facilitates searching for missing laboratory results. However, some participants mentioned that not all hospital laboratory results were available in OLIS. If they were, the amount of paper that clinics received from hospitals would decrease:

> *If I go to [the patient's] chart, I will see if their lab results are actually available through the EMR's access to OLIS.…If I can do that, then I don't need all that printed paper.* [P3, age 61-70 years]

Theme 8: Governance and Privacy Laws (Macrolevel)

Some participants were concerned about the lack of leadership in addressing poor EMR infrastructure, namely, lack of direct links with hospitals and pharmacies. According to one participant:

> *The fact that we can't get stuff from hospital…There's no technical problem. There's no leadership that puts together the infrastructure and secures it to do it the way it's supposed to be done. That's all we're missing, leadership…the government can fix two things. One, they could tell the people who supply the software whom they certify, that they have to provide turnkey end-to-end service. And number two, the government actually can help create the connectivity between us and the pharmacies, us and the hospitals.* [P6, age 51-60 years]

Furthermore, 2 participants were concerned about the security and privacy of patient charts because of legislation allowing the Ontario government to access patient data.

Theme 9: Funding (Macrolevel)

A total of 2 other participants noted that they did not receive enough government funding to cover all the EMR system expenses. As one participant said:

> *[The program] didn't cover everything but it was great, but then they stopped that…then this ongoing and maintaining, it's all out of our pockets.* [P4, age 41-50 years]

Theme 10: Maturity of Electronic Medical Record Use

Participants were directly asked how their use of EMRs for performing clinical tasks had evolved since adoption. The CAF does not have a category to account for the different maturity stages of the user, so a new category was developed. The CAF describes factors that impact the success of EMR adoption at a moment in time, whereas the new theme describes how these factors evolve over time.

Overall, 2 participants stated that their use of EMRs for performing clinical tasks had not evolved effectively since adoption. They noted flaws such as technical errors with the laboratory requisition feature; poor feature design for prescribing medication doses; and excessive scanner use because of the inability to electronically transfer documents among the EMR and some hospitals and pharmacies, which was needed to support continuity of care over time. Such flaws limited these participants from using the system to its maximum capacity. As one participant explained:

> *There's way too much paper handling. Why is a person sitting at a scanner all day long? Why are we still waiting?* [P6, age 51-60 years]

However, most participants agreed that their use of the EMR system to perform clinical tasks had improved since its adoption. Several participants revealed the importance of using certain advanced EMR features (eg, electronic fax and Hospital Report Manager) to facilitate patient care delivery and reduce paper work. As one participant said:

> *We get features that now allow us to run almost a paperless office that did not exist when we first started* [P5, age 30-40 years]

As such, the use of advanced features to facilitate patient care delivery and reduce paper work demonstrates that these physicians' use of the EMR system is maturing as they are able to incorporate advanced EMR features into their workflow.

XSL•FO
RenderX

Furthermore, using the electronic fax and Hospital Report Manager is considered advanced EMR use as physicians have incorporated these features into their clinical process as a way to facilitate CDPM. These features allow physicians to electronically access patient's results and limit the need to scan paper documents into the EMR, thereby reducing the wait time of physicians accessing patient's results. Thus, these features can improve patient care by decreasing the wait time during an appointment as the physician searches for the patient's results or the possibility of human error when scanning paper documents into the EMR, such as support staff mismatching scanned results to a patient's chart.

Theme 10 shows the need to have a temporal dimension to EMR evaluation to see what types of emerging issues will arise over time. The CAF looks at a more generic set of adoption factors, whereas theme 10 highlights the need to identify specific factors that facilitate EMR use that will emerge over time.

Discussion

This study explores primary care physicians' use of EMR systems to support CDPM. Most participants highlighted factors that facilitated their use of advanced EMR features. However, participants continue to experience barriers.

Principal Findings and Comparison With Prior Work

Microlevel Factors

Most participants mentioned that system quality and information quality factors, such as quick response time for standardized tasks (eg, receiving blood test results), and the feature's provision of complete and relevant information facilitated their use of advanced EMR features. However, participants reported unreliability as a barrier (eg, EMR feature not working), and a few participants also found the drug database feature to be non–user friendly.

Studies have recommended involving users in system design to address such technical factors [2,31,33]. As suggested in one study, professional associations, such as OntarioMD, could influence vendors by imposing standards and publishing specifications so that EMR features would be designed to benefit physicians [5].

Several participants noted that insufficient technical support and inadequate user training on the part of the vendor was a barrier. In addition, lack of on-site technical support from the vendor created additional costs such as hiring staff to address technical issues. A program such as QIDSS [13] could help address this barrier by helping physicians make better use of EMR data to improve clinical performance.

User satisfaction emerged from the data in terms of participants' perceived usefulness of an EMR feature as well as its perceived impact on both productivity and quality of care. Although several participants noted that EMR features (eg, recall system and diabetic flow sheets) supported their quality of patient care, for others, certain EMR features (eg, data entry and cardiovascular risk feature) were inefficient and time consuming, thus a barrier to their productivity.

A systematic review recommended discussing the usefulness of a given EMR feature, demonstrating its ease of use, and having fellow physicians demonstrate the feature [34]. OntarioMD's Peer Leader program is a network of clinicians with several years of EMR experience. These individuals support practices in Ontario to advance their EMR use [35]. Such a program can help address the user satisfaction barriers identified in our study.

Mesolevel Factors

According to our findings, participants who were IT leaders and had more EMR experience were more likely than others to exploit the EMR system. These findings are consistent with the diffusion of innovations theory, which describes how characteristics of potential adopters (eg, expertise and perception of innovation) influence the success of innovation adoption [10]. Furthermore, a commonly cited infrastructure barrier was the inability to directly transfer documents among the EMR system and hospitals and pharmacies. This barrier has also been identified in other studies [5,36].

Macrolevel Factors

Lack of leadership in addressing poor interoperability among EMR systems and hospitals and pharmacies is an important macrolevel factor discussed by a few participants. A grounded theory study conducted in Ontario also noted the lack of connectivity among clinical EMRs and hospital laboratories [5]. The study recommended that OntarioMD could influence software development via standards and publishing future requirements and through financial support to improve the interoperability among EMR systems and other health care entities [5].

Legislation and funding also emerged as issues in the data. Some participants were uneasy regarding the security and privacy of patient charts because of legislation that allows the Ontario government to access patient data. Other studies have also shown that concerns about privacy and security of patient data are a barrier to EMR use because of the potential legal problems [34,37,38].

In addition, participants who were not part of an FHT practice felt that government funding was not sufficient to cover EMR expenses. These findings confirm those of other studies in which barriers related to insufficient funding influenced the adoption and use of EMRs [2,5,39].

Maturity of Electronic Medical Record Use

Most participants thought that their use of EMR systems had improved since adoption with the support of advanced EMR features (eg, electronic fax and Hospital Report Manager). Studies that assessed clinicians' use of EMR systems found that longer EMR use led to improved outcomes (eg, greater expertise and improved patient care) [14,15]. Some of the key factors explored in this study could be measured over time to assess the different maturity stages of physicians' use of advanced EMR features.

Key factors such as reliability, functionality, and user-friendliness of the EMR feature; technical support and user training; user satisfaction; productivity; return on value; and

XSL•FO
RenderX

infrastructure could be assessed as part of the mature use of an EMR system either quantitatively using surveys or qualitatively through interviews. One possible method would be ranking the progress of each key factor for each advanced feature and the progress of mature use of these advance features. For example, for the advanced feature OLIS, its reliability, functionality, and user-friendliness could be ranked using a Likert scale that ranges from 0 to 5, where 0 indicates that the user strongly disagrees that OLIS is reliable, functional, and user friendly. Similarly, the progress of mature use can be assessed using a 5-point Likert scale, where 0 shows that the user strongly disagrees that the feature is fully integrated within their clinical workflow (eg, feature is not being used) and 5 implies that the user strongly agrees that the feature is fully integrated within their clinical workflow (eg, feature is used to access patient's current and past test results to enable treatment decisions and, if applicable, results are shared with the patient at the point of care). A longitudinal analysis of a clinic would need to be done to measure the progress of these key factors over time and the progress of mature use of these advanced EMR features. Thus, the maturity of EMR use dimension extends the CAF by incorporating postadoption factors perceived by physicians to influence their use of advanced features and the effects of these factors over time to reflect the different maturity stages of the user.

An application of this extended CAF would be to evaluate the progress of advanced EMR feature use among primary care physicians. Another would be for physicians to identify potential factors within their practice that influence their use of advanced EMR features in reaching maturity and to make recommendations for improvements.

Furthermore, the extended CAF could be used by key stakeholders, such as Canada Health Infoway and OntarioMD, to assess the progress of advanced EMR feature use to inform future policies designed to sustain the momentum of advanced EMR feature use.

Limitations and Strengths

One limitation of our study is the composition of the participant sample. OntarioMD assisted with recruiting participants by reaching out only to its peer leaders. Peer leaders are typically super users who could be biased favorably toward EMRs. Another limitation is that no participants were located in a rural setting. This group might report other barriers or motives. Researcher bias because of using directed content analysis is another limitation, as researchers are likely to find evidence supportive of their theory. Finally, participants might have answered questions a certain way to please the researcher [24]. Doing an audit trail minimized biased results.

In addition, as the type of EMR software investigated was dependent on the software used by participants, the study only involved 3 types of EMR software: PS Suite, IndiviCare, and Nightingale On Demand. This may have prevented us from observing other advanced EMR features available in other EMR software. Moreover, the EMR software we investigated were all OntarioMD certified, which provided additional benefits (eg, access to Hospital Report Manager, OLIS, and EMR funding eligibility). Other factors might have emerged had we investigated non–OntarioMD-certified EMR systems.

A key strength of this study is that physicians were interviewed in person, providing a deeper understanding of their responses and allowing them to demonstrate certain EMR features. This, in turn, allowed us to observe the barriers and facilitating factors experienced by participants. In addition, the credibility of this study was enhanced by coauthors auditing the results and 2 team members independently analyzing transcripts.

Conclusions

In this study, 9 primary care physicians in Ontario discussed barriers and facilitating factors that influenced their use of advanced EMR features. This study also extended the CAF through the emergence of a new dimension regarding the maturity of users' EMR use. The extended CAF can be used to support key stakeholders in tracking the use of advanced EMR features, which would support future policies. A future research direction could be the development tools (eg, survey or interview guide) to formally evaluate the extended CAF. Overall, our findings show that although primary care physicians' use of EMR systems has improved, barriers remain and need to be addressed to further enhance the physicians' use of advanced EMR features to facilitate CDPM.

Acknowledgments

The authors thank all participating family physicians and patients for their cooperation. The authors also thank OntarioMD for its support in recruiting participants.

Authors' Contributions

RR, the primary investigator, conceived, led, and coordinated the development and writing of the manuscript; RR and SY independently analyzed the transcripts; SY participated throughout the development and writing of the manuscript; RR, SY, and CK audited the data analysis findings; and SY, CK, and JM reviewed and made substantial contributions to the manuscript, contributing intellectual content and feedback on the drafts of the paper. All authors read and approved the final paper.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Electronic medical record maturity model.
[DOCX File , 66 KB - medinform_v7i4e13318_app1.docx]

Multimedia Appendix 2
Clinical adoption framework.
[PNG File , 100 KB - medinform_v7i4e13318_app2.png]

Multimedia Appendix 3
Interview guide.
[DOCX File , 102 KB - medinform_v7i4e13318_app3.docx]

References

1. World Health Organization. 2018. Chronic Diseases and Health Promotion URL: http://www.who.int/chp/about/integrated_cd/en/ [accessed 2018-12-02]
2. Lau F, Price M, Boyd J, Partridge C, Bell H, Raworth R. Impact of electronic medical record on physician practice in office settings: a systematic review. BMC Med Inform Decis Mak 2012 Feb 24;12:10 [FREE Full text] [doi: 10.1186/1472-6947-12-10] [Medline: 22364529]
3. Hsiao C, Marsteller JA, Simon AE. Electronic medical record features and seven quality of care measures in physician offices. Am J Med Qual 2014;29(1):44-52. [doi: 10.1177/1062860613483870] [Medline: 23610232]
4. Woolthuis EP, de Grauw WJ, van Gerwen WH, van den Hoogen HJ, van de Lisdonk EH, Metsemakers JF, et al. Identifying people at risk for undiagnosed type 2 diabetes using the GP's electronic medical record. Fam Pract 2007 Jun;24(3):230-236. [doi: 10.1093/fampra/cmm018] [Medline: 17510087]
5. Shaw N. The role of the professional association: a grounded theory study of Electronic Medical Records usage in Ontario, Canada. Int J Inf Manag 2014;34(2):200-209. [doi: 10.1016/j.ijinfomgt.2013.12.007]
6. Schoen C, Osborn R, Squires D, Doty M, Rasmussen P, Pierson R, et al. A survey of primary care doctors in ten countries shows progress in use of health information technology, less in other areas. Health Aff (Millwood) 2012 Dec;31(12):2805-2816. [doi: 10.1377/hlthaff.2012.0884] [Medline: 23154997]
7. Greiver M, Barnsley J, Glazier RH, Moineddin R, Harvey BJ. Implementation of electronic medical records: theory-informed qualitative study. Can Fam Physician 2011 Oct;57(10):e390-e397 [FREE Full text] [Medline: 21998247]
8. Miller RH, West C, Brown TM, Sim I, Ganchoff C. The value of electronic health records in solo or small group practices. Health Aff (Millwood) 2005;24(5):1127-1137. [doi: 10.1377/hlthaff.24.5.1127] [Medline: 16162555]
9. Bassa A, del Val M, Cobos A, Torremadé E, Bergoñón S, Crespo C, et al. Impact of a clinical decision support system on the management of patients with hypercholesterolemia in the primary healthcare setting. Dis Manag Health Outcomes 2005;13(1):65-72. [doi: 10.2165/00115677-200513010-00007]
10. Rogers EM. Diffusion of Innovations. Fifth Edition. New York: Free Press; 2003.
11. Samoutis G, Soteriades ES, Kounalakis DK, Zachariadou T, Philalithis A, Lionis C. Implementation of an electronic medical record system in previously computer-naïve primary care centres: a pilot study from Cyprus. Inform Prim Care 2007;15(4):207-216 [FREE Full text] [doi: 10.14236/jhi.v15i4.660] [Medline: 18237477]
12. Adaji A, Schattner P, Jones K. The use of information technology to enhance diabetes management in primary care: a literature review. Inform Prim Care 2008;16(3):229-237 [FREE Full text] [doi: 10.14236/jhi.v16i3.698] [Medline: 19094410]
13. The Association of Family Health Teams of Ontario. 2015. What Is A Quality Improvement Decision Support Specialist (QIDSS)? URL: https://www.afhto.ca/news-events/news/what-quality-improvement-decision-support-specialist-qidss [accessed 2018-11-12]
14. Jones M, Koziel C, Larsen D, Berry P, Kubatka-Willms E. Progress in the enhanced use of electronic medical records: data from the Ontario experience. JMIR Med Inform 2017 Feb 22;5(1):e5 [FREE Full text] [doi: 10.2196/medinform.6928] [Medline: 28228372]
15. Leung V, Hagens S, Zelmer J. Drug information systems: evolution of benefits with system maturity. Healthc Q 2013;16(2):43-48. [Medline: 24863449]
16. Lau F, Hagens S, Muttitt S. A proposed benefits evaluation framework for health information systems in Canada. Healthc Q 2007;10(1):112-6, 8. [Medline: 17326376]
17. Lau F, Price M, Keshavjee K. From benefits evaluation to clinical adoption: making sense of health information system success in Canada. Healthc Q 2011;14(1):39-45. [doi: 10.12927/hcq.2011.22157] [Medline: 21301238]
18. Dixon DR. The behavioral side of information technology. Int J Med Inform 1999 Dec;56(1-3):117-123. [doi: 10.1016/s1386-5056(99)00037-4] [Medline: 10659940]
19. Callen JL, Braithwaite J, Westbrook JI. Contextual implementation model: a framework for assisting clinical information system implementations. J Am Med Inform Assoc 2008;15(2):255-262 [FREE Full text] [doi: 10.1197/jamia.M2468] [Medline: 18096917]

XSL•FO
RenderX

20. Canada Health Infoway: Digital Health in Canada. 2013. A Framework and Toolkit for Managing Ehealth Change URL: https://www.infoway-inforoute.ca/en/component/edocman/1659-a-framework-and-toolkit-for-managing-ehealth-change-2/view-document?Itemid=0 [accessed 2018-10-06]
21. OntarioMD. Optimize and Advance EMR Use. 2017. URL: https://www.ontariomd.ca/products-and-services/emr-progress-assessment/emr-maturity-model [accessed 2018-09-20]
22. OntarioMD. 2019. Company Overview URL: https://www.ontariomd.ca/about-us/our-organization [accessed 2019-01-03]
23. eHealth Observatory: University of Victoria. 2019. Clinical Adoption Framework URL: https://ehealth.uvic.ca/methodology/models/CAF.php [accessed 2019-03-16]
24. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
25. Telus Health. 2019. PS Suite EMR URL: https://www.telus.com/en/health/health-professionals/clinics/ps-suite [accessed 2019-01-04]
26. Telus Health. 2019. Support Information for Nightingale Users URL: https://www.telushealth.co/support-information-nightingale-users/ [accessed 2019-01-03]
27. INDIVICA. 2019. IndiviCare 4 URL: http://indivica.ca/ [accessed 2019-01-02]
28. OSCAREMR. 2019. URL: https://oscar-emr.com/ [accessed 2019-01-03]
29. Larsen D. OntarioMD. 2015. Hospital Report Manager: Expansion and Success URL: https://www.ontariomd.ca/articlesdocumentlibrary/emr-approved-j%20sgl%20pages.pdf [accessed 2018-11-21]
30. OntarioMD. 2019. What is Ontario Laboratories Information System (OLIS) Deployment? URL: https://www.ontariomd.ca/products-and-services/olis-deployment [accessed 2019-01-04]
31. Ludwick DA, Doucette J. Primary care physicians' experience with electronic medical records: barriers to implementation in a fee-for-service environment. Int J Telemed Appl 2009;2009:853524 [FREE Full text] [doi: 10.1155/2009/853524] [Medline: 19081787]
32. QSR International. 2018. NVivo URL: https://www.qsrinternational.com/nvivo/home [accessed 2018-11-18]
33. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. J Am Med Assoc 2005 Mar 9;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
34. Boonstra A, Broekhuis M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. BMC Health Serv Res 2010 Aug 6;10:231 [FREE Full text] [doi: 10.1186/1472-6963-10-231] [Medline: 20691097]
35. OntarioMD. 2017. Peer Leader Program URL: https://www.ontariomd.ca/products-and-services/peer-leader-program [accessed 2019-01-05]
36. Ramaiah M, Subrahmanian E, Sriram RD, Lide BB. Workflow and electronic health records in small medical practices. Perspect Health Inf Manag 2012;9:1d [FREE Full text] [Medline: 22737096]
37. Simon SR, Kaushal R, Cleary PD, Jenter CA, Volk LA, Orav EJ, et al. Physicians and electronic health records: a statewide survey. Arch Intern Med 2007 Mar 12;167(5):507-512. [doi: 10.1001/archinte.167.5.507] [Medline: 17353500]
38. Loomis GA, Ries JS, Saywell RM, Thakker NR. If electronic medical records are so great, why aren't family physicians using them? J Fam Pract 2002 Jul;51(7):636-641. [Medline: 12160503]
39. Meade B, Buckley D, Boland M. What factors affect the use of electronic patient records by Irish GPs? Int J Med Inform 2009 Aug;78(8):551-558. [doi: 10.1016/j.ijmedinf.2009.03.004] [Medline: 19375381]

Abbreviations

CAF: Clinical Adoption Framework
CDPM: chronic disease prevention and management
EMR: electronic medical record
FHO: family health organization
FHT: family health team
HIS: health information system
IT: information technology
OLIS: Ontario Laboratories Information System
QIDSS: quality improvement decision support specialist

XSL•FO
RenderX

Edited by G Eysenbach; submitted 09.01.19; peer-reviewed by F Lau, T Jamieson, J Ryan, Y Gong; comments to author 31.01.19; revised version received 27.06.19; accepted 31.10.19; published 29.11.19.

Please cite as:
Rahal RM, Mercer J, Kuziemsky C, Yaya S
Primary Care Physicians' Experience Using Advanced Electronic Medical Record Features to Support Chronic Disease Prevention and Management: Qualitative Study
JMIR Med Inform 2019;7(4):e13318
URL: http://medinform.jmir.org/2019/4/e13318/
doi:10.2196/13318
PMID:31782742

©Rana Melissa Melissa Rahal, Jay Mercer, Craig Kuziemsky, Sanni Yaya. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 29.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Key Factors Affecting Ambulatory Care Providers' Electronic Exchange of Health Information With Affiliated and Unaffiliated Partners: Web-Based Survey Study

John C Pendergrass[1*], PhD; Ranganathan Chandrasekaran[2*], PhD

[1]Operations Management and Information Systems, Northern Illinois University, DeKalb, IL, United States

[2]Center for Health Information Management and Systems, University of Illinois at Chicago, Chicago, IL, United States

[*]all authors contributed equally

Corresponding Author:
John C Pendergrass, PhD
Operations Management and Information Systems
Northern Illinois University
740 Garden Rd
Barsema Hall 328
DeKalb, IL, 60115
United States
Phone: 1 8157531332
Email: jpendergrass@niu.edu

Abstract

Background: Despite the potential benefits of electronic health information exchange (HIE) to improve the quality and efficiency of care, HIE use by ambulatory providers remains low. Ambulatory providers can greatly improve the quality of care by electronically exchanging health information with affiliated providers within their health care network as well as with unaffiliated, external providers.

Objective: This study aimed to examine the extent of electronic HIE use by ambulatory clinics with affiliated providers within their health system and with external providers, as well as the key technological, organizational, and environmental factors affecting the extent of HIE use within and outside the health system.

Methods: A Web-based survey of 320 ambulatory care providers was conducted in the state of Illinois. The study examined the extent of HIE usage by ambulatory providers with hospitals, clinics, and other facilities within and outside their health care system–encompassing seven kinds of health care data. Ten factors pertaining to technology (IT [information technology] Compatibility, External IT Support, Security & Privacy Safeguards), organization (Workflow Adaptability, Senior Leadership Support, Clinicians Health-IT Knowledge, Staff Health-IT Knowledge), and environment (Government Efforts & Incentives, Partner Readiness, Competitors and Peers) were assessed. A series of multivariate regressions were used to examine predictor effects.

Results: The 6 regressions produced adjusted R-squared values ranging from 0.44 to 0.63. We found that ambulatory clinics exchanged more health information electronically with affiliated entities within their health system as compared with those outside their health system. Partner readiness emerged as the most significant predictor of HIE usage with all entities. Governmental initiatives for HIE, clinicians' prior familiarity and knowledge of health IT systems, implementation of appropriate security, and privacy safeguards were also significant predictors. External information technology support and workflow adaptability emerged as key predictors for HIE use outside a clinic's health system. Differences based on clinic size, ownership, and specialty were also observed.

Conclusions: This study provides exploratory insights into HIE use by ambulatory providers within and outside their health care system and differential predictors that impact HIE use. HIE use can be further improved by encouraging large-scale interoperability efforts, improving external IT support, and redesigning adaptable workflows.

(JMIR Med Inform 2019;7(4):e12000) doi:10.2196/12000

XSL•FO
RenderX

KEYWORDS

health information exchange; ambulatory care information systems; ambulatory care facilities

Introduction

Background

Health information exchange (HIE) is the electronic sharing of patient-level clinical health information among health care organizations, providers, and practice settings. HIE allows sharing of health information across organizational and geographic boundaries, making critical patient information available whenever and wherever needed [1,2]. HIE provides clinicians timely access to more complete patient information that is scattered across a fragmented US health care network. Ultimately, HIE can greatly improve quality of care, enhance patient safety, and reduce overall costs.

Evidence about the benefits of HIE is encouraging [3]. Studies on HIE outcomes have reported modest to moderate reductions in duplicate laboratory and radiology testing [4,5], better identification of medication discrepancies, and improved medication reconciliation [6]. HIE use has also decreased hospital readmissions and associated costs [7]. Faster access to patient information via HIE has enhanced emergency care and reduced costs [8]. From a business value viewpoint, effective use of HIEs has enhanced the resource utilization and productivity of providers, thereby improving their overall efficiency [9].

Despite such benefits, HIE usage has been low among ambulatory care providers [10]. In contrast to 76% hospitals who used HIE in 2014, only 42% of ambulatory care physicians engaged in any kind of HIE—within or outside their system. Furthermore, only 26% did so to share health information with external providers [11,12]. The relatively low HIE use among ambulatory clinics and the gap between HIE usage by hospitals and ambulatory care providers are concerning as the potential benefits of HIE will be hard to realize. Even among ambulatory clinics, there seem to be notable variations in the extent of electronic HIE within their health care system and with external providers. A health care system for an ambulatory clinic includes affiliated hospitals, providers, and other health entities who are connected through common ownership, or joint management, or some agreeable business arrangement [13].

HIE use in ambulatory care clinics can be complicated because each setting may have varying information needs, different levels of electronic health record (EHR) adoption, technological sophistication, and resource constraints. Although many technological, environmental, and organizational factors could affect the use of HIE [14], specific factors that affect HIE usage by ambulatory care providers remain underexplored. The influence of specific factors that cause variations in HIE usage within and outside the system is also not known. Exploring these factors is important to ensure the success of HIEs and to further promote their usage among ambulatory care providers. Our study addresses this gap by exploring the association between the key factors and HIE usage by ambulatory care providers. Our research objectives are 2-fold: (1) to assess the extent of HIE usage by ambulatory clinics to exchange health information with internal and external providers and (2) to examine the associations between key technology-organization-environment (TOE) factors and HIE usage within and outside the health care system.

Key Factors Affecting Health Information Exchange Usage

Figure 1 presents a summary of key factors associated with ambulatory clinics' HIE use—within or outside their system. The factors have been grouped based on the TOE framework [15] that has been widely used to understand the contextual factors associated with information technology (IT) usage in organizations.

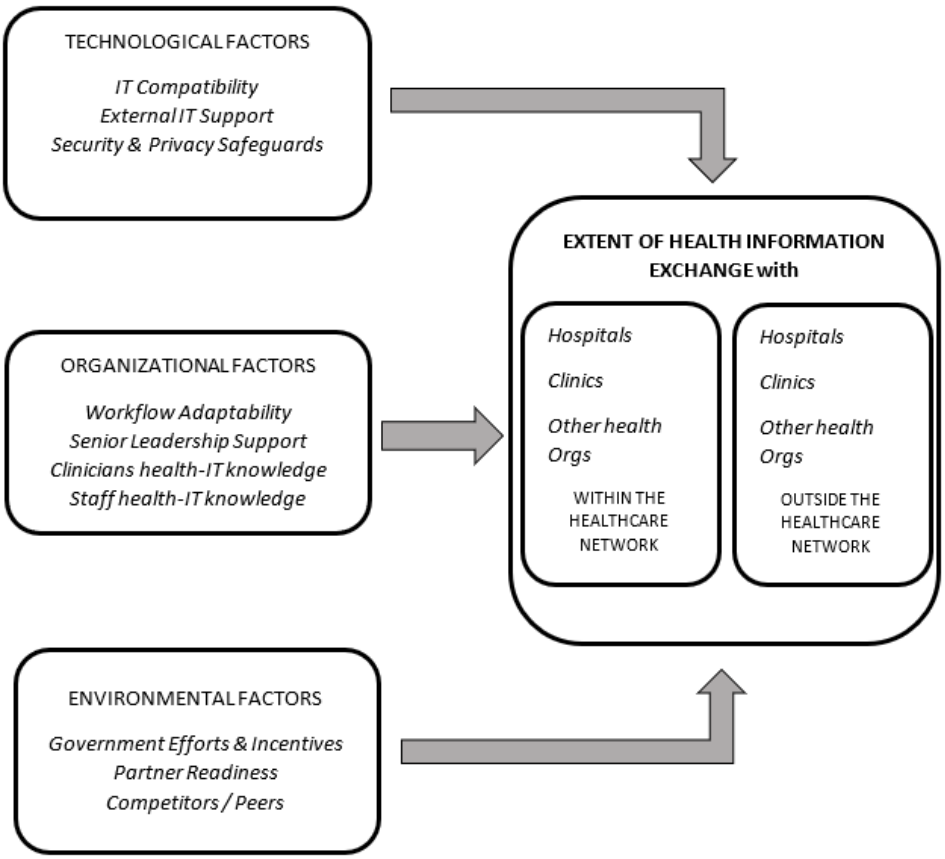
The 2009 Health Information Technology for Economic and Clinical (HITECH) Act included a health care organization's ability to engage in HIE as a criterion for Meaningful Use—the federal standard for incentives provided to clinicians and health care organizations for using EHR systems. Incentive payments for meaningful use in stage 2 require demonstrating actual HIE usage rather than merely possessing the ability to electronically exchange patient information. These efforts have resulted in a proliferation of HIE initiatives by hospitals as well as ambulatory care providers that range from minimally meeting the meaningful use requirements to more intense exchange of health care information [16,17]. Another key environmental factor influencing HIE use is the readiness level of partners. Ambulatory care clinics typically serve patients who also receive care from other partner hospitals, specialty clinics or laboratories and pharmacies, and these organizations could influence HIE use [18]. Ambulatory clinics can face interoperability challenges if partners differ in technological sophistication or use incompatible EHR systems. Competing clinics or providers can also influence HIE use [17,19]. Using HIE can help ambulatory clinics distinguish themselves from other competing clinics. However, in highly competitive markets, providers may be reluctant to share data because of fear of loss of patients to rival organizations [20]. Even among those sharing data, providers can restrict the type and amount of health care information that is shared [19].

Effective use of HIE is influenced by the ambulatory provider's internal capabilities. The primary internal capabilities include workflow processes [16,21] and senior leadership support for HIE [22]. To enhance HIE usage, providers need to effectively incorporate HIE into current workflows and routines [23]. Workflows tend to differ across practice settings [21], and fusing them with HIE can be challenging [10]. Integrating HIE into current workflows can be achieved by (1) tweaking and redesigning existing workflows or (2) customizing the HIE-related information systems to fit in with the clinic-specific workflows. As the latter alternative can be costly, workflows in smaller physician practices will need considerable redesign for integration with HIE systems, and this can be a significant factor inhibiting HIE use in resource-constrained ambulatory settings [18]. Organizational leadership has been consistently found to be important to spearhead HIE initiatives among

XSL•FO
RenderX

hospitals [24] and ambulatory clinics. HIE initiatives have been successful when senior leadership actively engages in HIE efforts and acts as organizational champions [25].

Figure 1. Key predictors of health information exchange use by ambulatory clinics: technology-organization-environment framework. IT: information technology; HIT: health information technology; Orgs: organizations.



A provider's IT infrastructure maturity and sophistication can greatly affect their HIE use. IT infrastructure includes the technological components as well as knowledge levels of key organizational stakeholders to effectively exploit IT to enhance business performance [26]. The fundamental building block for an interoperable, electronic HIE is a robust technological infrastructure and associated EHR system. In addition to flexible and compatible IT infrastructure, recognition of HIE's potential and the knowledge levels of clinicians and administrative staff is an important factor associated with HIE usage [27]. An additional factor that can augment HIE usage is the external technical support provided to HIE users [27]. Ambulatory care providers have historically relied on external consultants and EHR vendors for training and technical support. With an evolving marketplace, identifying and gaining access to health care knowledgeable IT support has been challenging for several health care providers. Availability of quality and timely IT support can augment HIE use. An often-repeated concern in HIE relates to security and privacy of health information that is being exchanged. Implementing technological safeguards can greatly promote HIE use both within a health care system and with outside providers [17].

Building upon the TOE framework and prior studies, we conducted a survey study to assess the associations between these factors on HIE usage by ambulatory care providers within and outside their health care system.

Methods

Setting and Sample

This study draws from a population of ambulatory health care clinics from the state of Illinois. We partnered with 2 Illinois-based health information technology (HIT) regional extension centers (RECs) for recruiting ambulatory clinics to participate in our study. RECs are organizations that have received funding under the HITECH Act to assist health care providers with the selection and implementation of HIT. The primary constituency of RECs consists of small- and mid-sized practices that seek education and assistance in evaluating and implementing HIT. In addition, a list of Illinois rural health clinics and other ambulatory care providers was obtained from the Illinois Department of Public Health. The data were collected through a Web-based survey of office-based physicians in the practice. The Institutional Review Board of the University of Illinois at Chicago approved the study protocol. All participants provided informed consent. Respondents were provided a financial incentive to complete the Web survey.

Survey Questionnaire

The survey questionnaire was constructed using items from several existing instruments that measured HIE usage and perceptions about the key TOE factors associated with HIE usage. We also used relevant items on the contextual factors from relevant HIT and management information systems

XSL•FO
RenderX

literature. Multimedia Appendix 1 lists the survey items along with sources from literature. As most of the questionnaire items for survey were adapted from the prior literature, we gave careful consideration to the content validity of the measures. A total of 5 subject matter experts carefully assessed the survey items and wording of the items in the questionnaire. On the basis of their feedback, minor changes were made to the wording and design of the questionnaire. Then, the questionnaire was pretested with 6 clinicians from ambulatory clinics, and based on their feedback, minor changes were made to the survey.

In the survey, we specified HIE to mean any electronic form of data exchange—excluding fax machines—including any computer-to-computer interface (such as an EHR system) or a Web portal. To assess HIE usage, we asked respondents about their level of usage (coded as 0 for no usage, 1 for partial usage, and 2 for completely using electronic exchange) for sharing different types of health information (patient demographics, referrals, clinical orders, care summaries, physician notes or medication lists, radiology results, and laboratory results) within and outside their health care system. Respondents indicated HIE usage (0-2) for each health information type for 3 partner contexts—hospitals, clinics, and other health care organizations—both within and outside their system. For each ambulatory clinic, aggregate HIE usage scores were computed for each of the 3 partner contexts, that is, we computed 6 scores (ranging from 0 to 14) indicating HIE usage with hospitals, clinics, and others within their health care system, and similarly for the 3 partner contexts outside a clinic's health care system.

Questions on the predictors (survey items are shown in Multimedia Appendix 1) captured the perceptions of respondents on the influence of technological factors (eg, compatibility of IT systems, access to external IT support, and security safeguards for exchanging health information), organizational factors (eg, vision of senior leadership on HIE, adaptability of workflows, and clinician and staff HIT knowledge), and environmental factors (eg, government regulations and incentives, competition, and readiness of partners) on HIE usage. Questions about perceptions used 5-point Likert scale responses ranging from *1, strongly disagree*, to *5, strongly agree*. Mean scores were computed for each of the predictors (Table 1). We also included questions on the ambulatory clinic's demographic information such as the number of clinicians, practice specialty (family medicine, pediatrics, urgent care, obstetrics and gynecology, surgical, and other), type of setting (solo, group, primary care, and specialty care), and ownership status of the clinic (provider owned, hospital or system owned, or mixed).

Table 1. Descriptive statistics of technology-organization-environment predictors.

Predictor factors	Value, mean (SD)
IT[a] compatibility	3.65 (0.89)
External IT support	3.37 (0.69)
Security safeguards	3.81 (0.88)
Senior leadership support	3.35 (0.76)
Workflow adaptability	3.27 (0.65)
Clinician HIT[b] knowledge	3.77 (0.85)
Staff HIT knowledge	3.67 (0.74)
Government initiatives	3.35 (0.71)
Competitor and peer influence	3.01 (0.83)
Partner readiness	3.48 (0.94)

[a]IT: information technology.

[b]HIT: health information technology.

Data Analysis

Demographic characteristics of respondent clinics were assessed using frequencies and percentages. For perceptive measures, wherever multiple items were defined a priori to reflect a factor, we took respondents' average scores across respective items to create a summary score for each factor: influence of government, competition, and peers; HIT knowledge levels of clinicians and staff; workflow adaptability; security safeguards; and external IT support (see Table 1). All the statistical analyses were performed using IBM SPSS version 22 running on a Windows 10 platform.

For multiitem measures, we assessed the reliability (ie, internal consistency) and validity (ie, convergent and discriminant). Cronbach alpha is a standard way to perform a reliability analysis [28], and a range of .70 to .80 is considered acceptable. The Cronbach alpha values of our constructs ranged from .710 to .795 and hence were acceptable. Then, we performed a principal component analysis to assess validity. To assess convergent validity, the degree to which a measure is correlated with other measures that it is theoretically predicted to correlate with, we evaluated the loading of each item onto their specified factor. All loadings were above alpha=.70. We compared the coefficients of the indicators with the standard errors (where the loadings should be at least twice as much as the standard error) and assessed the *t* statistic, which were all significant at P=.05. To assess discriminant validity, we checked cross-factor correlations against the square root of the average variance extracted (AVE) of each factor. As correlations were all smaller

XSL•FO
RenderX

than the square root of AVE, we concluded that discriminant validity was not a problem.

To explore the associations between the TOE factors and HIE usage, we performed a series of linear multivariate regressions. A series of 6 regressions were run with HIE usage as the dependent variable: one for each of the three partner contexts (hospitals, clinics, and others), both within and outside a health care system. Collinearity issues were also diagnosed by the examined variance inflation factor scores, which were all below permissible levels.

Results

Descriptive

Of the 2949 recruitment emails and faxes sent, 383 completed the survey for an overall response rate of 12.98% (383/2949). Moreover, 63 clinics indicated that they did not use any kind of HIE at all and were excluded from our analysis, reducing our usable sample size to 320.

Table 2 shows respondent demographics by number of clinical providers, clinic ownership, practice specialty, and type of setting. Over half (53.0%, 171/320) of the clinics have 10 or fewer providers, 43.0% (136/320) are wholly provider owned, over two-thirds (68.4%, 219/320) are either family medicine or pediatric clinics, and nearly half (48.1%, 154/320) constitute a primary clinic setting (either solo or group practice).

Table 2. Characteristics of respondent ambulatory clinics.

Clinic characteristics	Value, n (%)
Number of providers	
1-2	41 (13)
3-5	36 (11)
6-10	94 (29)
More than 10	149 (46.6)
Practice setting	
Solo primary care	68 (21)
Solo specialty care	104 (32.5)
Group primary care	86 (27)
Group specialty care	62 (19)
Practice type	
Family medicine	101 (31.6)
Pediatrics	118 (36.9)
Urgent care	52 (16)
Surgical	15 (5)
Obstetrics/gynecology	25 (8)
Others	9 (3)
Clinic ownership	
Wholly physician/provider owned	136 (43.0)
Wholly owned by hospital, health care system, HMO[a], etc	123 (38.9)
Partially owned by hospital, health care system, HMO, etc	47 (15)
Others	10 (3)

[a]HMO: health maintenance organization.

Table 3 reports the nature of health information electronically shared by Illinois ambulatory clinics by data type for the three contexts both within and outside of the health care system. The mean and standard deviation of the independent variables is shown as well. Over half of the respondent clinics indicated sharing basic patient demographic information with providers within and outside of their system. However, only around one-third of ambulatory care providers exchanged care summary documents or referrals with other providers within and outside their system. About 40% of clinics indicated exchanging laboratory results and radiology reports with other providers. Electronic exchange of referrals ranged from 44.4% (142/320) to 52.5% (168/320) within the health care system and 35.9% (115/320) to 39.4% (136/320) outside the health care system.

XSL•FO
RenderX

Table 3. Number and percentage of providers exchanging information electronically by data type.

Measure	Within health care system			Outside health care system		
	Hospitals	Clinics	Others[a]	Hospitals	Clinics	Others[a]
Independent variable, mean (SD)	4.92[b] (3.56)	5.17[b] (3.67)	5.28[b] (3.69)	4.35[b] (3.06)	4.57[b] (3.11)	4.68[b] (2.95)
Patient demographics, n (%)	169 (52.8)	178 (55.6)	168 (52.5)	176 (55.0)	171 (53.4)	186 (58.1)
Referrals, n (%)	142 (44.4)	168 (52.5)	163 (50.9)	115 (35.9)	126 (39.4)	118 (36.9)
Clinical orders, n (%)	101 (31.6)	125 (39.1)	103 (32.2)	91 (28)	109 (34.1)	112 (35.0)
Clinical/summary care records, n (%)	107 (33.4)	124 (38.8)	121 (37.8)	98 (31)	118 (36.9)	120 (37.5)
Medication history and/or physician notes, n (%)	121 (37.8)	139 (43.4)	110 (34.4)	96 (30)	116 (32.3)	119 (37.2)
Laboratory results/reports, n (%)	125 (39.1)	117 (36.6)	142 (44.4)	124 (38.8)	119 (37.2)	130 (40.6)
Radiology results/reports, n (%)	124 (38.8)	124 (38.8)	133 (41.6)	137 (42.8)	128 (40.0)	130 (40.6)

[a]Others refer to laboratories, pharmacies, and other care facilities.

[b]Paired sample *t* tests confirmed significant differences in means between the 2 groups, $P<.01$.

Predictors of Health Information Exchange Use

Table 4 presents the results of the 6 multivariate regression models, with panel 1 depicting the regression results for HIE use by Illinois clinics with hospitals, clinics, and other entities *within* their health care system and panel 2 reporting results for HIE use with hospitals, clinics, and others *outside* the Illinois clinics' health care system. All the 6 models had statistically significantly adjusted R-squared values ranging from 0.44 to 0.63. Thus, our models were able to explain 43% to 63% of the HIE usage by sample clinics. Among the environmental factors, as presented in Table 4, partner readiness is the strongest predictor of HIE usage across all the 6 regression models. Government efforts for promoting HIE are also significantly associated with HIE with internal and external entities. According to panel 1, competitor or peer health organizations exert a negative influence on a clinic's HIE use within the system.

Among the organizational factors examined, clinician's knowledge of HIT systems emerged as a strong predictor across all 6 regression models, indicating that clinics with providers who used electronic medical records (EMRs) and other health technologies more are likely to see greater HIE usage. Moreover, senior leadership support for HIE is also a significant predictor of HIE use in 5 of the 6 models. Per panel 2, workflow adaptability is a significant predictor of HIE use for exchanging information with external clinics and other health entities. Panel 2 also indicates knowledge of ambulatory clinic's staff to be negatively associated with HIE use with external clinics and external health entities. Taken together, providers' knowledge and experience with HIT systems are more important to promote HIE usage rather than that of support staff in a clinic.

Among the technological factors, implementation of appropriate security safeguards emerged as a significant predictor of HIE use across all 6 regression models. Per panel 2, external IT support to ambulatory clinics is an important predictor of electronically exchanging health information with *external* hospitals, clinics, and other health entities. Compatibility of existing IT infrastructure is a significant predictor for exchanging health information with other entities within the clinics' health care system.

Examining the clinic characteristics that were associated with HIE use, wholly provider-owned independent clinics, smaller clinics with 1 or 2 providers, family medicine practices, and solo primary care clinics seem to engage in more HIE use within as well as outside the health care system.

XSL•FO
RenderX

Table 4. Regression coefficients: key predictors of health information exchange use by ambulatory clinics.

Predictor variables	HIE[a] within health care system			HIE outside health care system		
	Hospitals	Clinics	Others	Hospitals	Clinics	Others
Environmental factors						
Government initiatives for HIE	0.17[b]	0.12[c]	0.10[d]	0.24[b]	0.27[b]	0.17[b]
Partner readiness	0.36[b]	0.24[b]	0.31[b]	0.24[b]	0.27[b]	0.21[b]
Peers and competitors	−0.24[b]	−21[b]	−0.15[c]	—[d]	—	—
Organizational factors						
Senior leadership support for HIE	0.13[e]	0.16[c]	—	0.16[c]	0.12[e]	0.21[b]
Workflow adaptability	—	—	—	—	0.09[e]	0.10[e]
Clinicians HIT[f] knowledge	0.14[c]	0.17[b]	0.14[c]	0.10[e]	0.10[c]	0.14[b]
Staff HIT knowledge	—	—	—	—	−0.11[c]	−0.12[c]
Technological factors						
IT[g] compatibility	—	—	0.14[c]	—	—	—
External IT support	—	—	—	0.13[c]	0.14[b]	0.09[e]
Security safeguards	0.13[c]	0.10[e]	0.11[e]	0.15[c]	0.17[b]	0.10[c]
Clinic characteristics						
Providers per clinic						
01-Feb	Reference	Reference	Reference	Reference	Reference	Reference
03-May	—	—	—	−0.11[e]	−0.10[e]	—
06-Oct	—	—	—	—	−0.15[c]	—
More than 10	—	—	—	—	−0.15[c]	—
Ownership						
Wholly or partly owned by hospital, HMO[h], or health system	Reference	Reference	Reference	Reference	Reference	Reference
Provider owned	—	0.10[e]	0.11[c]	—	—	0.36[b]
Specialty						
Family medicine	Reference	Reference	Reference	Reference	Reference	Reference
Pediatric care	−0.17[b]	−0.24[b]	−27[b]	−0.16[b]	−22[b]	−27[b]
Urgent care	−0.12[c]	−0.15[b]	−0.12[c]	−0.11[c]	−0.20[b]	−0.19[b]
Obstetrics/gynecology or surgical/others	−0.11[c]	−0.11[c]	−0.13[c]	−0.14[c]	−0.13[b]	−0.22[b]
Setting						
Solo primary care	Reference	Reference	Reference	Reference	Reference	Reference
Solo specialty care	−0.15[c]	−0.14[c]	−25[b]	−0.15[c]	−0.11[e]	−0.12[c]
Group primary care	—	−0.12[e]	−0.20[c]	−0.18[b]	−0.13[c]	−22[b]
Group specialty care	—	—	—	−0.13[e]	−0.11[c]	−0.12[c]
Adjusted R-squared	0.47	0.45	0.44	0.50	0.63	0.61
F test	12.46[b]	11.57[b]	10.84[b]	13.79[b]	22.79[b]	20.88[b]

[a]HIE: health information exchange.

[b]$P \leq .01$.

XSL•FO
RenderX

[c]$P \leq .05$.
[d]Not applicable.
[e]$P \leq .10$.
[f]HIT: health information technology.
[g]IT: information technology.
[h]HMO: health maintenance organization.

Discussion

Analysis

To our knowledge, this is the first empirical study to examine predictors of HIE within and outside a health care system in ambulatory settings. Using data from 383 clinics, we examined TOE factors associated with the use of electronic HIE by small- and mid-sized clinics across the state of Illinois. Usage was examined for exchange partners within the clinic's health care system, if any, and outside their health care system. For each system context, 3 categories of exchange partners were analyzed: hospitals, clinics, and other health care–related entities. Regression results varied by usage context, thereby revealing a differentiated understanding of the phenomena not previously examined.

Partner readiness emerged as the most significant predictor of HIE usage with all affiliates within as well as outside health care system. Governmental initiatives for HIE, clinicians' prior familiarity and knowledge of HIT systems, and implementation of appropriate security and privacy safeguards are also significant predictors of all the 6 contexts examined.

A key factor influencing decisions about the use of a health technology is the knowledge about the capabilities, limitations, and consequences of using that technology. For clinicians, knowledge about HIT systems improves with prior and continued use of technologies. A significant relationship between clinician HIT knowledge and HIE usage was indicated for all contexts both within and outside the health care system. When exchanging patient information in traditional ways (eg, fax), the workflow often utilizes administrative staff for performing these functions. However, the use of electronic HIE puts capabilities and access to information instantly at the hands of clinicians. If clinicians are required to utilize EMR systems, and different forms of electronic HIE are tied to the EMR, then usage is dependent on the clinician's knowledge of the EMR technology. In contrast, we found staff knowledge of HIT to be an insignificant predictor of electronic HIE usage within the health care system, but it was negatively related to HIE usage in 2 contexts outside the health care system. Staff knowledge might not be sufficient to promote electronic HIE, rather it is those clinicians who are adept at using EMR and other HIT systems who are likely to promote electronic HIE in ambulatory settings. Our survey assessed clinician perceptions regarding the knowledge of staff about HIT systems in general, and our findings show that generic knowledge of HIT systems might prove to be a negative factor for exchange of health information with unaffiliated entities. Our findings point to the need for additional HIE-specific training that may be required for staff in ambulatory clinics.

Peer practices, which include competition, indicated a negative and significant relationship for the 3 contexts within the clinic's health care system but no statistical significance outside the health care system. Our findings regarding peer influence inhibiting HIE usage within a health care system underscore competitive concerns for provider organizations' participation in HIE.

Workflow adaptability and external IT support exhibited significant positive association with HIT usage outside the clinic's health care system. HIE within a clinic's health care system might be relatively easier as the entities may not have much variability in terms of workflows and EMR systems. However, considerable variations in terms of HIT systems, processes, and workflows could impede electronic HIE with external partners. Managers and policy makers need to work toward more standardization of workflows and offer greater technical support to spur more HIE usage of clinics with partners outside their health care system.

Our findings confirm significant variations in HIE usage within and outside a clinic's health care system. This finding coupled with the need for external IT support for HIE use with external providers underscores a broader interoperability problem in exchanging health information. For instance, the leading EMR vendor Epic Systems has been criticized for limiting the data exchange capabilities and for charging providers additional fees to enable data exchange with providers who use non-Epic systems [29]. In response to such criticisms and subsequent congressional hearings [30], several new interoperability improvement efforts have been proposed. These initiatives, though well intended, could further increase the fragmentation. Deliberate *information blocking* practices by EHR vendors and providers where they knowingly interfere with electronic exchange of health information for competitive purposes can also impede HIE use [31]. Ambulatory clinics could be coerced into implementing systems and document exchange practices by larger providers or vendors [32]. Recent federal efforts aimed at implementing and improving interoperability standards are moving in the right direction to promote HIE use by clinical providers, for instance, the support by the Office of the National Coordinator for Health Information Technology for the development of solutions using the Fast Health care Interoperability Resources standard [33].

Limitations

A major limitation of this study stems from the sample being drawn from a single state. However, HIE efforts in the state of Illinois are not unlike those in many other states and, therefore, results may be typical. State-level efforts to improve HIE have varied across the country, and these could have differential impacts on HIE use by ambulatory providers. Another limitation pertains to self-reported data. We relied on perceptions of our

XSL•FO
RenderX

respondents to capture our variables. We were unable to obtain specific objective measures of HIE use. More objective measures of HIE such as the number of records exchanged or volume of data exchanged could be a fruitful extension of our research.

Conclusions

This study provides exploratory insights into HIE use by ambulatory providers within and outside their health care system and differential predictors that impact HIE use. By identifying key predictors, we have highlighted the importance of federal efforts, status of partners, and HIT knowledge of clinicians. HIE use by ambulatory providers can be further improved by encouraging large-scale interoperability efforts across the industry, improving external IT support, and redesigning adaptable workflows.

Acknowledgments

The Web survey was conducted with support from Chicago Health IT Regional Extension Center; Illinois Health Information Technology Regional Extension Center; and the Office of Health Information Technology, State of Illinois. This research was partly supported by the National Science Foundation program on Interdisciplinary Graduate Education and Research Traineeship (NSF-IGERT) grant on Electronic Security and Privacy provided to the University of Illinois at Chicago.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire items and sources.
[PDF File (Adobe PDF File), 312 KB - medinform_v7i4e12000_app1.pdf]

References

1. Kuperman GJ. Health-information exchange: why are we doing it, and what are we doing? J Am Med Inform Assoc 2011;18(5):678-682 [FREE Full text] [doi: 10.1136/amiajnl-2010-000021] [Medline: 21676940]
2. Adler-Milstein J, Jha AK. Health information exchange among US hospitals: who's in, who's out, and why? Healthc (Amst) 2014 Mar;2(1):26-32. [doi: 10.1016/j.hjdsi.2013.12.005] [Medline: 26250086]
3. Hersh WR, Totten AM, Eden KB, Devine B, Gorman P, Kassakian SZ, et al. Outcomes from health information exchange: systematic review and future research needs. JMIR Med Inform 2015 Dec 15;3(4):e39 [FREE Full text] [doi: 10.2196/medinform.5215] [Medline: 26678413]
4. Jung H, Vest JR, Unruh MA, Kern LM, Kaushal R, HITEC Investigators. Use of health information exchange and repeat imaging costs. J Am Coll Radiol 2015 Dec;12(12 Pt B):1364-1370 [FREE Full text] [doi: 10.1016/j.jacr.2015.09.010] [Medline: 26614881]
5. Yaraghi N. An empirical analysis of the financial benefits of health information exchange in emergency departments. J Am Med Inform Assoc 2015 Nov;22(6):1169-1172. [doi: 10.1093/jamia/ocv068] [Medline: 26117143]
6. Boockvar K, Ho W, Pruskowski J, DiPalo K, Wong J, Patel J, et al. Effect of health information exchange on recognition of medication discrepancies is interrupted when data charges are introduced: results of a cluster-randomized controlled trial. J Am Med Inform Assoc 2017 Nov 1;24(6):1095-1101. [doi: 10.1093/jamia/ocx044] [Medline: 28505367]
7. Vest J, Kern L, Silver M, Kaushal R, HITEC investigators. The potential for community-based health information exchange systems to reduce hospital readmissions. J Am Med Inform Assoc 2015 Mar;22(2):435-442. [doi: 10.1136/amiajnl-2014-002760] [Medline: 25100447]
8. Everson J, Kocher KE, Adler-Milstein J. Health information exchange associated with improved emergency department care through faster accessing of patient information from outside organizations. J Am Med Inform Assoc 2017 Apr 1;24(e1):e103-e110. [doi: 10.1093/jamia/ocw116] [Medline: 27521368]
9. Walker DM. Does participation in health information exchange improve hospital efficiency? Health Care Manag Sci 2018 Sep;21(3):426-438 [FREE Full text] [doi: 10.1007/s10729-017-9396-4] [Medline: 28236178]
10. Adler-Milstein J, Lin SC, Jha AK. The number of health information exchange efforts is declining, leaving the viability of broad clinical data exchange uncertain. Health Aff (Millwood) 2016 Jul 1;35(7):1278-1285. [doi: 10.1377/hlthaff.2015.1439] [Medline: 27385245]
11. Akhlaq A, McKinstry B, Muhammad KB, Sheikh A. Barriers and facilitators to health information exchange in low- and middle-income country settings: a systematic review. Health Policy Plan 2016 Nov;31(9):1310-1325. [doi: 10.1093/heapol/czw056] [Medline: 27185528]
12. Furukawa MF, King J, Patel V, Hsiao C, Adler-Milstein J, Jha AK. Despite substantial progress In EHR adoption, health information exchange and patient engagement remain low in office settings. Health Aff (Millwood) 2014 Sep;33(9):1672-1679. [doi: 10.1377/hlthaff.2014.0445] [Medline: 25104827]

XSL•FO
RenderX

13. The Agency for Healthcare Research and Quality. Defining Health Systems URL: https://www.ahrq.gov/chsp/chsp-reports/resources-for-understanding-health-systems/defining-health-systems.html [accessed 2018-12-10] [WebCite Cache ID 74ZMRoC62]
14. Eden KB, Totten AM, Kassakian SZ, Gorman PN, McDonagh MS, Devine B, et al. Barriers and facilitators to exchanging health information: a systematic review. Int J Med Inform 2016 Apr;88:44-51 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.01.004] [Medline: 26878761]
15. Baker J. The technology–organization–environment framework. In: Dwivedi Y, Wade M, Schneberger S, editors. Information Systems Theory: Explaining and Predicting Our Digital Society. New York, NY: Springer; 2010:231-245.
16. Kruse CS, Regier V, Rheinboldt KT. Barriers over time to full implementation of health information exchange in the United States. JMIR Med Inform 2014 Sep 30;2(2):e26 [FREE Full text] [doi: 10.2196/medinform.3625] [Medline: 25600635]
17. Yeager VA, Walker D, Cole E, Mora AM, Diana ML. Factors related to health information exchange participation and use. J Med Syst 2014 Aug;38(8):78. [doi: 10.1007/s10916-014-0078-1] [Medline: 24957395]
18. McCullough JM, Zimmerman FJ, Bell DS, Rodriguez HP. Electronic health information exchange in underserved settings: examining initiatives in small physician practices & community health centers. BMC Health Serv Res 2014 Sep 21;14:415 [FREE Full text] [doi: 10.1186/1472-6963-14-415] [Medline: 25240718]
19. Vest JR, Gamm LD. Health information exchange: persistent challenges and new strategies. J Am Med Inform Assoc 2010;17(3):288-294 [FREE Full text] [doi: 10.1136/jamia.2010.003673] [Medline: 20442146]
20. Adler-Milstein J, Jha AK. Sharing clinical data electronically: a critical challenge for fixing the health care system. J Am Med Assoc 2012 Apr 25;307(16):1695-1696. [doi: 10.1001/jama.2012.525] [Medline: 22535851]
21. Unertl K, Johnson K, Lorenzi N. Health information exchange technology on the front lines of healthcare: workflow factors and patterns of use. J Am Med Inform Assoc 2012;19(3):392-400 [FREE Full text] [doi: 10.1136/amiajnl-2011-000432] [Medline: 22003156]
22. Fontaine P, Ross SE, Zink T, Schilling LM. Systematic review of health information exchange in primary care practices. J Am Board Fam Med 2010;23(5):655-670 [FREE Full text] [doi: 10.3122/jabfm.2010.05.090192] [Medline: 20823361]
23. Rudin R, Volk L, Simon S, Bates D. What affects clinicians' usage of health information exchange? Appl Clin Inform 2011 Jan 1;2(3):250-262 [FREE Full text] [doi: 10.4338/ACI-2011-03-RA-0021] [Medline: 22180762]
24. Korst LM, Aydin CE, Signer JM, Fink A. Hospital readiness for health information exchange: development of metrics associated with successful collaboration for quality improvement. Int J Med Inform 2011 Aug;80(8):e178-e188 [FREE Full text] [doi: 10.1016/j.ijmedinf.2011.01.010] [Medline: 21330191]
25. Feldman SS, Schooley BL, Bhavsar GP. Health information exchange implementation: lessons learned and critical success factors from a case study. JMIR Med Inform 2014 Aug 15;2(2):e19 [FREE Full text] [doi: 10.2196/medinform.3455] [Medline: 25599991]
26. Byrd TA, Turner D. Measuring the flexibility of information technology infrastructure: exploratory analysis of a construct. J Manag Inf Syst 1999;17(1):167-208. [doi: 10.1080/07421222.2000.11045632]
27. Vest JR, Ancker JS. Health information exchange in the wild: the association between organizational capability and perceived utility of clinical event notifications in ambulatory and community care. J Am Med Inform Assoc 2017 Jan;24(1):39-46. [doi: 10.1093/jamia/ocw040] [Medline: 27107436]
28. Cronbach LJ, Shavelson RJ. My current thoughts on coefficient alpha and successor procedures. Educ Psychol Meas 2004;64(3):391-418. [doi: 10.1177/0013164404266386]
29. Garber S, Gates SM, Keeler EB, Vaiana ME, Mulcahy AW, Lau C, et al. Redirecting innovation in US health care: options to decrease spending and increase value. Rand Health Q 2014;4(1):3 [FREE Full text] [Medline: 28083317]
30. Alexander L. US Government Publishing Office. America's Health IT Transformation: Translating the Promise of Electronic Health Records into Better Care URL: https://www.gpo.gov/fdsys/pkg/CHRG-114shrg93864/pdf/CHRG-114shrg93864.pdf [accessed 2018-12-13] [WebCite Cache ID 74cpjx7yB]
31. Downing K, Mason J. ONC targets information blocking. J American Health Inform Manag Assoc 2015 Jul;86(7):36-38. [Medline: 26642619]
32. Adler-Milstein J, Pfeifer E. Information blocking: is it occurring and what policy strategies can address it? Milbank Q 2017 Mar;95(1):117-135 [FREE Full text] [doi: 10.1111/1468-0009.12247] [Medline: 28266065]
33. AHA Data. 2014 AHA Annual Survey Information Technology Supplement Health Forum, LLC URL: https://www.ahadataviewer.com/Global/survey%20instruments/2014AHAITSurvey.pdf [accessed 2018-12-10] [WebCite Cache ID 74ZNGMnEX]

Abbreviations

AVE: average variance extracted
EHR: electronic health record
EMR: electronic medical record
HIE: health information exchange
HIT: health information technology

XSL•FO
RenderX

HITECH: Health Information Technology for Economic and Clinical
IT: information technology
REC: regional extension center
TOE: technology-organization-environment

Edited by G Eysenbach; submitted 21.08.18; peer-reviewed by K Colorafi, E Andrikopoulou, N Miyoshi; comments to author 11.10.18; revised version received 12.12.18; accepted 31.12.18; published 07.11.19.

Please cite as:
Pendergrass JC, Chandrasekaran R
Key Factors Affecting Ambulatory Care Providers' Electronic Exchange of Health Information With Affiliated and Unaffiliated Partners: Web-Based Survey Study
JMIR Med Inform 2019;7(4):e12000
URL: http://medinform.jmir.org/2019/4/e12000/
doi:10.2196/12000
PMID:31697241

©John C Pendergrass, Ranganathan Chandrasekaran. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 07.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Impacts of the Perceived Transparency of Privacy Policies and Trust in Providers for Building Trust in Health Information Exchange: Empirical Study

Pouyan Esmaeilzadeh[1], PhD

Department of Information Systems and Business Analytics, College of Business, Florida International University, Miami, FL, United States

Corresponding Author:
Pouyan Esmaeilzadeh, PhD
Department of Information Systems and Business Analytics
College of Business
Florida International University
Modesto A Maidique Campus
11200 SW 8th Street
Miami, FL, 33199
United States
Phone: 1 9543488482
Email: pesmaeil@fiu.edu

Abstract

Background: In the context of exchange technologies, such as health information exchange (HIE), existing technology acceptance theories should be expanded to consider not only the cognitive beliefs resulting in adoption behavior but also the affect provoked by the sharing nature of the technology.

Objective: We aimed to study HIE adoption using a trust-centered model. Based on the Theory of Reasoned Action, the technology adoption literature, and the trust transfer mechanism, we theoretically explained and empirically tested the impacts of the perceived transparency of privacy policy and trust in health care providers on cognitive and emotional trust in an HIE. Moreover, we analyzed the effects of cognitive and emotional trust on the intention to opt in to the HIE and willingness to disclose health information.

Methods: A Web-based survey was conducted using data from a sample of 493 individuals who were aware of the HIE through experiences with a (or multiple) provider(s) participating in an HIE network.

Results: Structural Equation Modeling analysis results provided empirical support for the proposed model. Our findings indicated that when patients trust in health care providers, and they are aware of HIE security measures, HIE sharing procedures, and privacy terms, they feel more in control, more assured, and less at risk. Moreover, trust in providers has a significant moderating effect on building trust in HIE efforts (P<.05). Results also showed that patient trust in HIE may take the forms of opt-in intentions to HIE and patients' willingness to disclose health information that are exchanged through the HIE (P<.001).

Conclusions: The results of this research should be of interest to both academics and practitioners. The findings provide an in-depth dimension of the HIE privacy policy that should be addressed by the health care organizations to exchange personal health information in a secure and private manner. This study can contribute to trust transfer theory and enrich the literature on HIE efforts. Primary and secondary care providers can also identify how to leverage the benefit of patients' trust and trust transfer process to promote HIE initiatives nationwide.

(JMIR Med Inform 2019;7(4):e14050) doi:10.2196/14050

KEYWORDS

cognitive trust in competence; cognitive trust in integrity; emotional trust; perceived transparency of privacy policy; trust in health care providers

XSL•FO
RenderX

Introduction

Background

Trust plays a significant role in the situations where there is a distance between consumers and vendors, such as in internet-dependent contexts [1]. Health information exchange (HIE) networks share health information electronically with other care providers to improve care coordination and enhance patient safety. HIE projects can help primary and secondary health care providers by connecting them via information exchange networks. Different sharing mechanisms are being used by public and private health care organizations to facilitate information exchange initiatives [2]. Existing studies in HIE indicate that the following 3 exchange models are mainly applied by health care entities to electronically transmit patient health information: direct, query-based, and patient-centered exchange [3]. Electronic data exchange between providers can also take place at a regional or national level [4]. In a nationwide HIE project, health care organizations can exchange patients' information across a huge network of providers consistent with nationally defined standards and contracts [5]. In a regional HIE initiative, medical records are shared electronically with unaffiliated hospitals or ambulatory providers in a particular region. A regional health information organization is a third party that enables information exchange across health care organizations within a community, county, or state HIE platform [6].

HIE initiatives utilize sharing mechanisms with which health information is mostly transmitted without a patient's close supervision; thus, patient trust in the HIE is the core in this setting where a great deal of security concerns and privacy risks may be involved [3]. Trust in HIE technology can predict the direction of patients' responses to the implementation of HIE in health care organizations, especially when patients perceive that they may not know everything about this sharing mechanism. The lack of awareness is mainly because of the distance imposed between patients and actual users (health care organizations), lack of direct interactions between patients and HIE networks, newness and evolving nature of HIE initiatives, and unfamiliar mechanisms used in the HIE system to share health information electronically. These characteristics create a setting that is more intangible than the traditional sharing methods (such as fax or mail). The mentioned reasons may make patient trust more critical in the HIE context.

At this point, the use of HIEs by patients or health care professionals is not at a stage of diffusion [7,8]. There is vigorous debate among the entities in the health care industry about the topic of opt-in versus opt-out of digital health records and electronic exchange of such information [9]. This debate argues whether health care providers or patients should have the right to decide whether the digital health information should be exchanged [10]. However, patients have the unconditional right to be aware of the data-handling practices of medical providers [11]. Public perspectives are important to researchers and policy makers because patients are one of the key stakeholders, and the widespread adoption of HIEs is not possible without their positive beliefs and attitudes toward this technology (such as trust factors). Therefore, it is noteworthy to determine whether consumers will choose to opt in to an HIE system if they are given the choice in the near future.

Human thoughts and decisions include cognition and emotion [12]; therefore, both beliefs and feelings should be investigated to better understand how a patient would trust and react to a system that is leveraged by other users (health care providers) to disseminate health information. Consistent with the trust transfer process [13], the level of trust in health care providers can be migrated to trust in HIE systems. Accordingly, patients' trust in HIE characteristics can be derived from a trusted physician who has certain association with the HIE [14]. A patient has to trust an HIE system before he or she is willing to make an opt-in decision or disclose personal health information. Consumers will rely on a technology because of not only the technology's efficiency and effectiveness but also the fair and honest relational exchange between the information technology (IT) and them [15]. Thus, trust can reflect its effects through a cognitive process (robust rational reasons) and an emotional procedure (strong affects and feelings).

In line with the previous research [16], individual trust is defined as the levels of trust in the specific characteristics of a trustee, such as competence, integrity, and benevolence. Extrapolating to the HIE context, it is expected that the patients should trust some HIE characteristics to opt in to the HIE and become more willing to disclose their personal health information. In the context of HIE, trust in competence refers to the trust in the HIE's abilities, technical capabilities, skills, and expertise embedded in the technology. This dimension of trust implies the extent to which patients rely on technologically competent performance of HIE to effectively disseminate health information among a wide variety of health organizations. Trust in integrity describes the belief that the agreement between the patients and an HIE network is reliable, the HIE system honestly fulfills predetermined promises, and the HIE adheres to a set of principles that the patient finds acceptable. Trust in benevolence pertains to the belief that HIE cares about patients beyond the expected commitments to genuinely act in the patients' interests. On the basis of this dimension, the HIE initiatives are believed to seek joint gain in dyadic relationships with patients aside from profit motives to openly follow patients' welfare. Emotional trust implies an emotional security that enables individuals to feel assured that an IT will be responsive in uncertain situations.

Previous research on how patients' trust in HIE is built is still scarce [17]. Despite the importance of patient trust in HIE, the nature of patient trust has not been thoroughly conceptualized, clearly measured, and fully delineated in this context. Previous studies mainly investigate different dimensions of cognitive trust (mostly trust in network design characteristics), and relatively little attention has been given to other trust dimensions [18]. Moreover, the difference between the different levels and dimensions of patient trust in the HIE context has not been analyzed. For a patient to trust an HIE network, the patient should feel assured that the HIE will not compromise personal health information and sensitive medical records and will not act unreasonably [19]. Sharing sensitive health information through a technology that is used by health care providers

XSL•FO
RenderX

requires a new lens for understanding health consumers' opt-in intention toward HIE. According to Kim et al [20], traditional IT research mostly focuses on organizational employees as users who adopt traditional IT for work-related purposes. In the contexts of many technology adoption studies, cognitive factors (eg, effort expectancy or facilitating conditions) can overshadow the effects of emotional variables (eg, emotional trust) on adoption decisions. The existing theories of IT adoption (such as technology acceptance model and unified theory of acceptance and use of technology) are mostly cognitive oriented and focus on users' intention to accept and use a technology. However, in the HIE context, consumers are not the main users. Patients are the beneficiaries of HIE, but they are not the final users. The users are the health care professionals (ie, physicians and nurses), and the decision to adopt HIE is made at the practice or hospital level.

Information system (IS) literature shows that people feelings about IT impact their adoption decisions [21]. Our study is an attempt to extend this research stream by describing 2 aspects of trust (cognitive and emotional) and examining their roles in consumers' opt-in intention and their willingness to disclose health information. The main point of this research is that when a technology (eg, an HIE) deals with sharing sensitive information and may exacerbate privacy concerns, patients will not only depend on cognitive factors to shape opt-in intentions and make information disclosure decisions. This study takes a trust-based perspective to investigate HIE adoption from the patients' standpoint. On the basis of the study by Chopra and Wallace [22], trust plays an important role in situations where 2 sides are dependent, and this dependency may cause risk. In the context of HIE, given the amount of information exchanged among health care organizations, patients depend on HIE to improve treatment process, enhance care coordination, and increase the quality of care before they actually experience the possible effects. In this setting, risk can arise because patients may be concerned that too much personal information is shared, or erroneous health information is exchanged among health care providers through HIEs [23]. Therefore, health consumers' reactions to HIE implementation largely depend on their trust in the HIEs.

To the best of our knowledge, the nature of trust and the differences between the dimensions of patient trust in HIE have not been clearly described. Few empirical studies examined the impact of trust in health care providers on building patient trust in HIE from a trust transfer mechanism. Moreover, patients' decisions about HIE (such as opt-in decision) may not be purely cognitive based because of the special context in which this sharing technology is implemented and used. In many IT adoption decisions at the individual level, consumers' affective reactions influence their choices [24]. In the HIE context, patients may not directly share their health information through exchange mechanisms, and they are distant from care providers who actually use these systems. Such a situation can downplay the pure impact of cognitive factors and give more weight to emotion because of the uncertainty associated with HIE and how this technology is used. Our study aims to advance the existing understanding of patient trust by defining and differentiating it in the HIE adoption setting from the patient's perception. This study uses a balanced perspective to take both aspects of human experience (cognitive and emotional) into account and show whether emotional factors affect the consumers' willingness to disclose health information and their intention to opt in to a technology designed to exchange their sensitive health information.

The purpose of the study was to contribute to the current literature in trust transfer and propose a practical solution to improving patient trust and opt-in rates for HIE. This study is conducted to contribute to the existing research by investigating how individual consumers develop trust in HIE and in what manner dimensions of trust will affect their resultant decisions related to HIE. This research is derived from the literature on trust transfer and IT adoption by articulating how perceived transparency of privacy policy and trust in health care providers impact opt-in intention and willingness to disclose health information through enhancing cognitive and emotional trust in HIE characteristics.

Theoretical Background and Related Literature

Previous literature highlights the role of privacy statement in trust building in other contexts, for example, Web-based shopping, website registration, and mobile internet use. The completeness and transparency of Web privacy statements influence Web-based consumers' perceptions and behavioral intentions to purchase products [25]. In electronic commerce (e-commerce) settings, the content of privacy statements is found as a significant factor to predict consumer trust in websites [26]. According to Callanan et al [27], user awareness of privacy policy has a direct effect on using mobile internet. The presence of a solid website privacy policy heightens the Web-based shoppers' trust and, in turn, reduces their privacy concerns [28]. Framing a rigorous privacy statement that shows organizational compliance with the personal data protection regulations can significantly influence the consumers' buying decisions [29].

As reported by Tsai et al [30], if Web-based retailers provide accessible and transparent privacy policy guidelines, consumers are more likely to pay a premium to purchase services from privacy protective websites. When a privacy statement is clearly presented by websites, consumers are more willing to read it carefully to get more Web-based services [31]. Recent studies indicate that adults are likely to avoid using mobile apps or opt out of Web-based services because of the absence of solid privacy statements [32]. If consumers are well informed about Web-based privacy terms and conditions, they provide more information to websites [33]. On the contrary, if no details are presented in privacy policies, customers are not aware of collecting and sharing procedures. Privacy policy dimensions contain details that empower customers by clarifying their rights and the options they may have to better control the use of health information. For instance, if they can opt out of information sharing with a third party, they will feel more control over their personal data, and this feeling makes Web-based services appear more trustworthy to them [34].

With the advance of technologies used for information exchange, a great number of consumers are anxious about the disclosure, transfer, and sale of personal information that organizations collect from them. Privacy policies should be framed to address

XSL•FO
RenderX

patients' privacy and security concerns. Privacy policy statements define how a health care organization collects, manages, uses, and disseminates personal health information (ranging from less sensitive to highly sensitive). Previous studies in the HIE context described that HIE privacy policies should be informative and comprehensive to reassure patients that exchanging their health information is a low-risk practice [35]. However, it is still not clear what type of contents, dimensions, and format an HIE privacy policy should cover to raise public awareness and build cognitive trust in HIE. Privacy policies are mainly devised based on the 5 dimensions of Fair Information Practice Principles: notice, access, choice, security, and enforcement [26]. Notice refers to the commitment of organizations to send timely announcements to consumers about their information collection practices before personal information is collected. Choice indicates that the consumers should be given the options about how the collected personal information would be used. Access means defining the consumers' rights to view their own personal data and check whether such data are accurate and complete. Security defines the required steps and actions that should be taken by the organizations to ensure security and integrity of the consumers' personal information. Enforcement articulates which national or international mechanisms, guidelines, and instruments are in place to enforce principles of privacy protection. Thus, HIE initiatives should clearly communicate their privacy policy standpoint to patients to increase the degree of trust.

A large number of IS studies treat trust as trusting beliefs [13,36]. Trusting beliefs are the cognitive beliefs shaped by the trustor based on the trustee's trust-related characteristics (ie, competence, integrity, and benevolence) [12]. This cognitive trust is the result of a rational process in which a trustor expects that a trustee will own the required attributes that are reliable. Thus, cognitive trust is developed by a conscious calculation of advantages leading to rational reasons to trust a trustee. A mechanism that helps develop cognitive trust is the trust transfer, which is a cognitive process that may arise from a trusted entity to another new context [37]. According to Stewart [38], the trust transfer process relies on the relationships and interactions between the source and target. In the HIE settings, health care providers can be considered as the source, whereas the target is HIE systems, and the interaction is the efforts made to develop a transparent privacy policy model for information exchange. Thus, patients may form same perceptions about HIE because this technology will be used by the trusted health care providers. Nevertheless, rational expectations are not adequate for individuals to make trust-related decisions [39]. Previous trust literature describes trust in IT as a combination of both reasoning (cognitive trust) and feeling (emotional trust) [12]. Emotional trust, which is an individual's evaluation of feeling and faith [40], is developed by emotional reactions to the trustee. In the context of dealing with an IT, emotional trust denotes whether an individual feels comfortable and secure about relying on the technology.

As cognitive and emotional trust are 2 different concepts, it is important to consider both types of trust in our research to portray a more comprehensive effect of trust on individuals' reactions to the HIE implementation. On the basis of previous studies [41], trust in HIE is defined as follows:

Cognitive Trust in Competence

An individual's rational beliefs about the technical expertise and ability of an HIE to exchange health information among health care entities.

Cognitive Trust in Integrity

An individual's rational beliefs related to the honesty of the exchange process.

Cognitive Trust in Benevolence

An individual's rational beliefs that an HIE system always considers the patient's interest.

Emotional Trust

An individual's feelings of assurance and security about relying on an HIE to share information across health care providers.

The 3 dimensions of trust are treated independently because they are conceptually and operationally different [13]. For instance, an HIE system may have the competence required to exchange information, but the consumers may be worried that the HIE might be designed to be biased by sharing sensitive information for other purposes (such as marketing). Alternatively, the consumers may perceive that an HIE network exhibits care to the patients, especially, in case new conditions of information sharing arise (when no agreement and commitment were made before), but the HIE does not have adequate technical capability. Trust in HIE's benevolence is not easy to evaluate because individuals may not be likely to form the beliefs that HIE networks show care and goodwill beyond the main tasks of sharing personal health information in competent and honest manners. Previous studies in other contexts also indicate that cognitive trust in benevolence may not apply to every technology [39]. As there is no bilateral interactions and close personal relationships between patients and an HIE system, the HIE is not considered as a social actor, and cognitive trust in the benevolence of HIE may not be conceivable. Therefore, consistent with the key tasks HIEs are designed to perform, only cognitive trust in competence and integrity were used in this study.

The main theoretical foundation applied in this study was the Theory of Reasoned Action (TRA) [42]. According to TRA, an individual's intention to perform a behavior is dependent on 2 variables: attitude and subjective norms. Attitude indicates an individual's positive or negative feelings about a behavior, and a subjective norm denotes an individual's perception about whether their significant others believe he or she should or should not engage in the behavior. Consistent with the study by Karahanna et al [43], the effect of subjective norms (which are normative beliefs) becomes more significant when there is a lack of experience with an IT. Furthermore, a subjective norm is a salient factor when a user perceives social pressure from important others to adopt a technology for his or her personal usage. In the context of HIE adoption, patients will not actually be able to use this technology and may only shape attitudes and form beliefs toward using a new system in health care organizations to manage information exchange among a wide

XSL•FO
RenderX

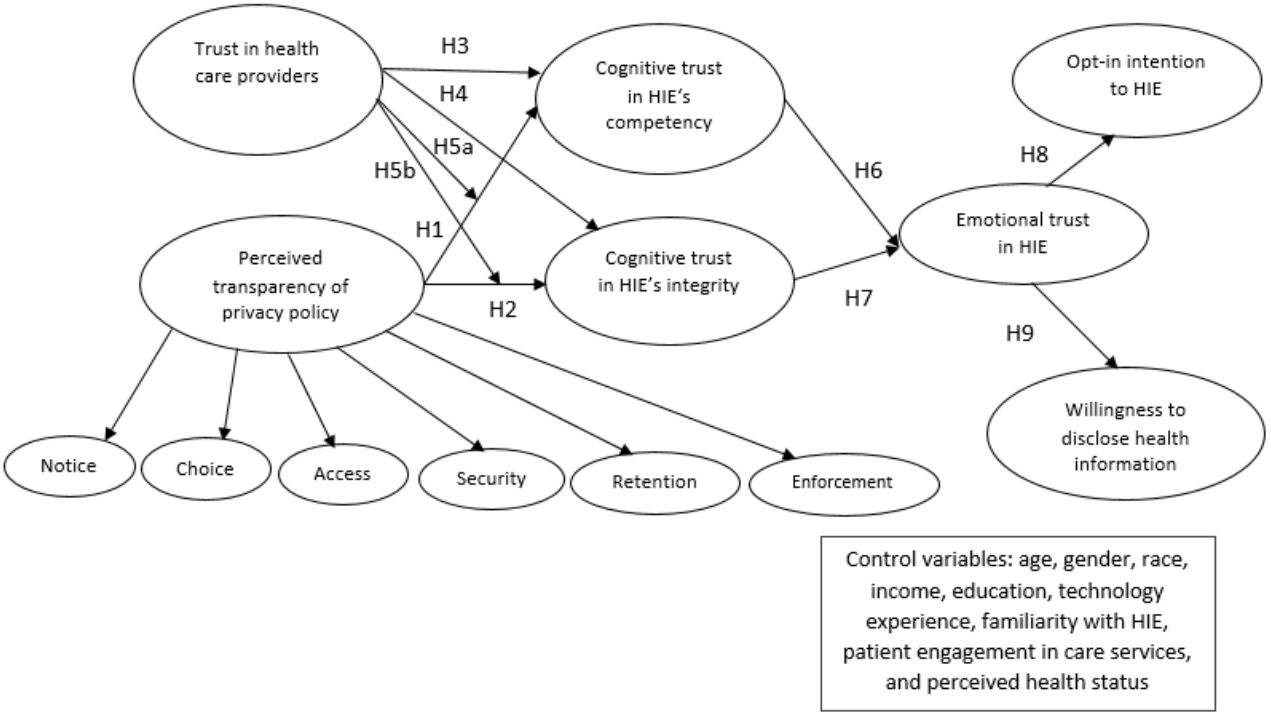
range of providers [11]. Therefore, as the main objective of this study was to investigate the opt-in intention and information disclosure willingness of individuals who have experience with HIE, the proposed model focused on attitude and not subjective norms.

Research Model

The following research model (Figure 1), which is mainly based on a belief-attitude-intention framework, explains the causal relationships. The links begin with perceived transparency of privacy policy and trust in health care providers (perceptions) to cognitive trust (trusting beliefs) and emotional trust (trusting attitude), and finally ends with opt-in intention to HIE as well as willingness to distribute health information (trusting intention). In this study, we focused on the intention rather than the adoption behavior, as there is a solid evidence in the IS literature that shows that intention is a strong predictor of behavior [44].

In this model, cognitive trust in competence and integrity is considered as beliefs, and emotional trust is conceptualized as an attitude. Patients may believe that the HIE is competent and honest in sharing their health information based on firm rational reasons. Emotional trust plays the role of attitude toward the HIE adoption behavior, as it is an evaluative affect (ie, feeling secure) about trusting in HIE.

Figure 1. Research model. H: hypothesis; HIE: health information exchange.



Hypotheses Development

Consumers' concerns in medical practices include high volume of collected health information, the possibility of privacy violations (eg, unauthorized access or hacked personal data), secondary use of medical records (eg, datamining purposes), lack of control over how medical records are collected, and how such information will be used [45-47]. Information privacy concerns may influence the validity and completeness of HIEs' patient databases, which may result in wasteful investment, inaccurate treatments, erroneous care planning, and higher mortality rates [48]. To avoid such issues, HIE networks should assure patients that their medical records would be well protected. Privacy issues will influence consumer beliefs about HIE initiatives. The degree of trust between patients and HIE efforts may attenuate the information privacy concerns.

According to Dimitropoulos and Rizk [49], privacy concern is defined by the extent to which the health care entities (eg, providers and organizations) could access, view, and share patient health information without obtaining a permission or consent. A factor that may mitigate privacy concerns related to information exchange efforts and help form patient trust is the transparency of privacy policies. Thus, privacy policies should be clearly presented by the health care organizations to build patient trust in the HIE's competence in protecting sensitive health information. The main objectives of the privacy policies are to enhance the understanding of how health information will be used inside or outside the organizations and decrease the concern that personal health information may be subject to improper access and would be used for unanticipated purposes [50]. The risk of information privacy misuse or unauthorized access highlights the importance of trust development before disclosing personal information. Previous studies emphasize that patients are concerned about losing control over the ways HIE systems handle their health information [51]. This concern mainly arises because of the lack of transparency of HIE information practices and policies. One of the best ways to address privacy concern and increase patient trust is through building a privacy policy with complete and transparent dimensions to clearly declare security tools and protection safeguards [27]. Comprehensible privacy policies should be developed by HIE initiatives to reduce the negative effects of

XSL•FO
RenderX

information privacy concerns and improve patients' cognitive trust in the HIE's technical competence. The dimensions and principles included in the policies should be informative and transparent enough to be able to advance patients' awareness of HIE data collection policies and information sharing practices. The more transparent the privacy policies are, the more they are likely to be reviewed and comprehended by patients, and only under this circumstance, patients are more willing to trust HIE's technical abilities to protect health information.

Nowadays, patients are very likely to seek medical treatments and care services from different physicians and providers. HIE systems provide networks in which patients' medical records are shared with a number of health care entities that are geographically scattered and use different privacy policies. In general, interoperable systems of data sharing between health care organizations are capable of improving completeness, reliability, and accuracy of medical records, which, in turn, ameliorate public health [52]. According to O'Kane et al [53], patients perceive that if the privacy policy is transparent, the electronic exchange of information among the health care providers is a more convenient and cost-effective sharing method, compared with the traditional data sharing efforts (eg, mail, phone, and fax transmission). With a clearly defined privacy policy, patients trust that a complete and flawless body of authorized information is shared electronically among health care entities through HIEs, and this is likely to help physicians generate better medical treatments and prescribe accurate medications. If privacy policy is perceived transparent, patients can learn about how health information is electronically shared between providers, what types of exchange mechanisms (eg, direct and look-up) are utilized to complete the sharing process, what types of sensitive information will be exchanged through the HIE, who will access and use the shared information, and for how long the information will be available to the authorized users. This perception may heighten trust in the HIE competence and encourage patients to believe that HIE technology is a real expert system in information-sharing area. Therefore, patients will become more familiar with HIE's main functions and obtain a cognitive picture of the sharing procedures and security mechanisms associated with HIE. This cognitive map becomes a tool for them to facilitate their decisions to support the use of HIE by health care entities to improve care quality and reduce health care bills [54]. Therefore, the transparency of privacy policy dimensions is a sound and rational reason for the consumer to trust in the HIE's competence. Accordingly, the following hypothesis is proposed:

> *H1: Perceived transparency of HIE's privacy policy will positively influence cognitive trust in HIE's competence.*

Different industries have diverse levels of compliance because of the various degrees of confirmation requirements and the different levels of information sensitivity [55]. Organizations operating in the health care industry should satisfy a higher level of compliance because they deal with highly sensitive health information and medical reports. Thus, stricter policy guidelines are imposed on the industry sectors that process and handle highly sensitive personal information. HIE projects can take advantage of a transparent and accessible privacy policy to resolve concerns associated with data safety and potential misuse to win patient trust in HIE's integrity, which, in turn, leads to competitive advantage. Privacy policies should be comprehensive and transparent enough to address all principles mentioned in the Health Insurance Portability and Accountability Act [49].

Notice principle articulates what health information is collected and exchanged, what the purpose of data exchange is, how such information will be used internally, and whether patient data will be disclosed to third parties. Choice principle delineates the consent process and permission requirements. This dimension provides options to patients regarding the use of their health data and the disclosure of such records to other third-party entities. For instance, by relying on this dimension, either patients are able to limit the exchange of personal information or voluntarily disclose their medical data for research purposes. Access principle entails granting the right to patients to obtain, review, and amend their personal information to ensure data accuracy and completeness. Security principle implies the adoption of reasonable measures and technical security steps to protect health information from unauthorized access, improper use, loss, unapproved alteration, or unanticipated disclosure during data exchange processes. Retention principle clarifies the acceptable duration of keeping and processing shared health information by health care providers. This dimension articulates the reasonable steps to permanently delete shared personal data if it is no longer required for the consented purpose. Enforcement principle highlights the self-regulation, such as privacy seals, that informs the public that the exchange procedures correspond to the legal requirements to protect information privacy [56]. Thus, highly transparent principles of privacy policies are able to demonstrate how safe, reliable, and dependent an HIE is and, in turn, increase patients' cognitive trust in the HIE's integrity.

The integrity of an HIE is the extent to which the HIE system is perceived to be honest and unbiased in the process of data sharing. However, an HIE system may be designed to adhere to a set of principles that are not acceptable by the patients. For instance, an HIE might collect, share, and use the patient' personal information for purposes other than care provision without obtaining an authorization. Health information might have been shared with unauthorized entities for secondary use (such as marketing and research) [57]. Unauthorized third parties may illegally access patients' sensitive medical records through HIE procedures and use such information for data mining purposes [18]. An HIE system with transparent privacy policy dimensions will be more effective in encouraging patients to trust a safe and credible mechanism that shares health information with authorized entities for legitimate purposes. Clear privacy policy dimensions attached to an HIE are likely to convince patients that the right amount of health information will be shared with authorized health care providers to meet relevant clinical purposes that are useful for patients' treatments. Patients perceive that HIE systems that offer a more transparent privacy policy may tell the truth by fulfilling the agreed promises without any deviations. Thus, these HIE efforts would be better in line with consumers' clinical preferences and heighten trust in integrity. An HIE privacy policy that is highly

XSL•FO
RenderX

transparent to customers is perceived to be aligned with their health care–related expectations than any other party's preferences. Compared with an HIE with low transparency, an HIE network with higher privacy transparency may be perceived to employ reliable procedures to grant access only to authorized users and apply honest exchange procedures to share health information for legitimate purposes. These reliable characteristics will increase the patients' perceptions that the HIE's procedures are unbiased.

According to Meinert et al [58], familiarity with the privacy policy statements can reduce the amount of risks and concerns related to an organization. Familiarity with the privacy policy of an HIE project can help patients develop a body of knowledge about what procedures are likely to be conducted and what mechanisms will be used in the future to exchange health information. This trust-related knowledge can increase the predictability power of patients to anticipate the HIE functions. If patients experienced some wrongdoing, dishonest procedures, deceptive information collection practices, unauthorized access, or illegal secondary use supported by an HIE system, they may predict that relying on the HIE systems is not wise. Consequently, they will think that the HIE network will also remain dishonest and untruthful in exchanging health information in the future. Thus, HIE's transparent privacy policy will promote the patient's trust in the HIE's integrity.

> *H2: Perceived transparency of HIE's privacy policy will positively influence cognitive trust in HIE's integrity.*

According to trust transfer theory, the trust transfer is a cognitive process in which the trust in one entity influences attitudes toward another phenomenon [59]. Trust transfer process describes that trust in a channel may affect the attitude toward a product or service offered in the same channel (intrachannel effects). Moreover, trust in a channel can be transferred to another channel because of perceived connections between them (interchannel effects) [37]. On the basis of the trust transfer theory, consumer trust in internet payment services may affect the level of consumer trust in mobile-based payment services [60]. In e-commerce settings, a study shows that trust can be transferred from the established and reputable websites to the unknown ones because of their links [38]. According to Lee et al [61], customers' trust in an offline bank is transferred to its Web-based banking services and, in turn, influences perceived website satisfaction. Customers' trust built over time in brick-and-mortar retailers is positively related to their level of trust in Web-based transactions before they visited their website [62].

As mentioned by Shin et al [63], trusted relationships between patients and providers play an important role in the acceptance of health informatics services. As HIE technology is mainly used by the health care providers to share personal information, trust in providers can form rational expectations that an HIE will also be a reliable and trustworthy sharing means [18]. Trusted interactions with the health care providers involved in the treatment process implies that the health care professionals will leverage a reliable, competent, and dependable mechanism for information exchange across organizations [64]. Heightened levels of trust in the providers can result in higher trust in the HIE's technical capabilities and integrity because the patients may perceive that the providers will act in the best interest of patients with minimum risks [35]. Therefore, trust in health care providers may initiate a conscious calculation of HIE advantages by evaluating competency and reliability of exchange procedures that will be used to minimize privacy and security risks [65]. Consistent with the findings of the study by Tang et al [66], trusting relationships with providers lead to rational reasons to participate in information-sharing initiatives. If patients believe that they can rely on health care providers, they become more likely to reason that a sharing mechanism used by them is also reliable and competent [67].

In the context of our study, trust transfer can be a key factor in the HIE context where the transfer of consumers' cognitive trust to the HIE's competence and integrity will take place because of their trust accumulated over time in health care providers. On the basis of the provided discussions on the trust transfer process, we proposed that trust in reliable and dependable health care providers can positively affect patients' cognitive trust in the HIE's competence and integrity. Thus, we hypothesize the following:

> *H3: The level of trust a patient has in health care providers positively affects their cognitive trust in the HIE's competence.*
>
> *H4: The level of trust a patient has in health care providers positively influences their cognitive trust in the HIE's integrity.*

Trust building is a process of interactions between involved parties and technology [68]. According to Lu et al [60], the level of trust transferred from the internet to mobile payment services moderates the relationship between trust in mobile payment and customers' behavioral intention. In the context of this study, we can argue that trust in health care providers also impacts the way privacy policy perceptions establish cognitive trust in the HIE's competence and integrity. The level of trust in providers may change the direction of the path between perceived transparency of privacy policy and patients' cognitive trust in the HIE. We proposed that the transparent privacy policy of the HIE initiatives can positively affect a patient's cognitive trust in the HIE competence and integrity, but these relationships may be variable depending on the level of trust a patient has in health care providers. A transparent privacy policy presented by an HIE network may encourage patients to believe that the exchange project has required reliable characteristics to protect health information. However, the strength of privacy policy-cognitive trust link will change contingent on the levels of trusting relationship between patients and providers. Moreover, when poor trusting relationships are established with the health care providers in society, less transparency will be perceived from the HIE privacy policy, and patients will have less rational reasons to trust in the HIE. The privacy policy and cognitive trust relationship is improved when patients hold a trusting perception about health care providers. Thus, we propose that trusted interactions with health care providers reinforce the relationship between the perceived transparency of HIE privacy policy and the level of cognitive trust in the HIE. This helps us develop our next hypotheses as follows:

XSL•FO
RenderX

H5a: The level of trust a patient has in health care providers moderates the relationship between perceived transparency of privacy policy and cognitive trust in the HIE's competence.

H5b: The level of trust a patient has in health care providers moderates the relationship between perceived transparency of privacy policy and cognitive trust in the HIE's integrity.

Cognitive trust in the HIE is delineated by 2 dimensions: the rational expectations about the HIE's ability to fulfil its obligations (cognitive trust in competence) and the rational reasons associated with the reliability of the HIE principles (cognitive trust in integrity). Emotional trust is defined as a patient's comfort and security feelings about relying on the HIE to disseminate health information. Consistent with the findings of the study by Curtin et al [69], emotion is mostly evoked by cognition. A study by Komiak and Benbasat [39] highlights the positive relationship between cognitive and emotional trust. They suggest that if individuals analyze that a recommendation agent (such as a Web-based personalization technology) is logically reliable, they, in turn, become more likely to rely on it emotionally. Extrapolating from the previous studies to the HIE context, we can also argue that cognitive trust in the HIE' competence and integrity is conceptualized as a belief. On the basis of the cognitive trust in competence, patients believe that the HIE is trustworthy because it has the required technological underpinning and competent exchange mechanisms to share health information among providers effectively and efficiently. Consistent with the cognitive trust in integrity, patients believe that the HIE is dependable for sharing health information because it holds reliable principles, truthful sharing standards, and honest promises. Consistent with TRA, these beliefs can strongly affect the attitude of patients toward the HIE efforts. Emotional trust is conceptualized as an attitude [42]. Emotional trust refers to an affective evaluation and feelings of relying on a trustee (such as a technology). In the context of HIE adoption, the higher the level of cognitive trust (both competence and integrity) in the HIE, the stronger the feelings of assurance, security, and comfort about the behavior of relying on the HIE. Therefore, the following hypotheses are developed:

H6: Cognitive trust in the HIE's competence will positively influence emotional trust.

H7: Cognitive trust in the HIE's integrity will positively influence emotional trust.

In this study, 2 related but different constructs are considered as dependent variables. Opt-in intention toward the HIE is the extent to which a patient is willing to rely on the HIE as a useful and reliable technology to be used by the health care entities to disseminate information. Willingness to disclose health information is the extent to which an individual is likely to share his or her sensitive health-related information with the health care organizations, with the knowledge that such information may be exposed to other providers through HIE systems. These 2 constructs are related because both of them are intention-based concepts; the first one is connected to adopting a technology (opt-in intention) and the second one is associated with a volunteer behavior (information disclosure). Nevertheless, they are different. The former variable deals with the notion that whether consumers are comfortable with the idea of having their health information shared through HIEs and whether to allow providers to use the system (if they are provided with the choice in the near future). The latter factor is the predictor of information disclosure behavior when the HIE systems are implemented by health care organizations. As patients typically cannot adopt an HIE, they can form attitudes, beliefs, and emotions about the concept of participating in sharing efforts. Therefore, in this context, the use should be evaluated through perceptual measures rather than actual opt-in behavior. A patient's feelings of security and a strong sense of comfort about relying on an HIE network can increase the intention to opt in to the HIE system. Thus, emotional trust in the HIE can encourage patients to have their medical records shared with relevant entities.

Information disclosure intention indicates the willingness of the individuals to voluntarily reveal personal information about themselves to others [70]. Information disclosure intention has an important effect on sharing behaviors in different Web contexts (eg, e-commerce and Web-based health communities) [71]. In the HIE context, patients may be likely to disclose their information with providers participating in an HIE network in exchange for disease prevention, reduced health care costs, and more accurate and timely treatment suggestions. Previous studies highlight the importance of privacy and security concerns in the context of HIE implementation [49,54]. Patients will hold a positive attitude toward an HIE network when their health records are collected, stored, and exchanged confidentially [72]. According to Wright et al [73], if a patient's privacy and security needs related to a data exchange mechanism are not met, he or she will become more likely to hide further health information from health care providers. Favorable attitude toward an HIE system is a result of a solid match between the HIE mechanisms and security or privacy requirements [3]. In this study, emotional trust is conceptualized as an attitude toward the HIE. In the presence of emotional trust, individuals are assured about the security of an HIE network and the privacy of their sensitive information that may be shared through this exchange means in the future. Thus, a high level of emotional trust in an HIE (ie, feeling secure about HIE use) will increase patients' opt-in intention toward it. Moreover, we expect that patients holding a favorable attitude toward an HIE are more likely to disclose personal health information to providers using the HIE in their practice.

H8: Emotional trust will positively influence opt-in intention toward the HIE.

H9: Emotional trust will positively influence willingness to disclose health information.

Consistent with TRA, our model only proposes indirect relationships between perceptions (perceived transparency of privacy statement and trust in health care providers) and attitude (emotional trust) through beliefs (cognitive trust).

XSL•FO
RenderX

Methods

Measurement Development

This study drew on the existing literature to measure the constructs included in the model, and minor changes were made to the instrument to fit the HIE context. Items measuring opt-in behavioral intentions were adapted from the studies by Venkatesh et al [44] and Angst and Agarwal [11]. The scales used to measure cognitive trust in the HIE's competency, cognitive trust in the HIE's integrity, and emotional trust in the HIE were adapted from the studies conducted by Komiak and Benbasat [39] and Mpinganjira [41]. To measure the 6 dimensions of the perceived transparency of privacy policy (ie, notice, choice, access, security, retention, and enforcement), we adapted the items reported by Chua et al [56] and Wu et al [26]. In this study, the perceived transparency of privacy policy was measured as a reflective second-order construct with 6 dimensions. The rationale behind this measurement is that the perceived transparency is reflective of the 6 dimensions and the expected interactions among them. According to Kayhan [74], reflective modeling is a better option than formative when first-order factors are expected to interact, correlate, or share a common theme. Thus, interrelationships among these factors is an important component of measuring the perceived transparency. For instance, notice principle, which defines the purpose of data exchange and explains what information is shared, may be related to security dimension that defines the security safeguards used to protect such information and the data transmission process. To measure trust in health care providers, we adapted the items reported by Moon [65] and Gefen et al [36]. Finally, the items indicating willingness to disclose health information were adapted from Zhang et al [75].

Once the initial questionnaire was developed based on previous research, we used an expert judgment approach to enhance the content validity of the survey. To check for the completeness, accuracy, readability, and format of the survey, the questionnaire was sent to 7 experts who are well published in the field of health informatics and HIE. The content validity index testing was used to analyze the feedback and suggestions. In this approach, the team of experts indicated whether each item on a scale was congruent with (or relevant to) the construct. Then, the percentage of items deemed to be relevant for each expert was computed, and finally, the average of the percentages across experts was taken. The average congruency percentage (ACP) was 92, which was higher than the threshold of 90% [76]. Therefore, the ACP was considered acceptable for the survey used in this study. We then removed the marked ambiguous words and modified the questions based on the experts' suggestions to ensure that they were clear and easy to understand for potential participants. Before conducting the main study, we conducted a pilot test with 137 graduate students at a large southeastern university in the United States to ensure the reliability and validity of the instrument. The Cronbach alpha was computed for each construct (perceived transparency of privacy statement, alpha=.96; trust in health care providers, alpha=.88; cognitive trust in the HIE's competence, alpha=.85; cognitive trust in the HIE's integrity, alpha=.90; emotional trust in the HIE, alpha=.91; opt-in intention toward the HIE, alpha=.92; and willingness to disclose health information, alpha=.92). All Cronbach alpha values were above the cutoff point of 0.7, which indicated that the instrument was internally consistent [77]. This study used 5-point Likert scales, with anchors ranging from 1=strongly disagree to 5=strongly agree. The final measure items used in this study are listed in Multimedia Appendix 1.

Data Collection Procedure

Data were collected in June 2018 from Amazon's Mechanical Turk (MTurk) to obtain a representative group of subjects. As the HIE is still not considered as a routine technology for many individuals, to get more solid and reliable findings, we specified an additional qualification that individuals had to meet to participate in the survey. We defined a screening question to include only those individuals who had visited a health care provider participating in an HIE network. Thus, the participants were aware of the HIE efforts, and their health information was shared through an HIE project when they took part in this study's data collection. The incentive for participation was a monetary reward (US $3). At the beginning of the Web-based survey, a detailed description of the HIE technology was provided to ensure that respondents completely comprehended the context and purpose of the study. The respondents were then asked a question about their level of familiarity with HIEs. To capture the dynamic trust transfer process and double check on whether their experience with the HIE projects met our criteria, before answering the main survey questions, they were requested to describe why and how they were familiar with HIEs. In total, 517 individuals attempted the survey. The respondents' answers to the familiarity question were analyzed to detect the main reasons they were aware of the HIE. Almost 94.9% (491/517) of the respondents were familiar with HIEs through *visiting a (or multiple) doctor who participated in an HIE network*. The remaining 5.0% (26/517) were aware of HIEs because of other reasons such as *through the internet searching/social media*, *reading health care magazines/newspaper*, *friends/family*, and *working in health care*. As we only focused on individuals who were familiar with HIEs because of visiting providers that actually shared their information through HIE networks, 16 potential participants were discarded and 501 met this condition.

As mentioned in previous studies, a general concern in data collection is the potential lack of attention and random responses [78]. Consistent with other studies, we used *captcha* questions to prevent and identify careless, hurried, or haphazard answers [79]. On the basis of the answers to these questions, 8 responses were dropped. This ratio is similar to those reported in previous studies that used MTurk for data collection [80]. Thus, concerns that Web-based respondents might reply randomly or haphazardly to complete the survey quickly were alleviated. After excluding responses that failed the response quality questions, the final set of usable and valid responses contained 493 samples. Moreover, the average completion time was 15.3 min that given the number of questions in the survey, suggested respondents spent an acceptable amount of time completing it.

Then, participants were requested to complete the survey by answering questions regarding the last time a health care provider used an HIE network to share their health information

XSL•FO
RenderX

with other entities (such as other hospitals, physician practices, laboratories, pharmacies, primary care, and emergency department). To ensure that their experience was recent enough, and so, they were able to remember its details, they were asked to indicate how many times they visited a (or multiple) doctor participating in an HIE project and when the most recent one was. Respondents had visited a (or multiple) physician involved in an HIE effort an average of 4.32 times during the previous year, and the most recent experience ranged from 2 months to a week ago. Relying on these screening questions and figures, the final sample fitted the study objective, which was investigating the trust of individuals (who were experienced with an HIE through providers who shared their records using the HIE) in the HIE and their opt-in intention toward it.

When testing the research model in this study, we controlled for consumer demographics and contextual factors such as income, age, education, race, gender, general technology experience, perceived health status, and engagement in the health care service, which are found and tested by the previous research as important factors in the adoption of HIEs. Therefore, it could be argued that by controlling the effects of aforementioned variables, the opt-in intention toward the HIE and willingness to disclose health information will mainly be measured based on the elements of cognitive and emotional processes linked with the health consumers' beliefs and attitudes toward electronic data exchange.

Instrument Validation

To validate the survey instrument, we performed confirmatory factor analysis on all the constructs to assess the measurement model. To do so, International Business Machines Corporation SPSS Amos (version 22) was used to test convergent and discriminant validity. According to Gefen et al [81], convergent validity can be tested by examining the standardized factor loading, composite reliability, and the average variance extracted (AVE). Table 1 shows the results of convergent validity test. All values of composite reliabilities were more than the threshold value of 0.7, which highlighted that the reliability of constructs was adequate [82]. According to Hair et al [83], a factor loading of ≥0.7 is acceptable. In this study, all reported standardized factor loadings were >0.7. The AVE of each construct was calculated using standardized factor loadings. All reported values of the AVE were also >0.5, which met the minimum requirement [84]. These measures indicated that the convergent validity of the measurement model was acceptable.

We also tested the discriminant validity of the constructs (Table 2). All the diagonal values were >0.7 and exceeded the correlations between any pair of constructs [85]. Therefore, the result indicates that the model fulfills the requirements of discriminant validity, and we can assume that the model also has adequate discriminant validity.

Although the correlations among constructs were not very noticeable (eg, a correlation of 0.483 between cognitive trust in the HIE's competence and integrity), we checked for multicollinearity by computing the variance inflation factor (VIF) and tolerance values for the predictor variables. The resultant VIF values were between 1.385 and 1.831, which were below the cutoff value of 5, and the tolerance values were in the range of 0.546 and 0.722, which were greater than the threshold of 0.1 [77]. Thus, the multicollinearity is not an issue in this research.

XSL•FO
RenderX

Table 1. Results of convergent validity.

Construct and respective items	Standardized factor loading (>0.7)	Composite reliability (>0.7)	Average variance extracted (>0.5)
Perceived transparency of privacy policy			
Notice			
1	0.81	0.927	0.716
2	0.85	__[a]	—
3	0.86	—	—
4	0.87	—	—
5	0.84	—	—
Choice			
1	0.83	0.92	0.696
2	0.83	—	—
3	0.85	—	—
4	0.86	—	—
5	0.8	—	—
Access			
1	0.81	0.884	0.657
2	0.81	—	—
3	0.77	—	—
4	0.85	—	—
Security			
1	0.81	0.913	0.723
2	0.86	—	—
3	0.87	—	—
4	0.86	—	—
Retention			
1	0.83	0.916	0.731
2	0.87	—	—
3	0.84	—	—
4	0.88	—	—
Enforcement			
1	0.88	0.903	0.757
2	0.87	—	—
3	0.86	—	—
Trust in health care providers			
1	0.81	0.872	0.695
2	0.84	—	—
3	0.85	—	—
Cognitive trust in the competency of health information exchange			
1	0.75	0.875	0.638
2	0.81	—	—
3	0.86	—	—
4	0.77	—	—
Cognitive trust in the integrity of health information exchange			

XSL•FO
RenderX

Construct and respective items	Standardized factor loading (>0.7)	Composite reliability (>0.7)	Average variance extracted (>0.5)
1	0.82	0.916	0.687
2	0.86	—	—
3	0.86	—	—
4	0.83	—	—
5	0.77	—	—
Emotional trust in health information exchange			
1	0.87	0.932	0.775
2	0.88	—	—
3	0.9	—	—
4	0.87	—	—
Opt-in intention to health information exchange			
1	0.81	0.926	0.758
2	0.89	—	—
3	0.88	—	—
4	0.9	—	—
Willingness to disclose health information			
1	0.89	0.929	0.766
2	0.83	—	—
3	0.9	—	—
4	0.88	—	—

[a]Not applicable.

XSL•FO
RenderX

Table 2. Results of discriminant validity (the main diagonal elements in italics denote the square roots of the average variances extracted, and the off-diagonal values represent the correlation coefficients between the constructs).

Construct	NOT[a]	CHO[b]	ACC[c]	SEC[d]	RET[e]	ENF[f]	THP[g]	CTC[h]	CTI[i]	EMT[j]	INT[k]	WILL[l]
NOT	*0.846*		—	—	—	—	—	—	—	—	—	—
CHO	*0.372*	0.*834*	—	—	—	—	—	—	—	—	—	—
ACC	*0.421*	0.479	0.*810*	—	—	—	—	—	—	—	—	—
SEC	*0.467*	0.421	0.447	0.*850*	—	—	—	—	—	—	—	—
RET	*0.358*	0.393	0.459	0.464	0.*854*	—	—	—	—	—	—	—
ENF	*0.411*	0.373	0.387	0.378	3.764	0.*870*	—	—	—	—	—	—
THP	0.381	0.323	0.515	0.446	3.862	0.485	*0.833*	—	—	—	—	—
CTC	0.396	0.356	0.496	0.399	3.898	0.454	0.317	*0.798*	—	—	—	—
CTI	0.367	0.464	0.322	0.323	3.670	0.517	0.396	0.483	*0.828*	—	—	—
EMT	0.434	0.378	0.505	0.356	4.254	0.523	0.261	0.358	0.474	*0.880*	—	—
INT	0.418	0.446	0.478	0.229	3.941	0.545	0.368	0.335	0.382	0.324	*0.870*	—
WILL	0.359	0.399	0.497	0.315	0.553	0.505	0.372	0.380	0.375	0.306	0.490	*0.875*

[a]NOT: notice.

[b]CHO: choice.

[c]ACC: access.

[d]SEC: security.

[e]RET: retention.

[f]ENF: enforcement.

[g]THP: trust in health care providers.

[h]CTC: cognitive trust in the competency of health information exchange.

[i]CTI: cognitive trust in the integrity of health information exchange.

[j]EMT: emotional trust in health information exchange.

[k]INT: opt-in intention to health information exchange.

[l]WILL: willingness to disclose health information.

Results

Descriptive Statistics

Table 3 depicts respondents' characteristics. The demographic characteristics show that the majority of the respondents were male (272/493, 55.1%), white (369/493, 74.8%), with a full-time job (338/493, 68.6%), and had a Bachelor's degree (257/493, 52.2%). Over 70% of respondents were aged between 20 and 39 years, and around 28% of the sample was aged >40 years. Our sample could be a representative of the actual demographics of the HIE users, as this is consistent with the age distribution in previous studies on the HIE [54,86].

Respondents of this study were fairly familiar with the general (eg, the internet and computer) and health care (eg, health tracking apps, Web-based patient community, and personal health record) technologies. Moreover, they were healthy enough to participate in the Web-based survey and were relatively engaged in their own care. Table 4 shows characteristics such as respondents' technology background, health status, and levels of engagement in care. The descriptive statistics of constructs used in the conceptual model are shown in Table 5.

XSL•FO
RenderX

Table 3. Sample (N=493) characteristics.

Variable and respective categories	Percentage
Gender	
Male	55.1
Female	44.9
Age (years)	
<20	0.4
20-29	35.8
30-39	35.8
40-49	13.3
50-59	8
≥60	6.6
Annual household income (US $)	
<25,000	13.3
25,000-49,999	32.3
50,000-74,999	24.3
75,000-99,999	16.4
≥100,000	13.7
Education	
Less than high school	1.8
High school graduate	12.8
Some college	19.5
2-year degree	8.4
Bachelor's degree	52.2
Graduate degree	5.3
Employment status	
Employed—full time	68.6
Employed—part time	16.8
Unemployed	6.2
Retired	5.3
Student	3.1
Race or ethnicity	
White	74.8
African American	8.4
Asian	9.7
Hispanic	4.9
Mixed	2.2
Participation in a Web-based patient community	
Yes	52.2
No	47.8
Using a personal health record	
Yes	63.7
No	36.3

XSL•FO
RenderX

Table 4. Sample technology background, engagement level, and health status.

Variable	Descriptive statistics, mean score (SD)
Perceived health status	3.98 (0.783)
Computer skills	4.41 (0.668)
Comfortable with using computers	4.61 (0.680)
Comfortable with using the internet	4.69 (0.581)
Comfortable with using mobile devices or apps for health purposes	4.33 (0.947)
Patient commitment	3.72 (0.68)
Therapeutic alliance	3.48 (0.79)

Table 5. Descriptive statistics of constructs (all measures are 5-point scales, with anchors 1=strongly disagree and 5=strongly agree).

Constructs	Descriptive statistics	
	Mean (SD)	Variance
Notice	3.75 (0.97)	0.94
Choice	3.62 (0.99)	0.99
Access	3.53 (0.95)	0.91
Security	3.62 (0.99)	0.99
Retention	3.51 (1)	1
Enforcement	3.64 (1.05)	1.11
Trust in health care providers	3.76 (1.01)	1.02
Cognitive trust in HIE's[a] competency	3.64 (1.05)	1.11
Cognitive trust in HIE's integrity	3.75 (0.97)	0.94
Emotional trust in HIE	3.62 (0.99)	0.99
Opt-in intention to HIE	3.64 (1.05)	1.11
Willingness to disclose health information	3.62 (1.05)	1.11

[a]HIE: health information exchange.

Control Variables

Factors that do not represent the core variables (ie, those included in the causal model) of this study, but which nevertheless may affect the interrelationships between the core variables, have been controlled for. These factors include age, gender, race, income, education, technology experience, familiarity with HIE, engagement in the care services, and perceived health status. Although the causal model seems to represent consumers' opt-in intention and determine their willingness to disclose health information, we found that the effects of control variables were not negligible. On the basis of the results, 2 dimensions of patient engagement in care (ie, patient commitment and therapeutic alliance) [87] directly influence both the HIE opt-in intention (beta=.204; $P<.01$) and disclosure willingness (beta=.127; $P<.05$) as contextual factors. This implies that factors that drive the patients to seek a greater understanding of their conditions will encourage them to opt in to the HIE and disclose health information. Moreover, the levels of patient's connection to the providers in the pursuit of care goals influence the trust building process and opt-in decision making. The findings also show that age (beta=−.141; $P<.01$), education level (beta=.112; $P<.05$), technology experience (beta=.132; $P<.01$), and HIE familiarity (beta=.247; $P<.001$) influence opt-in intention toward the HIE. These effects indicate that younger patients who are more familiar with the HIE networks and also have higher educational and technology experience backgrounds may have higher intentions toward the implementation of the HIE. Among the control variables, only education level affects the willingness to disclose health information (beta=.186; $P<.01$), meaning, individuals with higher levels of education are more likely to share their personal health information with providers. In contrast, no effects of gender, race, income, and perceived health status were found on both opt-in intention and willingness to disclose health data.

Structural Model

International Business Machines Corporation SPSS Amos (Version 22) was used to test the hypotheses within a structural equation modeling [88] framework. According to Ho [89], the goodness of fit statistics can evaluate the entire structural model and assess the overall fit. The findings indicated that the value of chi-square divided by degree of freedom for the model was $X^2/df=3507.2/1563=2.2$. The index values for confirmatory fit (0.914), normed fit (0.921), relative fit (0.923), and Tucker-Lewis (0.936) indices were above 0.9 and the

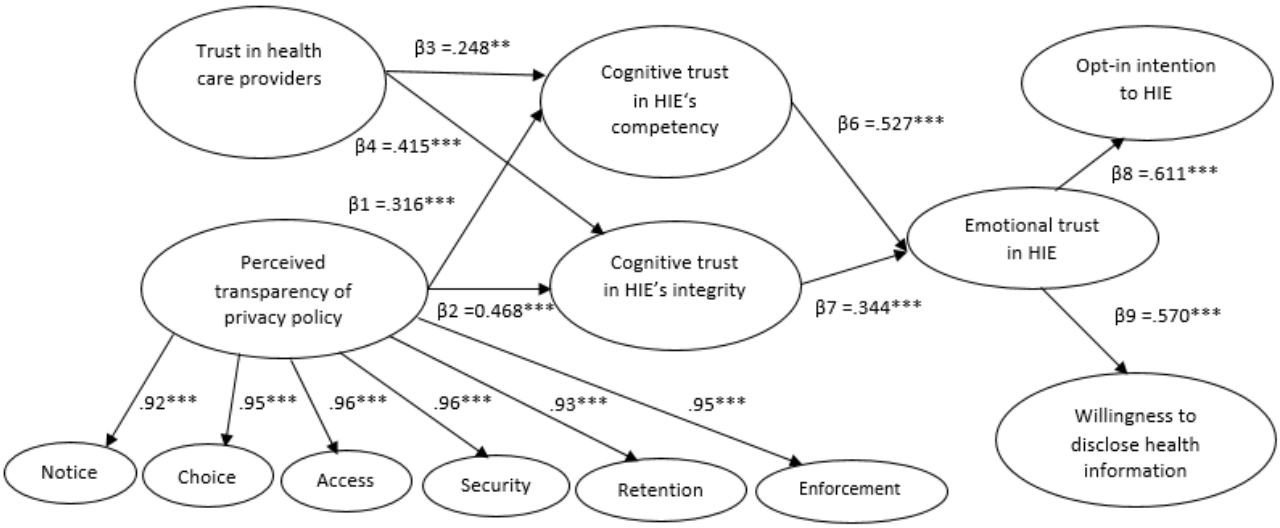
XSL•FO
RenderX

standardized root mean square residual (0.035) and root mean square error of approximation (0.047) were below 0.08 [90]. All these measures of fit were in the acceptable range, and only goodness of fit index (GFI; 0.851) and adjusted GFI (0.822) were marginal. On the basis of the study by Kline [91], at least 4 of the statistical values met the minimum recommended values, which supported a good fit between the hypothesized model and the observed data.

The results show that the perceived transparency of privacy policy is more accurately modeled and measured in the context of HIE as a second-order construct with 6 factors (ie, notice, choice, access, security, retention, and enforcement). The expectation of interactions is confirmed by the presence of significant positive correlations between the 6 dimensions. Moreover, the path values of the 6 indicators (notice: 0.92; choice: 0.95; access: 0.96; security: 0.96; retention: 0.93; and enforcement: 0.95) are significant (P<.001). Figure 2 displays the standardized path coefficients of the structural model under investigation and depicts the significant predictors of patients' opt-in intentions toward HIE and willingness to share health information.

The results of hypotheses testing are summarized in Table 6. The findings provide enough evidence to support H1, which indicates that the perceived transparency of privacy policy significantly increases cognitive trust in the HIE's competence (beta=.31; P<.01). The analysis also demonstrates that the perceived transparency of privacy policy is a significant antecedent of cognitive trust in the HIE's integrity (beta=.46; P<.001), and this positive linkage supports H2. Moreover, the R^2 scores for the 2 types of cognitive trust are 0.49 (cognitive trust competency) and 0.58 (cognitive trust in integrity), respectively. The results support H3 by showing the significant positive relationship between the trust in health care providers and the cognitive trust in the HIE's competence (beta=.24; P<.01). H4 is also supported where the higher level of trust in health care providers leads to higher cognitive trust in the HIE's integrity (beta=.41; P<.001). H6 argues the existence of a positive relationship between cognitive trust in the HIE's competence and emotional trust in the HIE, which is supported by the statistics (beta=.52; P<.001). Support is also found for H7, with cognitive trust in the HIE's integrity significantly affecting emotional trust in the HIE (beta=.34; P<.01). Furthermore, the R^2 score for emotional trust is 0.48. The findings provide solid evidence to support H8 by indicating that the higher the emotional trust in the HIE, the more likely patients are to allow health care providers to electronically exchange their health information using the HIE networks (beta=.61; P<.001). In addition, the positive effect of emotional trust to entice patients to disclose their health information is significant, supporting H9 (beta=.57; P<.001). Finally, the R^2 scores for opt-in intention and willingness to disclose health information are 0.56 and 0.51, respectively, reflecting that the model provides relatively strong explanatory power to predict the variance in the patients' willingness to release their health data and their intentions to opt in to HIE systems.

Figure 2. Model paths (**P<.01; ***P<.001). β: beta value; HIE: health information exchange.



XSL•FO
RenderX

Table 6. Results of hypotheses testing (all results supported the hypotheses).

Hypothesis	Path	Standardized coefficient	SE	Critical ratio
H1	PTPP[a] to CTC[b]	0.316[c]	0.077	11.265
H2	PTPP to CTI[d]	0.468[e]	0.074	11.152
H3	THP[f] to CTC	0.248[c]	0.063	4.649
H4	THP to CTI	0.415[e]	0.057	6.991
H6	CTC to EMT[g]	0.527[e]	0.066	6.806
H7	CTI to EMT	0.344[c]	0.082	8.195
H8	EMT to opt-in intention to health information exchange[h]	0.611[e]	0.062	12.669
H9	EMT to willingness to disclose health information[i]	0.570[e]	0.063	13.579

[a]PTPP: perceived transparency of privacy policy.

[b]CTC: cognitive trust in the competency of health information exchange (R^2=0.49).

[c]*P*<.01

[d]CTI: cognitive trust in the integrity of health information exchange (R^2=0.58).

[e]*P*<.001

[f]THP: trust in health care provider.

[g]EMT: emotional trust in health information exchange (R^2=0.48).

[h]R^2=0.56

[i]R^2=0.51

Moderating Effect of Trust in Health Care Providers

The moderating effect of trust in health care providers on the paths of perceived transparency of privacy policy to cognitive trust in the HIE's competence and cognitive trust in the HIE's integrity are significant at .05. To further interpret the interactions, separate regression analyses were conducted for subgroups of the sample. According to the approach of 1 standard deviation below and above the mean [89], the sample was split into 2 subgroups: low provider trust and high provider trust. Then, the relationship between perceived transparency of privacy policy and cognitive trust in the HIE's competence was regressed for each subgroup. The same analysis was conducted on the path of perceived transparency of privacy policy to cognitive trust in the HIE's integrity. The moderating test results indicate that the positive relationship between perceived transparency of privacy policy and cognitive trust in the HIE's competence is stronger among those who have well-established trusting relationships with health care providers. The findings also reveal that there is a significant difference in the relationship between perceived transparency of privacy policy and cognitive trust in the HIE's integrity between high versus low provider trust subgroups (critical ratio of difference between the 2 groups=2.51; *P*<.05). This finding supports our hypotheses (ie, H5a and H5b) that indicate that the relationship between perceived transparency of privacy policy and cognitive trust in the HIE's competence and the linkage of perceived transparency of privacy policy with cognitive trust in the HIE's integrity are positively moderated by the trust level in health care providers.

Discussion

Principal Findings and Implications for Research

The main findings of this study indicate that trust in health care providers and perceived transparency of privacy policy are significant predictors of cognitive trust in the HIE's competence and integrity. Our study also shows that the levels of trust in health care providers can moderate the relationships between perceived transparency of privacy policy and dimensions of cognitive trust in the HIE. Therefore, trusting relationships between patients and providers can strengthen the effects of privacy policy on cognitive trust in the HIE networks. Moreover, the results demonstrate that cognitive trust in the HIE can improve emotional trust in the HIE projects. Finally, emotional trust in the HIE efforts can encourage patients to opt in to HIE projects and become more willing to disclose their personal health information.

This research contributes several implications for theory. First, as trust has been proposed by previous studies as an important variable in the context of HIE rollout [92], it is crucial to add, and empirically test, different dimensions of trust in the HIE adoption research stream. In this study, the cognitive trust in competence, cognitive trust in integrity, and emotional trust are separated to provide a more complete evaluation of trust effects on patients' intentions to endorse the use of HIE and its possible impacts on their information disclosure decisions. In the context of HIE, previous studies mostly consider trust as trusting beliefs [18,35]. To offer more comprehensive insights, both cognitive and emotional dimensions of patient trust in the HIE are investigated. Our proposed model posits that cognitive trust is reflected as beliefs and emotional trust is conceptualized as attitude. In line with TRA and previous studies that use this

XSL•FO
RenderX

theory in other contexts, the findings show that cognitive trust in the HIE's competence and integrity (beliefs) significantly influences emotional trust in the HIE (attitude). This is consistent with psychology studies [69] suggesting that emotion is triggered by cognition, and it directly influences decision-making process.

This study differentiates between cognitive and emotional trust to contribute to the understanding of the process of patient trust formation in the HIE context. Cognitive trust in HIE is delineated by 2 dimensions: the rational expectations about the HIE's ability to fulfill its obligations (cognitive trust in competence) and the rational reasons associated with the reliability of the HIE principles (cognitive trust in integrity). Emotional trust is defined as a patient's comfort and security feelings about relying on the HIE to disseminate health information. This study also indicates that perceived transparency of privacy policy is able to resolve uncertainty associated with information-sharing processes, advance patient awareness of the HIE, and generate knowledge about how it operates. Then, patients' interpretation of their knowledge will directly affect cognitive trust in competence and integrity. In line with the findings of the study by Kahn et al [93], rational expectations and reasons that the HIE has the necessary characteristics to be relied upon will affect the degree to which patients feel in control, secure, and comfortable (emotional trust) about relying on the HIE to share their personal health information. On the basis of the significant relationships described in the model, when a patient is cognitively and emotionally involved with the HIE system, and trust is formed, he or she becomes more likely to disclose health information and allow health care providers to leverage this technology to share such information electronically with other health care parties.

Second, we draw upon the trust transfer theory to explain how patients' trusting beliefs in the HIE are formed in the context of HIE. This study examines trust transfer as a salient means of establishing initial trust in the HIE initiatives. The model developed in this study highlights how trust in health care providers (as the main users of HIE) is migrated to patients' cognitive and emotional trust in the HIE and how these trust factors will influence patients' opt-in decisions and their willingness to share personal health information. The results are consistent with the findings of the study by Lin et al [37], indicating that before a consumer accepts a technology, his or her past encounters and experiences may influence his or her beliefs about the new technology. The significant impact of trust in health care providers on the cognitive trust levels in the HIE's competence and integrity provides an empirical evidence on a dynamic of trust transfer process between health care professionals and health-related technologies (such as the HIE systems). This implies that well-established trust in health care providers strongly transfers to patients' cognitive trust in a technology designed to share sensitive health information across health care entities. Visiting reliable and responsible health care providers can lead to a rational process in which a patient expects that an HIE network is designed in a way to own the required features that are dependable. The significant interaction relationships between trust in health care providers and cognitive trust in the HIE's competence and integrity further validate the important role of previous experience with trusted providers in establishing positive beliefs and attitudes toward the HIE initiatives. The proposed model also highlights the importance of trust transfer mechanism in building cognitive trust in the HIE, emotional trust in the HIE, adoption behavior of patients, and their willingness to share their health information in the future.

Third, we figure out 3 factors to understand the trust transfer process: trust in source (trust in health care providers), trust in target (trust in the HIE), and the relationships between the source and the target (formulating a transparent privacy policy). According to the model, we propose that the trust in health care providers and the interactions between health care providers and the HIE (through the development of a transparent privacy policy to protect health information) can affect patient trust-building process. As patient trust in the HIE has become a critical factor for most of the HIE networks [94], the research on how this phenomenon is formed is of critical value. Our study delivers a comprehensive picture of the trust transfer process by highlighting the role of trust in source (providers), trust in target (HIE initiatives), and the interactions between source and target through creating a solid and comprehensive privacy policy for information exchange across various providers. In a study by Delgado-Ballester et al [95], perceived business tie is considered as the main predictor of trust in target. In contrast, Lin et al [37] argue that trust in source is the only key factor in the trust transfer process. Our study is an attempt to provide a full picture of trust transfer mechanism in the HIE context by indicating that both trust in health providers and the relationships between the providers and HIE can impact patients' trust in the HIE systems.

Our research identifies a factor to capture the interactions between the providers and HIE efforts. The perceived transparency of privacy policy is considered as the factor reflecting the tie and linkage between the providers and HIE networks. This enriches the trust transfer literature by showing that not only initial trust in providers can play an important role in developing patient trust in the HIE but also the beliefs that trusted providers will make a significant contribution to the development of a comprehensive and transparent privacy policy will meaningfully affect the patient trust-building process. This finding also implies that trust in health care providers has both direct and moderating effects on the cognitive trust in the HIE's competence and integrity. Relying on the moderating effect proposed in the model, our findings also answer the following question: will patients always trust an HIE system if a robust privacy policy with transparent components is provided and reliable security safeguards are leveraged to protect the health data in the information transmissions? The strong moderating effect of trust in health care providers on the relationships between the perceived transparency of privacy policy and the level of cognitive trust in the HIE's competence and integrity can address this question. This study provides empirical evidence that the greater levels of trust patients have in health care providers can reinforce their perceptions that a highly transparent and solid privacy policy is attached to the HIE

XSL•FO
RenderX

initiatives, and, in turn, their cognitive trust in the HIE will be improved.

Fourth, the results show that trust transfer factors have explained more than half of the variance in opt-in intention toward the HIE and individuals' willingness to disclose health information. Therefore, we can predict that trust transfer process is a strong explanatory mechanism to understand how patients' trust in the HIE efforts is built. We also believe that our model is not necessarily limited to the HIE but would be applicable to other technologies in the health care industry with similar characteristics, such as electronic health record and electronic prescribing systems. Finally, this study enriches the HIE literature by applying trust transfer theory to this research domain. Different from the traditional exchange mechanisms (such as mail or fax), the development of the HIE networks has largely pushed the information sharing among health care providers from conventional approaches to electronic exchange mechanisms. As patients are not the main users of the HIE systems and because of the distance imposed between patients and the actual users (providers), patients' past perceptions about health care providers may be transferred to the electronic exchange context. This study uses the trust transfer perspective to explain the theorization of the HIE adoption by capturing the dynamics of trust-building process.

Implications for Practice

There are also a number of important practical implications derived from this study. First, the significant role of trust in health care providers to predict cognitive trust in the HIE and the moderating effect of trust in providers in the relationship between privacy policy development and cognitive trust suggest that health care providers with good reputation can practically advance patient engagement in the HIE efforts. In contrast, patients are not likely to support and participate in the HIE efforts if these systems are developed or managed by providers with relatively poor reputation because of previous data breaches. Consistent with the findings of the study by Lu et al [60], consumers' initial lack of trust in health care providers can become a significant barrier in the implementation of the HIE projects. Thus, participations of trusted providers in the implementation of the HIE initiatives and contributions of reputable health care organizations to the development of comprehensive and transparent privacy policies should be highlighted in the HIE projects to win the cognitive and emotional trust of patients. Patient trust can be used as an enabler that allows a health care provider to expand from the traditional sharing methods to the HIE models (such as direct exchange or query-based exchange) [96]. Health care providers should look for opportunities to nurture their patients' trust in projects designed to exchange health information electronically. They should consider using tactics to increase the transparency and completeness of the HIE privacy policy and develop campaigns that leverage the power of image and reputation.

Second, HIE policy makers should establish a broad marketing strategy to enhance patients' perceptions about the accountability and accuracy of privacy policies, which can foster their trust in the HIE services. Research implications suggest that the HIE initiative managers should consider maximizing the transparency of privacy policy dimensions to induce consumers to read the privacy policy statements and make it a significant consideration in sharing personal information. The findings suggest the importance of educating consumers about the HIE mechanisms and sharing procedures to appeal to their cognitive and emotional trust. As our study shows the significant role of the perceived transparency of privacy policy in building cognitive trust in competence and integrity, a systematic strategy can be performed by health care entities to better demonstrate the dimensions of HIE's privacy statement. For instance, national educational programs, health conferences, and webinars that are easily accessible to a wide range of people can be administered to clearly publicize the key goals and policies of the national HIE efforts. Educational forums available on official health websites, Web-based tutorials accessible on patient portals or Web-based health communities, and computerized help programs can be used by health care organizations to improve the transparency of HIE efforts, broadcast their privacy policies, and increase public awareness on digital exchange mechanisms.

Third, according to the findings, the lack of public awareness about the expected benefits of HIE and the components of its privacy policy may impede the progress of sharing information between providers because of a lack of patients' cognitive and emotional trust in the HIE. This study suggests that both physicians and health care organizations (such as hospitals) can directly play an important role in persuading patients to give consent to sharing medical records using HIEs. Physicians' role may be more effective because they have face-to-face encounters with patients, and during consultations, they can enlighten the patients about the privacy policy of electronic sharing mechanisms. Hospitals can also influence how patients build trust in the HIE by educating them through brochures, leaflets, diagrams, and fact sheets that are comprehensible for an average person. These efforts should be able to clearly highlight why health information is shared, what types of information can be exchanged, how such information is shared from a point to another, what exchange mechanisms are used, who can access the medical data, what security safeguards will protect their records, and how often the transmission takes place.

Fourth, beside the educational programs designed to market expected benefits of the HIE for patients, the HIE administrators and health care organizations should attempt to meet patients' privacy policy expectations. According to the results, a comprehensive privacy statement that is able to address privacy policy requirements should have 6 related dimensions. The notice component should clearly state the type of health data collected and shared, specify the purposes of data exchange, identify any potential recipients of the data, explain how the shared personal information will be used, and indicate whether the exchange of the requested data is voluntary or required. The choice factor should provide transparent options for patients about how to put a limit on sharing personal information, give patients clear choice by asking for permission before disclosing health information to a third party, and clearly provide individual's choice of sharing health information under specific conditions (eg, in the case of emergency). The access component should describe whether individuals are able to access their

XSL•FO
RenderX

personal information and are able to correct inaccuracies in their personal information and state whether they have the right to delete their personal information from the HIE records. The security feature should clearly state the safeguards used to protect the data from unauthorized access and explain the required technology used to ensure cross-border data protection. The retention factor should clearly state the duration of keeping personal data, describe the time frame during which providers will access shared health information, and explain the reasonable approaches to ensure that the private health data are not kept longer than is necessary. Finally, the enforcement component should be clear enough to describe the actions that will be taken according to the law against who violate the privacy principles and provide a set of guidelines and enforcement mechanisms to assure that information sharing on the Web will abide by privacy laws.

Finally, the results show that the transparency of HIE privacy policies is important for patients, and the contents significantly affect their cognitive trust in the HIE competence and integrity. HIE privacy policies should not be merely prepared to meet legal requirements and protect health care entities from potential privacy lawsuits; most importantly, they should be able to address patients' privacy and security risks. Evidence suggests that in general, the contents of hospitals' privacy policies are prepared in a way that are not easily understandable by the majority of adults, and, in turn, patients are not usually interested in reading policy statements [97,98]. Presenting all required dimensions of privacy practices in the privacy policies, standardizing writing patterns, and improving the transparency of contents, along with the simplification of legal jargon, special expressions, and specialized language used to develop policy statements, can help patients better understand their rights and controls of their sensitive health information. HIE privacy statements should choose and focus on the contents that are able to resolve the most pressing privacy concerns. Administrators of the HIE initiatives can modify the contents and elements included in privacy policies based on the issues that rank high on their patients' concern list. The regulatory agencies can also play an important role by conducting educational workshops or training for the HIE organizations and health care providers on how to develop comprehensive privacy policies and running awareness campaigns to advance public understanding of information privacy and privacy practices of the HIE initiatives. By doing so, legal punishments and privacy violation penalties can be minimized and patients' trust in the HIE initiatives will be enhanced. Thus, the different dimensions and requirements of the HIE privacy policy point out the importance of health care providers' endeavor to prepare a detailed privacy policy framework for HIEs. The entities involved in the HIE efforts should also analyze the existing privacy statement's language, format, and wording to ensure the privacy policy clearly reflects the 6 components. The findings of this research would be useful for the HIE organizations to create robust, accessible, comprehensive, and transparent privacy statements for information exchange purposes to improve patient trust in the HIE initiatives.

Limitations and Future Research

Our research has some limitations that call for additional studies. The opportunity of further research on the development of patient trust in the HIE is rich, and we urge future studies to continue to test and improve the proposed model, theoretical logics, and hypotheses. The proposed model in this study may serve as a starting point in delineating the formation of patient trust in the HIE, and further research is required to investigate the processes of trust building and its different dimensions, antecedents, and outcomes. Consistent with previous research in other contexts [39], in this study, we separated the cognitive trust in the HIE competence from cognitive trust in the HIE integrity to better demonstrate the different role each dimension may play in the patient's decision-making process. Future studies can extend this model by measuring cognitive trust as a second-order construct. Moreover, another promising area for future research is to investigate the likely difference between patients' trust in individual providers and trust in health care organizations with large administrative systems to identify how this difference will influence the degree of cognitive and emotional trust in the HIE technology. Thus, the possible variance of trust transfer mechanism based on provider types can be further examined in future research.

In this study, we leveraged several measures and filters to recruit participants who were familiar with a regional, patient-centered HIE's functions and its privacy policy. However, as a self-rated sample of participants on MTurk was used, there is a small chance that some individuals were not completely aware of the HIE mechanisms and formed their own mental construal of the IT artifact. Therefore, we suggest that further studies use a different method to ensure that subjects are knowledgeable about the HIE efforts. For instance, future research can recruit informed patients who are directly referred by the providers participating in the HIE initiatives. Moreover, our study used a Web-based survey to recruit participants digitally. Thus, we only considered individuals who accessed the internet and were healthy enough to participate in the Web-based survey. Future studies can use other data collection means and sampling strategies to reach out to a sample that is generalizable to a wide range of health care consumers.

The variance of trust factors (cognitive and emotional) explained by the proposed model was around 50%. This implies that there are other variables that could be included in the model. Future research should examine other factors that may affect the trust-building process in the HIE context. Moreover, the intention to opt in to HIE and willingness to disclose health information may be affected by the sensitivity levels of health information. For instance, if a patient perceives that his or her health information is highly sensitive (eg, mental health information and sexual health diseases), he or she may prefer to hide it from health care providers and, in turn, become less likely to opt in to an HIE network that shares such personal information with other providers. Future studies can measure the possible effect of perceived health information sensitivity on the 2 behavioral intentions. In addition, this study does not focus on a specific group of patients with sensitive information (such as individuals living with HIV). Future research can extend this model to identify how the cognitive and emotional

XSL•FO
RenderX

trust will be developed in the context where health information is highly sensitive. Finally, the 2 behavioral intentions may follow a 2-stage model, in which intention to opt in to HIE may occur before the willingness to disclose health information. We recommend that future studies investigate the possible relationship between the 2 behavioral intentions and the impact of these intentions on actual patient behaviors.

Conclusions

Sharing personal information and dependence on a technology to conduct information exchange are a trust-related behavior. However, how patient trust develops in the HIE context, within which the HIE technology is implemented and used by health care organizations for sharing purposes, has not been clearly discussed yet. To fill this research gap, this study mainly draws upon the trust transfer mechanism to articulate patients' trust-building process. The cognitive trust in the HIE's competence and integrity, in addition to the emotional trust, is a fundamental part of patients' opt-in decisions and willingness to disclose their health information. The findings also show that patients' initial trust in health care providers and their perception about the role of trusted providers in the development of privacy policy are found significant to determine the opt-in intentions and willingness to share health information in the future. These results can contribute to trust transfer theory and enrich the literature on the HIE efforts. Practitioners can also identify how to leverage the benefit of patients' trust and trust transfer process to promote the HIE initiatives nationwide.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Online survey.
[PDF File (Adobe PDF File), 568 KB - medinform_v7i4e14050_app1.pdf]

References

1. Baptista G, Oliveira T. Understanding mobile banking: The unified theory of acceptance and use of technology combined with cultural moderators. Comp Human Behav 2015;50:418-430. [doi: 10.1016/j.chb.2015.04.024]
2. Lenert L, Sundwall D, Lenert ME. Shifts in the architecture of the Nationwide Health Information Network. J Am Med Inform Assoc 2012;19(4):498-502 [FREE Full text] [doi: 10.1136/amiajnl-2011-000442] [Medline: 22268218]
3. Campion TR, Edwards AM, Johnson SB, Kaushal R, HITEC investigators. Health information exchange system usage patterns in three communities: practice sites, users, patients, and data. Int J Med Inform 2013 Sep;82(9):810-820. [doi: 10.1016/j.ijmedinf.2013.05.001] [Medline: 23743323]
4. Adler-Milstein J, McAfee AP, Bates DW, Jha AK. The state of regional health information organizations: current activities and financing. Health Aff (Millwood) 2008;27(1):w60-w69. [doi: 10.1377/hlthaff.27.1.w60] [Medline: 18073225]
5. Dixon BE, Zafar A, Overhage JM. A framework for evaluating the costs, effort, and value of nationwide health information exchange. J Am Med Inform Assoc 2010;17(3):295-301 [FREE Full text] [doi: 10.1136/jamia.2009.000570] [Medline: 20442147]
6. Frisse ME, Johnson KB, Nian H, Davison CL, Gadd CS, Unertl KM, et al. The financial impact of health information exchange on emergency department care. J Am Med Inform Assoc 2012;19(3):328-333 [FREE Full text] [doi: 10.1136/amiajnl-2011-000394] [Medline: 22058169]
7. Devine E, Totten A, Gorman P, Eden K, Kassakian S, Woods S, et al. Health information exchange use (1990-2015): a systematic review. EGEMS (Wash DC) 2017 Dec 7;5(1):27 [FREE Full text] [doi: 10.5334/egems.249] [Medline: 29881743]
8. Eden KB, Totten AM, Kassakian SZ, Gorman PN, McDonagh MS, Devine B, et al. Barriers and facilitators to exchanging health information: a systematic review. Int J Med Inform 2016 Apr;88:44-51 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.01.004] [Medline: 26878761]
9. Cundy PR, Hassey A. To opt in or opt out of electronic patient records? Isle of Wight and Scottish projects are not opt out schemes. Br Med J 2006 Jul 15;333(7559):146 [FREE Full text] [doi: 10.1136/bmj.333.7559.146] [Medline: 16840483]
10. Wilkinson J. Patients should have to opt out of national electronic care records: what's all the fuss about? Br Med J 2006 Jul 1;333(7557):42-43 [FREE Full text] [doi: 10.1136/bmj.333.7557.42] [Medline: 16809715]
11. Angst CM, Agarwal R. Adoption of electronic health records in the presence of privacy concerns: the elaboration likelihood model and individual persuasion. Manag Inf Syst Q 2009;33(2):339-370. [doi: 10.2307/20650295]
12. Komiak SX, Benbasat I. Understanding customer trust in agent-mediated electronic commerce, web-mediated electronic commerce, and traditional commerce. Inform Technol Manage 2004;5(1/2):181-207. [doi: 10.1023/b:item.0000008081.55563.d4]
13. McKnight DH, Choudhury V, Kacmar C. Developing and validating trust measures for e-commerce: an integrative typology. Inform Sys Res 2002 Sep;13(3):334-359. [doi: 10.1287/isre.13.3.334.81]
14. Platt J, Kardia S. Public trust in health information sharing: implications for biobanking and electronic health record systems. J Pers Med 2015 Feb 3;5(1):3-21 [FREE Full text] [doi: 10.3390/jpm5010003] [Medline: 25654300]

XSL•FO
RenderX

15. Lu B, Zhang T, Wang L, Keller LR. Trust antecedents, trust and online microsourcing adoption: an empirical study from the resource perspective. Dec Supp Sys 2016 May;85:104-114. [doi: 10.1016/j.dss.2016.03.004]
16. Schoorman F, Mayer R, Davis JH. An integrative model of organizational trust: past, present, and future. Acad Manag Rev 2007;32(2):344-354. [doi: 10.5465/amr.2007.24348410]
17. Shen N, Bernier T, Sequeira L, Strauss J, Silver MP, Carter-Langford A, et al. Understanding the patient privacy perspective on health information exchange: a systematic review. Int J Med Inform 2019 May;125:1-12. [doi: 10.1016/j.ijmedinf.2019.01.014] [Medline: 30914173]
18. McGraw D, Dempsey JX, Harris L, Goldman J. Privacy as an enabler, not an impediment: building trust into health information exchange. Health Aff (Millwood) 2009;28(2):416-427. [doi: 10.1377/hlthaff.28.2.416] [Medline: 19275998]
19. Tripathi M, Delano D, Lund B, Rudolph L. Engaging patients for health information exchange. Health Aff (Millwood) 2009;28(2):435-443. [doi: 10.1377/hlthaff.28.2.435] [Medline: 19276000]
20. Kim H, Chan H, Chan Y, Gupta S. Understanding the balanced effects of belief and feeling on information systems continuance. In: Proceedings of the 2004 International Conference on Information Systems. 2004 Presented at: ICIS'04; December 12-15, 2004; Washington, DC, USA p. 24.
21. Hu X, Lin Z, Whinston AB, Zhang H. Hope or hype: on the viability of escrow services as trusted third parties in online auction environments. Inform Sys Res 2004;15(3):236-249. [doi: 10.1287/isre.1040.0027]
22. Chopra K, Wallace WA. Trust in Electronic Environments. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track 9 - Volume 9. 2003 Presented at: HICSS'03; January 6-9, 2003; Big Island, HI, USA, USA p. 331.1. [doi: 10.1109/hicss.2003.1174902]
23. Esmaeilzadeh P. The effects of public concern for information privacy on the adoption of health information exchanges (HIEs) by healthcare entities. Health Commun 2019 Sep;34(10):1202-1211. [doi: 10.1080/10410236.2018.1471336] [Medline: 29737872]
24. Derbaix CM. The impact of affective reactions on attitudes toward the advertisement and the brand: A step toward ecological validity. J Mark Res 1995;32(4):470-479. [doi: 10.1177/002224379503200409]
25. Capistrano EP, Chen JV. Information privacy policies: the effects of policy characteristics and online experience. Comput Stand Inter 2015;42:24-31. [doi: 10.1016/j.csi.2015.04.001]
26. Wu K, Huang SY, Yen DC, Popova I. The effect of online privacy policy on consumer privacy concern and trust. Comput Human Behav 2012;28(3):889-897. [doi: 10.1016/j.chb.2011.12.008]
27. Callanan C, Jerman-Blažič B, Blažič AJ. User awareness and tolerance of privacy abuse on mobile internet: an exploratory study. Telemat Inform 2016;33(1):109-128. [doi: 10.1016/j.tele.2015.04.009]
28. Rifon N, LaRose R, Choi SM. Your privacy is sealed: effects of web privacy seals on trust and personal disclosures. J Consumer Aff 2005;39(2):339-362. [doi: 10.1111/j.1745-6606.2005.00018.x]
29. Egelman S, Tsai J, Cranor LF. Timing is Everything?: The Effects of Timing and Placement of Online Privacy Indicators. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2009 Presented at: CHI'09; April 4-9, 2009; Boston, MA, USA p. 319-328. [doi: 10.1145/1518701.1518752]
30. Tsai JY, Egelman S, Cranor L, Acquisti A. The effect of online privacy information on purchasing behavior: an experimental study. Inform Sys Res 2011 Jun;22(2):254-268. [doi: 10.1287/isre.1090.0260]
31. Steinfeld N. 'I agree to the terms and conditions': (How) do users read privacy policies online? An eye-tracking experiment. Comput Human Behav 2016;55:992-1000. [doi: 10.1016/j.chb.2015.09.038]
32. Rainie L, Madden M. Pew Research Center. 2015 Mar 16. Americans' Privacy Strategies Post-Snowden URL: https://www.pewinternet.org/2015/03/16/americans-privacy-strategies-post-snowden/ [accessed 2019-10-10]
33. Milne GR, Culnan MJ. Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices. J Interact Mark 2004;18(3):15-29. [doi: 10.1002/dir.20009]
34. Aïmeur E, Lawani O, Dalkir K. When changing the look of privacy policies affects user trust: An experimental study. Comput Human Behav 2016;58:368-379. [doi: 10.1016/j.chb.2015.11.014]
35. Dimitropoulos L, Patel V, Scheffler S, Posnack S. Public attitudes toward health information exchange: perceived benefits and concerns. Am J Manag Care 2011 Dec;17(12 Spec No):SP111-SP116 [FREE Full text] [Medline: 22216769]
36. Gefen D, Karahanna E, Straub DW. Trust and TAM in online shopping: an integrated model. Manag Inf Syst Q 2003;27(1):51-90. [doi: 10.2307/30036519]
37. Lin J, Lu Y, Wang B, Wei KK. The role of inter-channel trust transfer in establishing mobile commerce trust. Electron Commer R A 2011 Nov;10(6):615-625. [doi: 10.1016/j.elerap.2011.07.008]
38. Stewart KJ. Trust transfer on the world wide web. Organ Sci 2003;14(1):5-17. [doi: 10.1287/orsc.14.1.5.12810]
39. Komiak SX, Benbasat I. The effects of personalization and familiarity on trust and adoption of recommendation agents. Manag Inf Syst Q 2006;30(4):941-960. [doi: 10.2307/25148760]
40. Rempel JK, Holmes JG, Zanna MP. Trust in close relationships. J Pers Soc Psychol 1985;49(1):95-112. [doi: 10.1037//0022-3514.49.1.95]
41. Mpinganjira M. Precursors of trust in virtual health communities: a hierarchical investigation. Inform Manag 2018;55(6):686-694. [doi: 10.1016/j.im.2018.02.001]

XSL•FO
RenderX

42. Fishbein M, Ajzen I. Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research. Boston, USA: Addison Wesley; 1975.
43. Karahanna E, Straub DW, Chervany NL. Information technology adoption across time: a cross-sectional comparison of pre-adoption and post-adoption beliefs. Manag Inf Syst Q 1999;23(2):183-213. [doi: 10.2307/249751]
44. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. Manag Inf Syst Q 2003;27(3):425-478. [doi: 10.2307/30036540]
45. Agaku IT, Adisa AO, Ayo-Yusuf OA, Connolly GN. Concern about security and privacy, and perceived control over collection and use of health information are related to withholding of health information from healthcare providers. J Am Med Inform Assoc 2014;21(2):374-378 [FREE Full text] [doi: 10.1136/amiajnl-2013-002079] [Medline: 23975624]
46. Chen Y, Xu H. Privacy Management in Dynamic Groups: Understanding Information Privacy in Medical Practices. In: Proceedings of the 2013 conference on Computer supported cooperative work. 2013 Presented at: CSCW'13; February 23-27, 2013; San Antonio, Texas, USA p. 541-552. [doi: 10.1145/2441776.2441837]
47. Perera G, Holbrook A, Thabane L, Foster G, Willison DJ. Views on health information sharing and privacy from primary care practices using electronic medical records. Int J Med Inform 2011 Feb;80(2):94-101. [doi: 10.1016/j.ijmedinf.2010.11.005] [Medline: 21167771]
48. Whiddett R, Hunter I, Engelbrecht J, Handy J. Patients' attitudes towards sharing their health information. Int J Med Inform 2006;75(7):530-541. [doi: 10.1016/j.ijmedinf.2005.08.009] [Medline: 16198142]
49. Dimitropoulos L, Rizk S. A state-based approach to privacy and security for interoperable health information exchange. Health Aff (Millwood) 2009;28(2):428-434. [doi: 10.1377/hlthaff.28.2.428] [Medline: 19275999]
50. Frohlich J, Karp S, Smith MD, Sujansky W. Retrospective: lessons learned from the Santa Barbara project and their implications for health information exchange. Health Aff (Millwood) 2007;26(5):w589-w591. [doi: 10.1377/hlthaff.26.5.w589] [Medline: 17670777]
51. Kim KK, Joseph JG, Ohno-Machado L. Comparison of consumers' views on electronic data sharing for healthcare and research. J Am Med Inform Assoc 2015 Jul;22(4):821-830 [FREE Full text] [doi: 10.1093/jamia/ocv014] [Medline: 25829461]
52. O'Donnell HC, Patel V, Kern LM, Barrón Y, Teixeira P, Dhopeshwarkar R, et al. Healthcare consumers' attitudes towards physician and personal use of health information exchange. J Gen Intern Med 2011;26(9):1019-1026 [FREE Full text] [doi: 10.1007/s11606-011-1733-6] [Medline: 21584839]
53. O'Kane A, Mentis H, Thereska E. Non-static nature of patient consenthifting privacy perspectives in health information sharing. In: Proceedings of the 2013 conference on Computer supported cooperative work. 2013 Presented at: CSCW'13; February 23-27, 2013; San Antonio, Texas, USA p. 553-562. [doi: 10.1145/2441776.2441838]
54. Park H, Lee S, Kim Y, Heo E, Lee J, Park JH, et al. Patients' perceptions of a health information exchange: a pilot program in South Korea. Int J Med Inform 2013;82(2):98-107. [doi: 10.1016/j.ijmedinf.2012.05.001] [Medline: 22658777]
55. Li Y, Stweart W, Zhu J, Ni A. Online privacy policy of the thirty Dow Jones corporations: Compliance with FTC Fair Information Practice Principles and readability assessment. Commun IIMA 2012;12(3):5 [FREE Full text]
56. Chua HN, Herbland A, Wong SF, Chang Y. Compliance to personal data protection principles: a study of how organizations frame privacy policy notices. Telemat Inform 2017;34(4):157-170. [doi: 10.1016/j.tele.2017.01.008]
57. Grande D, Mitra N, Shah A, Wan F, Asch DA. Public preferences about secondary uses of electronic health information. JAMA Intern Med 2013 Oct 28;173(19):1798-1806 [FREE Full text] [doi: 10.1001/jamainternmed.2013.9166] [Medline: 23958803]
58. Meinert D, Peterson D, Criswell J, Crossland MD. Privacy policy statements and consumer willingness to provide personal information. J Electron Comm Organ 2006;4(1):1-17. [doi: 10.4018/jeco.2006010101]
59. Wang N, Shen X, Sun Y. Transition of electronic word-of-mouth services from web to mobile context: a trust transfer perspective. Dec Supp Sys 2013;54(3):1394-1403. [doi: 10.1016/j.dss.2012.12.015]
60. Lu Y, Yang S, Chau PY, Cao Y. Dynamics between the trust transfer process and intention to use mobile payment services: a cross-environment perspective. Inform Manag 2011 Dec;48(8):393-403. [doi: 10.1016/j.im.2011.09.006]
61. Lee KC, Kang I, McKnight DH. Transfer from offline trust to key online perceptions: an empirical study. IEEE Trans Eng Manage 2007;54(4):729-741. [doi: 10.1109/tem.2007.906851]
62. Kuan H, Bock G. Trust transference in brick and click retailers: an investigation of the before-online-visit phase. Inform Manag 2007;44(2):175-187. [doi: 10.1016/j.im.2006.12.002]
63. Shin D, Lee S, Hwang Y. How do credibility and utility play in the user experience of health informatics services? Comput Human Behav 2017;67:292-302. [doi: 10.1016/j.chb.2016.11.007]
64. Kuperman GJ. Health-information exchange: why are we doing it, and what are we doing? J Am Med Inform Assoc 2011;18(5):678-682 [FREE Full text] [doi: 10.1136/amiajnl-2010-000021] [Medline: 21676940]
65. Moon LA. Factors influencing health data sharing preferences of consumers: a critical review. Heal Policy Technol 2017;6(2):169-187. [doi: 10.1016/j.hlpt.2017.01.001]
66. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 2006;13(2):121-126 [FREE Full text] [doi: 10.1197/jamia.M2025] [Medline: 16357345]

XSL•FO
RenderX

67. Esmaeilzadeh P. Healthcare consumers' opt-in intentions to Health Information Exchanges (HIEs): an empirical study. Comput Human Behav 2018;84:114-129. [doi: 10.1016/j.chb.2018.02.029]
68. Montague EN, Winchester WW, Kleiner BM. Trust in medical technology by patients and health care providers in obstetric work systems. Behav Inf Technol 2010 Sep;29(5):541-554 [FREE Full text] [doi: 10.1080/01449291003752914] [Medline: 20802836]
69. Curtin JJ, Patrick CJ, Lang AR, Cacioppo JT, Birbaume N. Alcohol affects emotion through cognition. Psychol Sci 2001;12(6):527-531. [doi: 10.1111/1467-9280.00397] [Medline: 11760143]
70. Lowry PB, Cao J, Everard A. Privacy concerns versus desire for interpersonal awareness in driving the use of self-disclosure technologies: the case of instant messaging in two cultures. J Manag Inform Sys 2011;27(4):163-200. [doi: 10.2753/mis0742-1222270406]
71. Dinev T, Hart P. An extended privacy calculus model for e-commerce transactions. Inform Syst Res 2006;17(1):61-80. [doi: 10.1287/isre.1060.0080]
72. Abdulnabi M, Al-Haiqi A, Kiah M, Zaidan A, Zaidan B, Hussain M. A distributed framework for health information exchange using smartphone technologies. J Biomed Inform 2017 May;69:230-250 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.013] [Medline: 28433825]
73. Wright A, Soran C, Jenter CA, Volk LA, Bates DW, Simon SR. Physician attitudes toward health information exchange: results of a statewide survey. J Am Med Inform Assoc 2010;17(1):66-70 [FREE Full text] [doi: 10.1197/jamia.M3241] [Medline: 20064804]
74. Kayhan VO. The nature, dimensionality, and effects of perceptions of community governance. Inform Manag 2015;52(1):18-29. [doi: 10.1016/j.im.2014.10.004]
75. Zhang X, Liu S, Chen X, Wang L, Gao B, Zhu Q. Health information privacy concerns, antecedents, and information disclosure intention in online health communities. Inform Manag 2018;55(4):482-493. [doi: 10.1016/j.im.2017.11.003]
76. Polit DF, Beck CT. The content validity index: are you sure you know what's being reported? Critique and recommendations. Res Nurs Health 2006;29(5):489-497. [doi: 10.1002/nur.20147] [Medline: 16977646]
77. Hair JF, Ringle CM, Sarstedt M. PLS-SEM: indeed a silver bullet. J Market Theory Pract 2011;19(2):139-152. [doi: 10.2753/mtp1069-6679190202]
78. Huang JL, Curran PG, Keeney J, Poposki EM, DeShon RP. Detecting and deterring insufficient effort responding to surveys. J Bus Psychol 2012;27(1):99-114. [doi: 10.1007/s10869-011-9231-8]
79. Mason W, Suri S. Conducting behavioral research on Amazon's mechanical turk. Behav Res Methods 2012;44(1):1-23. [doi: 10.3758/s13428-011-0124-6] [Medline: 21717266]
80. Boyer O'Leary M, Wilson JM, Metiu A. Beyond being there: the symbolic role of communication and identification in perceptions of proximity to geographically dispersed colleagues. Manag Inf Syst Q 2014;38(4):1219-1243. [doi: 10.25300/misq/2014/38.4.13]
81. Gefen D, Straub D, Boudreau M. Structural equation modeling and regression: guidelines for research practice. Commun Assoc Inform Sys 2000;4(1):7. [doi: 10.17705/1cais.00407]
82. Chin WW. The partial least squares approach to structural equation modeling. In: Marcoulides GA, editor. Modern Methods for Business Research. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers; 1998:295-336.
83. Hair J, Black W, Babin B, Anderson R, Tatham R. Multivariate Data Analysis. Sixth Edition. London, England: Pearson; 2006.
84. Segars AH. Assessing the unidimensionality of measurement: a paradigm and illustration within the context of information systems research. Omega 1997;25(1):107-121. [doi: 10.1016/s0305-0483(96)00051-5]
85. Fornell C, Tellis G, Zinkhan GM. Validity assessment: A structural equations approach using partial least squares. Chicago, IL, USD: Proceedings, american marketing association educators? conference; 1982.
86. Patel VN, Dhopeshwarkar RV, Edwards A, Barrón Y, Sparenborg J, Kaushal R. Consumer support for health information exchange and personal health records: a regional health information organization survey. J Med Syst 2012 Jun;36(3):1043-1052. [doi: 10.1007/s10916-010-9566-0] [Medline: 20703633]
87. Higgins T, Larson E, Schnall R. Unraveling the meaning of patient engagement: a concept analysis. Patient Educ Couns 2017 Jan;100(1):30-36. [doi: 10.1016/j.pec.2016.09.002] [Medline: 27665500]
88. Thackeray R, Crookston BT, West JH. Correlates of health-related social media use among adults. J Med Internet Res 2013 Jan 30;15(1):e21 [FREE Full text] [doi: 10.2196/jmir.2297] [Medline: 23367505]
89. Ho R. Handbook of Univariate and Multivariate Data Analysis and Interpretation with SPSS. Boca Raton, FL, USA: CRC Press Taylor & Francis Group; 2006.
90. Byrne BM. Structural equation modeling: perspectives on the present and the future. Int J Test 2001;1(3-4):327-334. [doi: 10.1080/15305058.2001.9669479]
91. Kline RB. Principles and Practice of Structural Equation Modeling. New York, NY, USA: The Guilford Press; 2015.
92. Unertl K, Johnson K, Lorenzi NM. Health information exchange technology on the front lines of healthcare: workflow factors and patterns of use. J Am Med Inform Assoc 2011;19(3):392-400 [FREE Full text] [doi: 10.1136/amiajnl-2011-000432] [Medline: 22003156]

XSL•FO
RenderX

93. Kahn D, Pace-Schott E, Hobson J. Emotion and cognition: feeling and character identification in dreaming. Conscious Cogn 2002 Mar;11(1):34-50. [doi: 10.1006/ccog.2001.0537] [Medline: 11883987]
94. Heath M, Appan R, Gudigantala N. Exploring health information exchange (HIE) through collaboration framework: normative guidelines for it leadership of healthcare organizations. Inform Syst Manag 2017;34(2):137-156. [doi: 10.1080/10580530.2017.1288524]
95. Delgado-Ballester E, Hernández-Espallardo M. Effect of brand associations on consumer reactions to unknown on-line brands. Int J Elect Comm 2008;12(3):81-113. [doi: 10.2753/jec1086-4415120305]
96. Esmaeilzadeh P, Mirzaei T. Comparison of consumers' perspectives on different health information exchange (HIE) mechanisms: an experimental study. Int J Med Inform 2018 Nov;119:1-7. [doi: 10.1016/j.ijmedinf.2018.08.007] [Medline: 30342677]
97. Paasche-Orlow MK, Jacob DM, Powell JN. Notices of Privacy Practices: a survey of the Health Insurance Portability and Accountability Act of 1996 documents presented to patients at US hospitals. Med Care 2005;43(6):558-564. [doi: 10.1097/01.mlr.0000163646.90393.e4] [Medline: 15908850]
98. Singh RI, Sumeeth M, Miller J. A user-centric evaluation of the readability of privacy policies in popular web sites. Inf Syst Front 2011;13(4):501-514. [doi: 10.1007/s10796-010-9228-2]

Abbreviations

ACP: average congruency percentage
AVE: average variance extracted
e-commerce: electronic commerce
GFI: goodness of fit index
HIE: health information exchange
IS: information system
IT: information technology
MTurk: Mechanical Turk
TRA: Theory of Reasoned Action
VIF: variance inflation factor

Edited by G Eysenbach; submitted 18.03.19; peer-reviewed by J Rowley, L Dimitropoulos; comments to author 04.09.19; revised version received 11.09.19; accepted 28.09.19; published 26.11.19.

Please cite as:
Esmaeilzadeh P
The Impacts of the Perceived Transparency of Privacy Policies and Trust in Providers for Building Trust in Health Information Exchange: Empirical Study
JMIR Med Inform 2019;7(4):e14050
URL: http://medinform.jmir.org/2019/4/e14050/
doi:10.2196/14050
PMID:31769757

©Pouyan Esmaeilzadeh. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 26.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Developing a Reproducible Microbiome Data Analysis Pipeline Using the Amazon Web Services Cloud for a Cancer Research Group: Proof-of-Concept Study

Jinbing Bai[1,2], MSN, PhD; Ileen Jhaney[3], MSPH; Jessica Wells[1,2], PhD

[1]Nell Hodgson Woodruff School of Nursing, Emory University, Atlanta, GA, United States

[2]Cancer Prevention and Control Program, Winship Cancer Institute, Emory University, Atlanta, GA, United States

[3]Winship Research Informatics Shared Resource, Winship Cancer Institute, Emory University, Atlanta, GA, United States

Corresponding Author:
Jinbing Bai, MSN, PhD
Nell Hodgson Woodruff School of Nursing
Emory University
1520 Clifton Road NE
Atlanta, GA, 30322
United States
Phone: 1 404 7272466
Email: jbai222@emory.edu

Abstract

Background: Cloud computing for microbiome data sets can significantly increase working efficiencies and expedite the translation of research findings into clinical practice. The Amazon Web Services (AWS) cloud provides an invaluable option for microbiome data storage, computation, and analysis.

Objective: The goals of this study were to develop a microbiome data analysis pipeline by using AWS cloud and to conduct a proof-of-concept test for microbiome data storage, processing, and analysis.

Methods: A multidisciplinary team was formed to develop and test a reproducible microbiome data analysis pipeline with multiple AWS cloud services that could be used for storage, computation, and data analysis. The microbiome data analysis pipeline developed in AWS was tested by using two data sets: 19 vaginal microbiome samples and 50 gut microbiome samples.

Results: Using AWS features, we developed a microbiome data analysis pipeline that included Amazon Simple Storage Service for microbiome sequence storage, Linux Elastic Compute Cloud (EC2) instances (ie, servers) for data computation and analysis, and security keys to create and manage the use of encryption for the pipeline. Bioinformatics and statistical tools (ie, Quantitative Insights Into Microbial Ecology 2 and RStudio) were installed within the Linux EC2 instances to run microbiome statistical analysis. The microbiome data analysis pipeline was performed through command-line interfaces within the Linux operating system or in the Mac operating system. Using this new pipeline, we were able to successfully process and analyze 50 gut microbiome samples within 4 hours at a very low cost (a c4.4xlarge EC2 instance costs $0.80 per hour). Gut microbiome findings regarding diversity, taxonomy, and abundance analyses were easily shared within our research team.

Conclusions: Building a microbiome data analysis pipeline with AWS cloud is feasible. This pipeline is highly reliable, computationally powerful, and cost effective. Our AWS-based microbiome analysis pipeline provides an efficient tool to conduct microbiome data analysis.

(JMIR Med Inform 2019;7(4):e14667) doi:10.2196/14667

KEYWORDS

Amazon Web Services; cloud computation; microbiome; pipeline; sequence analysis

Introduction

Big data and data-driven analysis has become a primary driver of precision health [1,2]. The human microbiota and their genomes, collectively called the human microbiome, is one form of big data [3]. The human body harbors trillions of microbes, including bacteria, viruses, fungi, and archaea [4,5], which vary from host to host and across body sites within a single host [6,7]. The human microbiome plays a critical role

XSL•FO
RenderX

in human health and disease [8,9]. With advances in next-generation sequencing technology and the rise of shotgun metagenomics and metabolomic techniques, microbiome data sets have rapidly expanded, especially following the initiatives of the Human Microbiome Project [10] and the American Gut Project [11]. Computation and analysis of big data sets in local infrastructures via traditional computational methods (eg, use of personal computers and local computational clusters) often requires prolonged run times, delaying further analytic work that needs to be performed and postponing the translation of research findings into clinical practice [12]. Another shortcoming of classical data analysis methods is the difficulty involved in sharing the data and findings among research collaborators. Advances in cloud computing have provided the technical capabilities to help resolve difficulties posed by standard computational methods [12,13].

Beyond data storage, assessing human microbiome data sets requires bioinformatic tools that enable deeper mining, the deciphering of the mechanistic connections among the microbes, and the potential functions of these communities. To examine significant associations between metadata (eg, demographic and clinical variables) and DNA sequencing data, special bioinformatic and statistical tools for conducting microbiome analyses are needed [14]. It was not until recently that researchers have developed software for microbiome data analysis (eg, Quantitative Insights Into Microbial Ecology [QIIME] and Mothur) [15]. One popular bioinformatics tool, QIIME 2, can be installed natively within a conda environment through a docker or a cloud platform. The Amazon Web Services (AWS) cloud provides an invaluable computational environment for running bioinformatics tools, such as QIIME 2, without the overhead of implementing and supporting a large-scale computing infrastructure [12]. Cloud-based computational pipelines have been developed for a variety of data analysis, including CloudNeo [16] and RNA-Sequencing (RNA-Seq) Analysis Pipeline [17], for next-generation sequencing data, and Clustered Regularly Interspaced Short Palindromic Repeats Cloud (CRISPRcloud) for the deconvolution of pooled screening data [18]. The development of a comprehensive microbiome data analysis pipeline, including data storage, computations, and analysis, along with its testing using microbiome data sets from actual studies, would help researchers further investigate the impact of the microbiome on human health and disease (eg, cancer, metabolic syndromes, and neurodegenerative disorders) [8,19-21]. A reliable and validated microbiome data analysis pipeline operating through the AWS cloud could be used to provide a consistent communication platform for research collaborators to share information on data processing, data analysis, and research findings. Thus, the AWS pipeline could increase both the reproducibility of microbiome studies and the proficiency of the research team [22].

For clinical scientists, several challenges need to be overcome before conducting microbiome projects: (1) collective storage space for big microbiome data sets needs to be created so that the data and results can be easily shared within the research team; (2) centralized data computing capabilities need to be established to foster the replicability of results across all current and future projects; and (3) the cost for cloud computing services needs to be determined so that the team can cost-effectively study the human microbiome. The AWS cloud has become a popular platform for big data storage, high performance computing, and analytics [16-18,23]. Thus, the purpose of this study was to develop a microbiome data analysis pipeline using the AWS cloud service (MAP-AWS) and conduct a proof-of-concept test for microbiome data computation and analysis with this newly developed MAP-AWS.

Methods

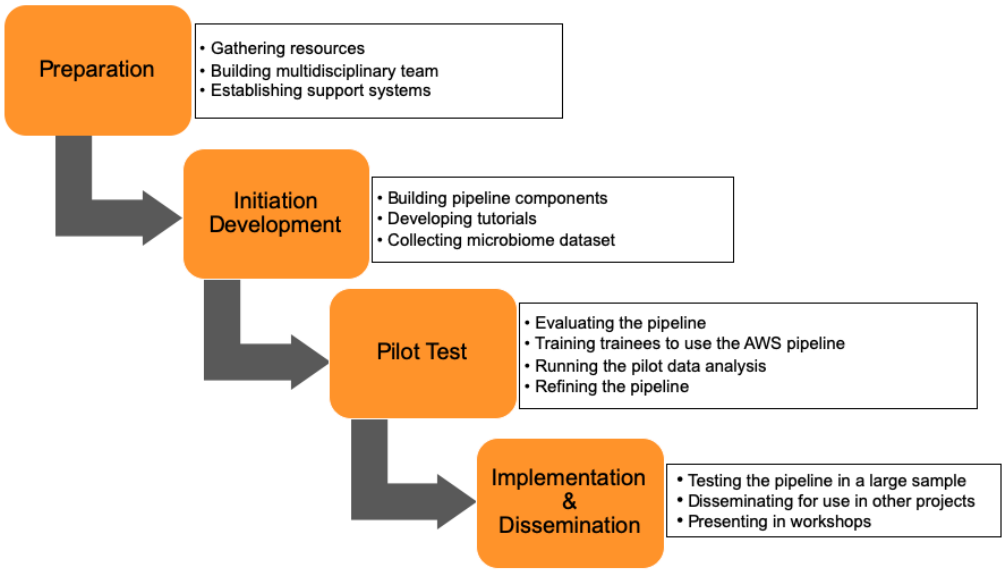
Overview

The process of developing and testing the MAP-AWS comprised of four stages, as illustrated in Figure 1. We first collected resources regarding microbiome data analysis and the use of MAP-AWS, built a multidisciplinary research team, and lined up available support systems from our institution and the AWS Support Center. Second, we initiated the development of the microbiome analysis pipeline, including a tutorial, with support of the Research Informatics Team from the Winship Cancer Institute at Emory University (Atlanta, Georgia, USA).

Third, we began pilot testing: we ran a small vaginal microbiome data set (n=19), refined the pipeline and the tutorial, and retested the pipeline with a larger data set from the gut microbiome (n=50). Last, we disseminated the MAP-AWS within our institutional research groups via presentations and workshops and obtained feedback regarding this newly developed MAP-AWS.

XSL•FO
RenderX

Figure 1. Design process of the microbiome analysis pipeline.



Preparation

Through our previous work analyzing microbiome data sets [22,24], we found that assembling a team with the necessary skill sets, ensuring financial feasibility, and establishing system resources were essential components of developing a system for big data analysis. Building a multidisciplinary team with specific expertise is key to successfully deploying AWS cloud for use in microbiome data processing, computing, and analysis. The MAP-AWS team was built within a multidisciplinary group of nurses, physicians, biostatisticians, epidemiologists, and microbiologists from the Schools of Nursing, Public Health, and Medicine, and the Winship Cancer Institute at Emory University. One nurse researcher, primarily trained in the human microbiome and cancer science, led the team and formed extensive collaborations with several other team members: one research informatics analyst from the Winship Cancer Institute, one biostatistician from the School of Public Health, and one predoctoral student from the School of Nursing. All team members had considerable experience in microbiome data analysis [24,25], including previous QIIME 2 training, active participation in microbiome-related internal and external workshops, and publications on the human microbiome and its impact on human health [24,25].

Recruiting information technology (IT) resources early in the preparation phase is critical to ensuring that all IT resources are coordinated by the time they are needed, and support services for troubleshooting activities can be provided in a timely manner. IT staff can help coordinate and organize the diverse data sets that will be used, including the DNA sequencing data, metadata, project-related information, and research protocols. Assignment of policies and permissions for access to all AWS resources is a critical role of the Research Informatics IT group. Assistance with AWS command-line interface (CLI) was provided through the IT support group as well. We set up AWS accounts for each team member intensively involved in developing the MAP-AWS with the support of our IT support group and the online AWS Support Center.

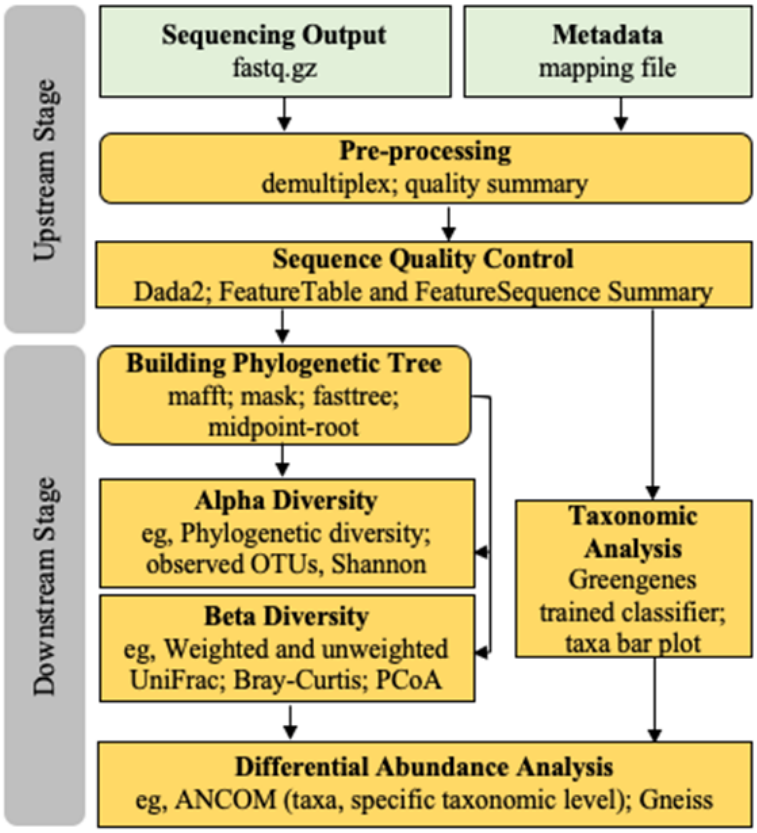
Development of the Microbiome Data Analysis Pipeline Using Amazon Web Services

AWS cloud provides various options for microbiome data storage, computing, and analysis. For data storage, the Amazon Simple Storage Service (S3) buckets were used. For microbiome data computing, Amazon Elastic Compute Cloud (EC2) instances (ie, virtual servers in the AWS cloud) were primarily used, and for microbiome data analysis and specific bioinformatics, statistical packages that included QIIME 2 [22] and RStudio (RStudio, Boston, Massachusetts), were installed within the EC2 instances. For most of the EC2 instances we created we opted to use the Linux operating system, which has an optimized central processing unit (CPU), memory, and storage configurations [26]. During this specific developmental stage, we produced a step-by-step tutorial on how to run processes on microbiome data sets within AWS. This tutorial included the topics: logging into AWS, data importation and storage, data analysis using QIIME 2, and exportation of analysis results. This newly developed MAP-AWS provides a complete workflow to run microbiome data analysis in AWS.

The use of QIIME 2 for microbiome data analysis has been tested by our group in a variety of computer systems, such as the Linux operating system, the Mac operating system (OS), and AWS cloud [24]. The QIIME 2 pipeline generates the bacterial community's information for each sample [22], and this process includes two phases which are referred to as the upstream and downstream stages (Figure 2). The upstream stage consists of importing 16S rRNA sequences, ensuring sequence quality control, constructing a feature table, and generating a phylogenetic tree which illustrates the ecologic similarities of the bacterial taxa present in a sample [24]. The feature table describes the features present and the number of samples associated with each feature in the sample set. The downstream stage consists of taxonomic, diversity, and abundance analysis [14,22]. In this stage, statistics and interactive visualizations of the data are used to display the findings via figures and tables [27].

XSL•FO
RenderX

Figure 2. QIIME 2 workflow. QIIME: Quantitative Insights Into Microbial Ecology; OTU: operational taxonomic unit; PCoA: principal coordinates analysis; ANCOM: analysis of composition of microbiomes.



Testing the Microbiome Data Analysis Pipeline Using Amazon Web Services

To test the feasibility of the MAP-AWS, we undertook three rounds of testing using a case-based formative approach to refine the microbiome analysis pipeline. In round one, we trained two novice students to use the MAP-AWS to determine where major changes were needed to improve the usability of the content and presentation formats of the tutorial and pipeline. Then, we demonstrated the MAP-AWS to a group of cancer scientists to get feedback regarding the content and presentation of the tutorial. In round two, we conducted a pilot test of the workflow in AWS cloud with a small training data set (19 deidentified vaginal microbiome samples from women with gynecologic cancers), which we had prepared for the purpose of training research scientists in microbiome data analysis [24]. This step enabled us to identify and troubleshoot issues before running a larger microbiome data set. In round three, two team members (JB and IJ) independently analyzed the same vaginal microbiome data set (ie, 16S rRNA V3-V4 gene sequences with corresponding metadata) using the MAP-AWS with the same Greengenes classifiers to determine the reproducibility of the pipeline. Final findings were compared between the two team members. Lastly, we ran a larger sample (50 deidentified gut microbiome samples) sequenced by the Emory Integrated Genomics Core. For each project, we regularly tracked costs and the processing times using the built-in QIIME 2 provenance feature that captures system environment variables, including processing time and system versions (ie, Linux and QIIME).

Dissemination of the Microbiome Data Analysis Pipeline Using Amazon Web Services

After testing and refining the MAP-AWS processes and tutorial, we expanded the use of this pipeline to other microbiome projects within our team, including a gut microbiome and colorectal carcinogenesis study which involved sequence and metadata import, data quality control, results analysis, and model building. We disseminated our pipeline through presentations and workshops.

Ethical Consideration

All the microbiome data we used in this study have been deidentified and no Institutional Review Board (IRB) approval is needed.

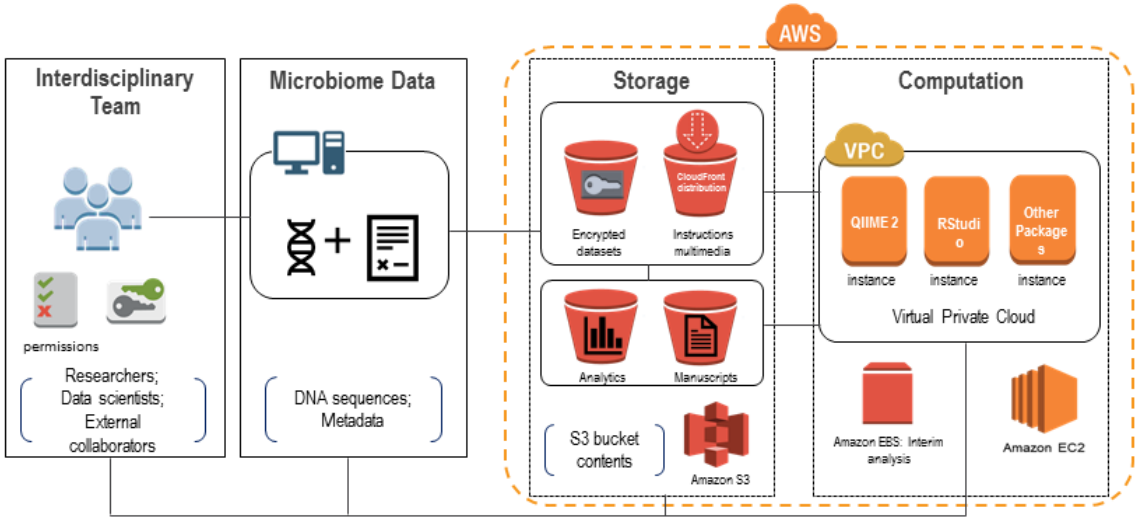
Results

Description of the Developed Microbiome Data Analysis Pipeline Using Amazon Web Services

The main components of the MAP-AWS include a multidisciplinary research team, bacterial sequences and corresponding metadata, Amazon S3 buckets for microbiome data storage, Linux EC2 instances (with QIIME 2 and RStudio installed) to run microbiome data analysis, and security keys to create and manage the use of encryption (Figure 3). With our platform, microbiome data analysis can be performed using AWS's CLI within the Linux operating system or in the Mac OS system.

XSL•FO
RenderX

Figure 3. The microbiome data analysis pipeline using AWS. AWS: Amazon Web Services; S3: Simple Storage Service; VPC: virtual private cloud; QIIME: Quantitative Insights Into Microbial Ecology; EBS: Elastic Block Store; EC2: Elastic Compute Cloud.



We primarily used S3 buckets for storage and EC2 instances for analysis. With our MAP-AWS, we used two specific types of storage. The first was Amazon Elastic Block Store (EBS), which is closely integrated with our EC2 instance. EBS is used to store hard drive contents of EC2 instances, as well as snapshots of these instances [28]. The other storage class used was Amazon S3 bucket, which is simply cloud object storage. We carefully planned permissions for storage, encryption, and EC2 instance access. The AWS EC2 provides virtual machines that are optimized for running CPU-intensive cloud-based applications [28]. Depending upon the analysis and virtual server purpose, our EC2 instances were configured for general purpose or optimized specifically for memory, computational power, or storage. For each EC2 instance, we were able to specify random access memory (RAM), virtual CPUs, storage, and network performance.

A Secure Shell (SSH) client (either a Mac OS terminal or MobaXterm for Windows) was used to securely connect to our EC2 instances, enabling remote access to a terminal through which Linux commands could be entered to process data. The AWS CLI was installed and used extensively to interact with our AWS resources and infrastructure.

Next, we set up encryption to ensure our S3 buckets were secured from unauthorized access. We assigned encryption at the bucket level so that all objects moved into the bucket were automatically encrypted.

Testing of the Microbiome Data Analysis Pipeline Using Amazon Web Services

Feasibility

Two undergraduate nursing students were trained to use the MAP-AWS tutorial we developed and were interviewed after they finished performing analysis of the training microbiome data set (the vaginal microbiome samples). They both successfully completed the training data set analysis under the guidance of the tutorial and our team member (IJ). Both undergraduates were positive about the use of the MAP-AWS and the tutorial, supporting the feasibility of the MAP-AWS.

Reproducibility

All steps performed using the MAP-AWS in this study were tested with support from the QIIME 2 Development Team and the AWS Support Center. A total of three incidents involving the S3 bucket and EC2 instances needed addressing and were resolved by the AWS Support Center during the MAP-AWS development process. Two trained microbiome team members (JB and IJ) independently analyzed the same vaginal microbiome data set (ie, 16S rRNA V3-V4 gene sequences with the corresponding metadata, n=19) using the MAP-AWS. Comparisons of the final findings showed identical results in upstream and downstream analyses (Figure 2) [24], supporting the reproducibility of the MAP-AWS for microbiome data analysis.

Cost and Efficiency

The 50 gut microbiome samples were successfully processed within 4 hours with the MAP-AWS and subsequently processed for microbiome diversity, taxonomy, and abundance analyses using QIIME 2, version 2018.4 (Table 1). Performing microbiome data analysis for the same data set with typical client-server architecture took >6 hours. These running times were retrieved from the QIIME 2 provenance. We duplicated our efforts on a more recent QIIME 2, version 2019.4, and the running times for completion were congruent with previous results. Compared with standard methods of microbiome data analysis, the MAP-AWS processed these samples efficiently and at a low cost. We used a c4.4xlarge EC2 instance, which costs $0.80 per hour. This pricing level is similar to the Nephele pipeline published in 2017, a c3.4xlarge EC2 instance at the time costing $0.84 per hour [23].

XSL•FO
RenderX

Table 1. Running Time for the Gut Microbiome Sample Analysis (n=50).

Stage, step	MAP-AWS[a]	Traditional methods
Upstream stage		
Data import	4 s	1 min 48 s
Quality control (ie, Dada2)	3 h 22 min 27 s	6 h 38 min 49 s
Phylogeny	32 s	<1 min
Downstream stage		
Taxonomy analysis	1 min 23 s	3 min 27 s
Diversity analysis	2 s	8 s

[a]MAP-AWS: Microbiome Data Analysis Pipeline Using Amazon Web Services.

Discussion

Key Findings

This paper developed a microbiome data analysis pipeline by using AWS cloud and conducted a proof-of-concept test for microbiome data storage, processing, and analysis. This pipeline is highly reliable, computationally powerful, and cost effective. This study was a proof of concept for building and testing a newly developed pipeline (MAP-AWS) for microbiome data analysis. This pipeline is efficient and highly cost effective. It will provide a convenient environment to share analysis tools and results between collaborators. To accurately assess and utilize this data, we rely on the development of tools, pipelines, and standard operating procedures to handle big data effectively and efficiently via the AWS cloud. Microbiome pipelines using on-demand EC2 instances showed a great capacity for microbiome data analysis at a low cost. This pipeline improved productive and insightful collaboration with clinical scientists across different institutions to help the multidisciplinary research team continue the collaborative use of AWS.

With growing interest in evaluating the human microbiome and deciphering its relationship with health and disease, more efficient and cost-effective tools are needed for microbiome big data analysis. The purpose of this study was to develop and evaluate the MAP-AWS platform for use by clinical scientists. We described how researchers can construct their own microbiome data analysis pipeline using AWS. The AWS cloud can significantly expedite the microbiome analysis process and provide a collaborative platform for sharing data and results among research collaborators. The MAP-AWS tool successfully completed all microbiome processing and analysis steps both efficiently and reproducibly. The MAP-AWS not only maintains essential reproducibility of processing steps and analyses but also facilitates the efficiency and cost-effectiveness of microbiome data analysis in contrast with basic, commonly used methods of microbiome data analysis [12].

Compared with standard processing for big data analysis, the AWS cloud brings extensive benefits to current microbiome data analysis, including optimized computational capabilities, flexible EC2 instance configurations, and robust security and policies for all resources. Although common server and desktop environments can provide microbiome processing capabilities, AWS brings a supportive systems environment for storage, computational, and analytical capabilities. For instance, many methods in the microbiome platform benefit from compute-optimized processing since their focus is serving high performance computing targeted for compute-intensive applications. The MAP-AWS includes an integrated tool with a combined tutorial for using AWS tools (such as S3 bucket retrieval and EC2 instances use) and performing raw data processing, advanced QIIME 2 and RStudio analysis, and data sharing and management between researchers. This MAP-AWS platform establishes a common environment for sharing analysis tools and results between project managers and researchers across institutions. Given the appropriate permissions, researchers internal to the University and external collaborators can reliably rerun analyses and share findings. It is easy to deploy the microbiome data platform in multiple regions around the world with just a few clicks.

AWS cloud has been widely adopted for whole-genome sequencing (WGS) analysis tasks. For large-scale WGS analyses, AWS was shown to be an efficient and affordable WGS analysis tool [29]. Specifically, Wang and colleagues evaluated the performance of GT-WGS with a 55×WGS data set (400 gigabyte fastq sequences), provided by the genome-wide complex trait analysis (GCTA) 2017 competition, and found that their system took only 18.4 minutes to finish the analysis and that the cost of the whole process was only $16.50 (United States Dollars) [29]. Likewise, our initial microbiome pilot study was completed quickly (within 4 hours) using the MAP-AWS, in contrast with 2-3 days for runs using local computers. Thus, implementing MAP-AWS can significantly improve computing efficiency and speed up the translation of research findings into clinical practice.

Several EC2 pricing models exist, including on-demand, reserved, and spot instances. Users can increase or decrease their computing capacity according to the real-time demand of their applications with on-demand instances and by paying by a specified hourly rate. We tested our microbiome pipeline using on-demand instances, showing a great capacity for microbiome data analysis at a low cost.

One of the biggest challenges facing researchers is the ability to integrate and correlate the massive amounts of data produced by these protocols and identify biologically relevant information that can be used to formulate testable hypotheses. As a proof-of-concept test for the utilization of AWS in microbiome

XSL•FO
RenderX

data analysis, our findings support its value and affordability. Our MAP-AWS efficiently integrated and correlated significant amounts of omics data stored and utilized in a cloud-based environment and provided a streamlined platform for communication between researchers. Microbiome research is on the precipice of producing large data sets of great magnitude. To accurately assess and utilize this data, investigators must rely on the development of tools, pipelines, and standard operating procedures to handle big data effectively and efficiently via the AWS cloud. Together, researchers, clinicians, and computer scientists, with the help of AWS cloud computing services, are poised to revolutionize microbiome research and its applications in human health.

Our microbiome data analysis pipeline was undertaken within a cancer nursing research group and tested with data sets of small sample sizes. The technical pipeline should also be applicable to other microbiome data sets, such as oral or skin microbiome data. A dysbiotic human microbiome is associated with a variety of human disease susceptibility [21,30], including endocrine-related disorders (eg, diabetes [31] and inflammatory bowel diseases [32]) and neurodevelopmental disorders (eg, autism spectrum disorders [33] and Alzheimer's disease [34]). Therefore, the MAP-AWS can be extended to analyze the microbiome data of various chronic diseases and conditions. Our goal is to further test our MAP-AWS using large data sets. In addition, the current pipeline is primarily embedded with QIIME 2 and RStudio, which limits the use of other microbiome analysis packages like Mothur [35]. As QIIME 2 is gaining more attention as a bioinformatics tool, the MAP-AWS is an ideal example of conducting microbiome data analysis with the AWS cloud. As there is increased access to deidentified microbiome data sets, such as the Human Microbiome Project [10], American Gut [11], and the Qiita platform [36], the MAP-AWS will provide our clinical scientists and clinicians a new cloud-based tool to understand the role of the microbiome in quality of care and patient outcomes.

Conclusions

This study was a proof of concept for building and testing a newly developed pipeline (MAP-AWS) for microbiome data analysis. This pipeline is efficient and highly cost effective. It will provide a convenient environment to share analysis tools and results between collaborators. The long-term goal for this platform is to continue the collaborative use of AWS among clinical scientists across different institutions to make our multidisciplinary research team more productive and insightful. A larger-scale testing of the MAP-AWS across different clinical conditions will enhance communications between multidisciplinary researchers and confirm our proposed efficiencies for running a microbiome pipeline in a cloud-based environment.

Acknowledgments

This research project was supported by the Amazon Web Services Cloud Credits for Research program.

Conflicts of Interest

None declared.

References

1. Prosperi M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. BMC Med Inform Decis Mak 2018 Dec 29;18(1):139 [FREE Full text] [doi: 10.1186/s12911-018-0719-2] [Medline: 30594159]
2. Agarwal V, Zhang L, Zhu J, Fang S, Cheng T, Hong C, et al. Impact of Predicting Health Care Utilization Via Web Search Behavior: A Data-Driven Analysis. J Med Internet Res 2016 Sep 21;18(9):e251 [FREE Full text] [doi: 10.2196/jmir.6240] [Medline: 27655225]
3. Ursell LK, Metcalf JL, Parfrey LW, Knight R. Defining the human microbiome. Nutr Rev 2012 Aug;70 Suppl 1:S38-S44 [FREE Full text] [doi: 10.1111/j.1753-4887.2012.00493.x] [Medline: 22861806]
4. Knight R. Follow Your Gut: The Enormous Impact Of Tiny Microbes (TED Books): Simon & Schuster / Ted; 2019.
5. Sender R, Fuchs S, Milo R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. Cell 2016 Jan 28;164(3):337-340 [FREE Full text] [doi: 10.1016/j.cell.2016.01.013] [Medline: 26824647]
6. Spor A, Koren O, Ley R. Unravelling the effects of the environment and host genotype on the gut microbiome. Nat Rev Microbiol 2011 Apr 16;9(4):279-290. [doi: 10.1038/nrmicro2540] [Medline: 21407244]
7. Morgan XC, Huttenhower C. Chapter 12: Human microbiome analysis. PLoS Comput Biol 2012 Dec 27;8(12):e1002808 [FREE Full text] [doi: 10.1371/journal.pcbi.1002808] [Medline: 23300406]
8. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet 2012 Mar 13;13(4):260-270 [FREE Full text] [doi: 10.1038/nrg3182] [Medline: 22411464]
9. Lynch SV, Pedersen O. The Human Intestinal Microbiome in Health and Disease. N Engl J Med 2016 Dec 15;375(24):2369-2379. [doi: 10.1056/nejmra1600266]
10. Gevers D, Knight R, Petrosino JF, Huang K, McGuire AL, Birren BW, et al. The Human Microbiome Project: a community resource for the healthy human microbiome. PLoS Biol 2012 Aug 14;10(8):e1001377 [FREE Full text] [doi: 10.1371/journal.pbio.1001377] [Medline: 22904687]
11. American Gut Project. 2019. The American Gut Project Dataset URL: http://americangut.org/resources/ [accessed 2018-01-20]

XSL•FO
RenderX

12. Navas-Molina JA, Hyde ER, Sanders JG, Knight R. The microbiome and big data. Current Opinion in Systems Biology 2017 Aug;4:92-96. [doi: 10.1016/j.coisb.2017.07.003]
13. Nature Publishing Group. Microbiota meet big data. Nat Chem Biol 2014 Aug;10(8):605. [doi: 10.1038/nchembio.1604] [Medline: 25036299]
14. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010 May 11;7(5):335-336 [FREE Full text] [doi: 10.1038/nmeth.f.303] [Medline: 20383131]
15. Navas-Molina J, Peralta-Sanchez J, Gonzalez A, McMurdie P, Vazquez-Baeza Y, Xu Z, et al. Advancing our understanding of the human microbiome using QIIME. Methods Enzymol 2013;531:371-444. [doi: 10.1016/b978-0-12-407863-5.00019-8]
16. Bais P, Namburi S, Gatti D, Zhang XJ, Chuang JH. CloudNeo: a cloud pipeline for identifying patient-specific tumor neoantigens. Bioinformatics 2017 Oct 01;33(19):3110-3112 [FREE Full text] [doi: 10.1093/bioinformatics/btx375] [Medline: 28605406]
17. D'Antonio M, D'Onorio De Meo P, Pallocca M, Picardi E, D'Erchia AM, Calogero RA, et al. RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application. BMC Genomics 2015;16(Suppl 6):S3. [doi: 10.1186/1471-2164-16-s6-s3]
18. Jeong H, Kim S, Rousseaux M, Zoghbi HZ, Liu Z. CRISPRcloud: a secure cloud-based pipeline for CRISPR pooled screen deconvolution. Bioinformatics 2017 Sep 15;33(18):2963-2965 [FREE Full text] [doi: 10.1093/bioinformatics/btx335] [Medline: 28541456]
19. Claesson MJ, Clooney AG, O'Toole PW. A clinician's guide to microbiome analysis. Nat Rev Gastroenterol Hepatol 2017 Oct 9;14(10):585-595. [doi: 10.1038/nrgastro.2017.97] [Medline: 28790452]
20. Rajagopala SV, Vashee S, Oldfield LM, Suzuki Y, Venter JC, Telenti A, et al. The Human Microbiome and Cancer. Cancer Prev Res 2017 Jan 17;10(4):226-234. [doi: 10.1158/1940-6207.capr-16-0249]
21. Schroeder BO, Bäckhed F. Signals from the gut microbiota to distant organs in physiology and disease. Nat Med 2016 Oct 6;22(10):1079-1089. [doi: 10.1038/nm.4185] [Medline: 27711063]
22. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 2019 Aug 24;37(8):852-857. [doi: 10.1038/s41587-019-0209-9] [Medline: 31341288]
23. Weber N, Liou D, Dommer J, MacMenamin P, Quiñones M, Misner I, et al. Nephele: a cloud platform for simplified, standardized and reproducible microbiome data analysis. Bioinformatics 2018 Apr 15;34(8):1411-1413 [FREE Full text] [doi: 10.1093/bioinformatics/btx617] [Medline: 29028892]
24. Bai J, Jhaney I, Daniel G, Watkins Bruner D. Pilot Study of Vaginal Microbiome Using QIIME 2™ in Women With Gynecologic Cancer Before and After Radiation Therapy. Oncol Nurs Forum 2019 Mar 01;46(2):E48-E59. [doi: 10.1188/19.ONF.E48-E59] [Medline: 30767956]
25. Bai J, Hu Y, Bruner DW. Composition of gut microbiota and its association with body mass index and lifestyle factors in a cohort of 7-18 years old children from the American Gut Project. Pediatr Obes 2019 Apr 11;14(4):e12480. [doi: 10.1111/ijpo.12480] [Medline: 30417607]
26. Amazon Web Services. AWS. 2019. Getting started with Amazon EC2 Linux instances URL: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EC2_GetStarted.html [accessed 2019-08-15]
27. Navas-Molina J, Peralta-Sanchez J, Gonzalez A, McMurdie P, Vazquez-Baeza Y, Xu Z, et al. Advancing Our Understanding of the Human Microbiome Using QIIME. Microbial Metagenomics, Metatranscriptomics, and Metaproteomics 2013;531:371-444. [doi: 10.1016/b978-0-12-407863-5.00019-8]
28. Madhyastha TM, Koh N, Day TKM, Hernández-Fernández M, Kelley A, Peterson DJ, et al. Running Neuroimaging Applications on Amazon Web Services: How, When, and at What Cost? Front Neuroinform 2017 Nov 03;11:63 [FREE Full text] [doi: 10.3389/fninf.2017.00063] [Medline: 29163119]
29. Wang Y, Li G, Ma M, He F, Song Z, Zhang W, et al. GT-WGS: an efficient and economic tool for large-scale WGS analyses based on the AWS cloud service. BMC Genomics 2018 Jan 19;19(Suppl 1):959 [FREE Full text] [doi: 10.1186/s12864-017-4334-x] [Medline: 29363427]
30. Slattery J, Macfabe DF, Frye RE. The Significance of the Enteric Microbiome on the Development of Childhood Disease: A Review of Prebiotic and Probiotic Therapies in Disorders of Childhood. Clin Med Insights Pediatr 2016 Oct 09;10:CMPed.S38338. [doi: 10.4137/cmped.s38338]
31. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 2012 Oct 04;490(7418):55-60. [doi: 10.1038/nature11450] [Medline: 23023125]
32. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol 2012 Apr 16;13(9):R79 [FREE Full text] [doi: 10.1186/gb-2012-13-9-r79] [Medline: 23013615]
33. Hsiao E, McBride S, Hsien S, Sharon G, Hyde E, McCue T, et al. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell 2013 Dec 19;155(7):1451-1463 [FREE Full text] [doi: 10.1016/j.cell.2013.11.024] [Medline: 24315484]
34. Hill JM, Clement C, Pogue AI, Bhattacharjee S, Zhao Y, Lukiw WJ. Pathogenic microbes, the microbiome, and Alzheimer's disease (AD). Front Aging Neurosci 2014;6:127 [FREE Full text] [doi: 10.3389/fnagi.2014.00127] [Medline: 24982633]

XSL•FO
RenderX

35. Xia Y, Sun J. Hypothesis Testing and Statistical Analysis of Microbiome. Genes Dis 2017 Sep;4(3):138-148 [FREE Full text] [doi: 10.1016/j.gendis.2017.06.001] [Medline: 30197908]
36. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, et al. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods 2018 Oct 1;15(10):796-798 [FREE Full text] [doi: 10.1038/s41592-018-0141-9] [Medline: 30275573]

Abbreviations

AWS: Amazon Web Services
CLI: command-line interface
CPU: central processing unit
CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats
EBS: Elastic Block Store
EC2: Elastic Compute Cloud
GCTA: genome-wide complex trait analysis
IRB: Institutional Review Board
IT: information technology
MAP-AWS: microbiome data analysis pipeline using Amazon web services
OS: operating system
QIIME: Quantitative Insights Into Microbial Ecology
RAM: random access memory
RNA-Seq: RNA-Sequencing
S3: Simple Storage Service
SSH: Secure Shell
WGS: whole-genome sequencing

Edited by C Lovis; submitted 09.05.19; peer-reviewed by A Benis, G Kolostoumpis, L Zhang; comments to author 14.07.19; revised version received 20.08.19; accepted 22.08.19; published 11.11.19.

Please cite as:
Bai J, Jhaney I, Wells J
Developing a Reproducible Microbiome Data Analysis Pipeline Using the Amazon Web Services Cloud for a Cancer Research Group: Proof-of-Concept Study
JMIR Med Inform 2019;7(4):e14667
URL: http://medinform.jmir.org/2019/4/e14667/
doi:10.2196/14667
PMID:31710301

©Jinbing Bai, Ileen Jhaney, Jessica Wells. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 11.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Deep Learning Approach for Managing Medical Consumable Materials in Intensive Care Units via Convolutional Neural Networks: Technical Proof-of-Concept Study

Arne Peine[1,2], MD, MHBA; Ahmed Hallawa[1,3], MSc; Oliver Schöffski[4], Prof Dr, MPH; Guido Dartmann[2,5], Prof Dr; Lejla Begic Fazlic[5], PhD; Anke Schmeink[2,6], Prof Dr; Gernot Marx[1,2], Prof Dr, MD, FRCA; Lukas Martin[1,2], MD, MHBA

[1]Department of Intensive Care Medicine and Intermediate Care, University Hospital Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany

[2]Clinomic GmbH, Aachen, Germany

[3]Chair for Integrated Signal Processing Systems, Rheinisch-Westfälische Technische Hochschule Aachen University, Aachen, Germany

[4]Chair of Health Management, School of Business, Economics and Society, Friedrich-Alexander-University Erlangen-Nürnberg, Nürnberg, Germany

[5]Research Area Distributed Systems, Trier University of Applied Sciences, Trier, Germany

[6]Research Area Information Theory and Systematic Design of Communication Systems, Rheinisch-Westfälische Technische Hochschule Aachen University, Aachen, Germany

Corresponding Author:
Arne Peine, MD, MHBA
Department of Intensive Care Medicine and Intermediate Care
University Hospital Rheinisch-Westfälische Technische Hochschule Aachen
Pauwelsstr 30
Aachen, 52074
Germany
Phone: 49 241 800
Email: apeine@ukaachen.de

Abstract

Background: High numbers of consumable medical materials (eg, sterile needles and swabs) are used during the daily routine of intensive care units (ICUs) worldwide. Although medical consumables largely contribute to total ICU hospital expenditure, many hospitals do not track the individual use of materials. Current tracking solutions meeting the specific requirements of the medical environment, like barcodes or radio frequency identification, require specialized material preparation and high infrastructure investment. This impedes the accurate prediction of consumption, leads to high storage maintenance costs caused by large inventories, and hinders scientific work due to inaccurate documentation. Thus, new cost-effective and contactless methods for object detection are urgently needed.

Objective: The goal of this work was to develop and evaluate a contactless visual recognition system for tracking medical consumable materials in ICUs using a deep learning approach on a distributed client-server architecture.

Methods: We developed Consumabot, a novel client-server optical recognition system for medical consumables, based on the convolutional neural network model MobileNet implemented in Tensorflow. The software was designed to run on single-board computer platforms as a detection unit. The system was trained to recognize 20 different materials in the ICU, while 100 sample images of each consumable material were provided. We assessed the top-1 recognition rates in the context of different real-world ICU settings: materials presented to the system without visual obstruction, 50% covered materials, and scenarios of multiple items. We further performed an analysis of variance with repeated measures to quantify the effect of adverse real-world circumstances.

Results: Consumabot reached a >99% reliability of recognition after about 60 steps of training and 150 steps of validation. A desirable low cross entropy of <0.03 was reached for the training set after about 100 iteration steps and after 170 steps for the validation set. The system showed a high top-1 mean recognition accuracy in a real-world scenario of 0.85 (SD 0.11) for objects presented to the system without visual obstruction. Recognition accuracy was lower, but still acceptable, in scenarios where the objects were 50% covered (P<.001; mean recognition accuracy 0.71; SD 0.13) or multiple objects of the target group were present (P=.01; mean recognition accuracy 0.78; SD 0.11), compared to a nonobstructed view. The approach met the criteria of absence

XSL•FO
RenderX

of explicit labeling (eg, barcodes, radio frequency labeling) while maintaining a high standard for quality and hygiene with minimal consumption of resources (eg, cost, time, training, and computational power).

Conclusions: Using a convolutional neural network architecture, Consumabot consistently achieved good results in the classification of consumables and thus is a feasible way to recognize and register medical consumables directly to a hospital's electronic health record. The system shows limitations when the materials are partially covered, therefore identifying characteristics of the consumables are not presented to the system. Further development of the assessment in different medical circumstances is needed.

(JMIR Med Inform 2019;7(4):e14806) doi:10.2196/14806

KEYWORDS

convolutional neural networks; deep learning, critical care; intensive care; image recognition; medical economics; medical consumables; artificial intelligence; machine learning

Introduction

A large amount of medical materials are consumed in the daily routine of intensive care units (ICUs). It is estimated that 85% of their healthcare costs are captured by three cost blocks, namely staff, clinical support services, and consumable medical materials [1,2]. The latter include, for example, sterile disposable material (eg, venous catheters or scalpels), material for body care (eg, absorbent pads, disposable flaps), or small materials (eg, needles, swabs or spatulas). A study of the International Programme for Resource Use in Critical Care (IPOC) quantified the daily cost for disposables in four European countries between 139.5€ (152.9 United States Dollars [USD]) (104.9€ [115 USD]–177.2€ [194.2 USD]) and 29.6€ (32.4 USD) (17.5€ [19.2 USD]–59.7€ [65.4 USD]), for drugs and fluids between 183.3€ [200.9 USD] (150.6€ [165 USD]–217.4€ [238.3 USD]) and 65.3€ (71.6 USD) (42.2€ [46.2 USD]–91.5€ [100.3 USD]) per patient [1]. Materials are often stored centrally in a departmental location, such as a materials warehouse or a centralized room, from which only the daily required material is taken and stored in proximity to patients. This is particularly relevant for infectious patients, as possibly contaminated material may have to be disposed of for hygienic reasons when the patient is discharged.

Due to the complexity of an ICU treatment, the financing of ICUs is often based on a flat-rate reimbursement scheme [3], resulting in a fixed daily hospital reimbursement for each day on the unit, not taking into account the reason for admission, the disease, or the resulting expenditure. In this scheme, individual medical or nursing measures are only remunerated with an additional fee in special cases (eg, blood transfusions or very complex interventions). As the consumption of the above-mentioned material has also been financed from the reimbursed lump-sum payment, this means the consumption of disposable medical material can hardly be recorded on a patient-related basis. It is largely unknown how many materials are needed for a single patient with a specific disease, so it is therefore not possible to carry out analyses in this respect even though these questions are highly relevant in daily practice. This is particularly true for storage and investment, as suboptimal management results in unnecessarily high storage maintenance costs. Furthermore, from a scientific perspective, timely and accurate documentation of medical consumables is critical, as it is especially noticeable in daily scientific work when medical measures are not documented in a timely manner. For example, the administration of an infusion solution is usually not documented until several minutes after the start of the procedure [4]. This makes retrospective data analysis (eg, in the field of machine learning) considerably more difficult as action and reaction are often critically time-linked, thus making the scientific evaluation of measures and data analysis significantly more time-consuming [5].

In some hospitals, this problem is solved by scanning a material-specific barcode when a disposable material is used at the patient site, which enables patient-specific billing. However, this means each article needs to be marked with an individual code. The use of wireless radio-frequency identification (RFID) has been assessed in the circumstances of patient care, however this requires the installation of specialized reader infrastructure and related management systems [6]. However, applying the previous techniques for identification is impossible for a relevant part of the materials (such as swabs, needles) because of their physical structure or their low price in relation to the tagging technique. An intelligent, cost-effective solution is urgently needed for this problem, and the developed solution must suit the specific needs of the ICU with the following characteristics: (1) recognizes materials without explicit labelling (eg, barcodes, radio frequency labelling); (2) fulfills the high standards for quality and hygiene of ICUs (ideally minimizing touch-interactions); and (3) has minimal consumption of resources (eg, cost, time, training, and computational power) [7]. Most notably, the system must fulfill the high data protection requirements of the sensitive area of intensive care medicine.

The aim of this work is to develop and evaluate Consumabot, a novel client-server recognition system for medical consumable materials based on a convolutional neural network, as an approach to solve the above-mentioned challenges in the sector of intensive care medicine. We first described the technical background (hardware and machine learning), taking the specific limitations and challenges of the ICU sector into account. In a proof-of-concept study, we then evaluated the performance of the system in the adverse circumstances of a real ICU environment, assessing the feasibility of the application in a real-world hospital setting.

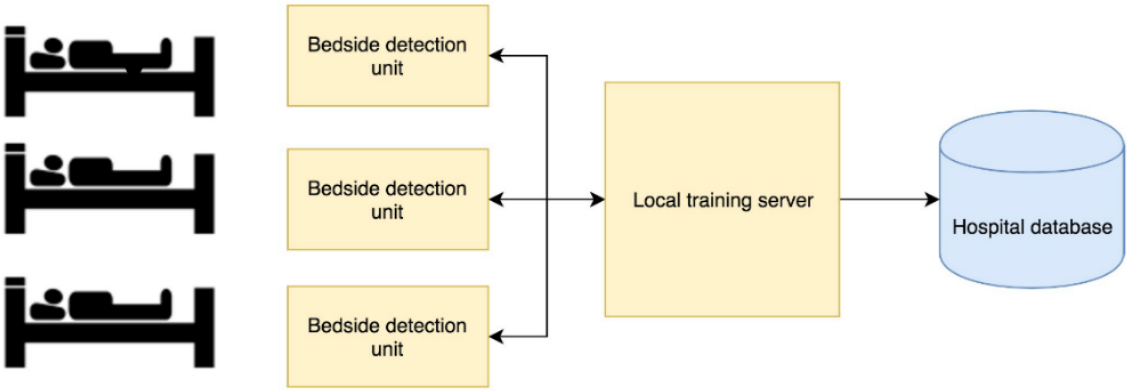
XSL•FO
RenderX

Methods

General Layout

We performed a thorough analysis on the technical, medical, and economic circumstances of ICUs and defined the specific requirements of the system. This included the identification of the need for a visual and contact-free recognition of the consumables, and for detection in proximity to the patient bed to facilitate assignment to the individual patient. Based on these considerations, the distributed concept of Consumabot was developed as follows: multiple low-cost detection units are located close to the patient bed, then these units are wirelessly connected to a local training server with high computational power for model training. This server has a direct connection to the hospital database and the electronic health record (EHR) backend (Figure 1).

Figure 1. Client-server setup between bedside detection units, local training server, and hospital database of the Consumabot system.

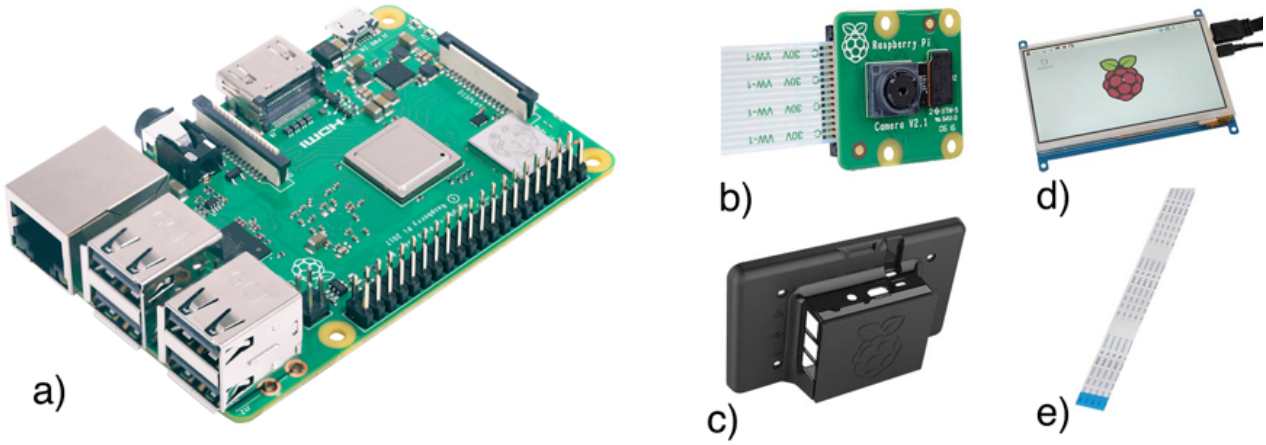


Hardware Setup

For the bedside detection units (clients), the commercially available single board computer Raspberry Pi was chosen as the hardware platform. Raspberry Pi is an inexpensive and widespread, small, single-board computer, initially developed to promote teaching of basic computer science in schools and in developing countries. The system architecture was developed at the University of Cambridge and is now being promoted by the charitable foundation the Raspberry Pi Foundation [8]. With low power consumption, no need for active cooling, and designed for continuous operation over several weeks, it is the ideal platform for the bedside detection units. Due to the popularity of the system, versatile extensions are available (eg, a powerful camera and hygienic housings). Indeed, Raspberry Pi has seen widespread use in different Internet of Things (IoT) applications in the healthcare sector [9-11].

In this work, we specifically used the Raspberry Pi 3 Model B for the detection units (clients), equipped with a quad-core processor ARM Cortex A53, a network and wireless network card, and one gigabyte of random accessible memory. The official camera module of the Raspberry Pi Foundation version 2.1, with 8-megapixel resolution (Figure 2b), as well as a touch screen monitor module with 7-inch display for interaction with the software (Figure 2d) were used. The whole setup was installed into a stable polyethylene housing (Figure 2c), and the camera was positioned using flat ribbon cable (Figure 2e). The local recognition modules could not be used for the training of the neural network since the computational capabilities of the processors are too low, thus resulting in a long training time.

Figure 2. Hardware setup of the recognition module on the Raspberry Pi computational platform: a) Raspberry Pi 3 Model B; b) camera module version 2.1; c) Polyethylene housing; d) touch screen monitor module with 7 inch display diagonal; and e) Flat ribbon cable.



XSL•FO
RenderX

Consequently, a more computationally capable training server was set up. This server was equipped with an Intel Xeon Gold 6140 processor with four dedicated processor cores, had 40 gigabytes of storage space on a single-state hard disk and had 320 gigabytes of storage space on a conventional magnetic hard disk for the resulting training data. Both the recognition module and the server used the open source operating system Linux in the Debian distribution [12].

Training Setup and Machine Learning Scheme

The software backbone of Consumabot was developed in Python, a programming language often used in the field of machine learning. One major advantage of Python is the availability of a wide range of machine learning tools like *NumPy*, used in data preprocessing, *scikit-learn*, used for data mining, or *Keras* as a high-level neural network interface. For modelling the machine learning backend, we adapted the model of a convolutional neural network (CNN) [13]. The convoluted (folded) structure makes CNNs particularly suitable for processing visual information, especially in the fields of image recognition and classification [14-16]. However, manual development of a neural network is very time-consuming, so software frameworks in which essential mathematical and preprocessing steps have already been developed are often used. Consumabot uses the software library *TensorFlow*, a software framework that simplifies the programming of data stream–centered procedures [17], and several adapted programming code elements for retraining image classifiers were included into Consumabot's source code [18]. Since the training of a full neural network is a complex and computationally intensive process, we applied a technique called transfer learning, a machine learning method where a model developed for a task is reused as the starting point for a model on a second task [19]. In transfer learning, basic processing image recognition steps, such as the recognition of edges, objects, and picture elements, are already trained in many iteration steps while the classification task is only assigned to the neural network in its final step, which is analogous to the training of an infant. However, it is important to note that the quality of the classification depends on the specificity of the training of the respective net. Thus, a neural network trained on images achieves better results in this domain than in the domain of something like natural language processing, resulting in the need to choose a task-specific, suited, pretrained network.

In the first step of the training the bottlenecks were generated, and they are the layer of the network that is located directly below the output layer [20]. Since each image is used several times, the bottlenecks do not change during the training as they can be created once and stored temporarily. In the second step, a set of random images from the training data set, with associated bottlenecks, were selected and placed in the output layer. The network-classified predictions are then compared with the correct classification, adjusting the weighting of the layers backwards (backpropagation) [21]. Classification accuracy and training progress were tracked with Tensorboard, a software designed for monitoring the training of neural networks [17,21].

Classification and Feedback

The internal structure of the system must take the particularities of the ICU environment into account. The detection module's camera took a picture every second and stored it on a memory card. This image from the camera was presented to the trained model of the recognition unit, which then predicted the probabilities of the recognized objects. If anything other than an empty surface was detected then the recognized object with the greatest probability was selected and presented to the user. After pressing the *store in database* button on the screen, the material was then registered in the database of the training server. This resulted in an additional confirmation step of human classification, as only images confirmed to have been classified correctly were included in further training. To facilitate online learning, at regular intervals the stored images of correctly recognized materials were transferred as training data to the training server and the model of the neural network was trained. The resulting model (the retrained graph) was distributed among the recognition units, which enabled an improved recognition of the desired consumables. Finally, the database of the control server was used to further process the data for either analysis or optimization.

On-Site Study

After finishing the training, we installed the hardware on an ICU within the University Hospital Aachen. Testing the system in real ICU conditions is obligatory due to the specific lighting and environmental conditions. In a test series, the 20 objects specified in Textbox 1 were classified using the camera of the recognition unit. Different rotations and orientations of the objects were chosen to correspond to the realistic field of application. To simulate the adverse circumstances of typical clinical workflow, we simulated a total of three scenarios: (1) scenario one, where the material was presented without any visual obstruction to the detection unit; (2) scenario two, where the material was 50% covered to simulate a visual obstruction during the routine clinical workflow; and (3) scenario three, where a secondary material (skin disinfection bottle) was present in the visual field while the material was presented without visual obstruction.

Each material was presented 10 times to the system. An object was classified as correct if it correctly appeared on the screen as the most probable classification (top-1 accuracy).

A full video of the hardware setup and training process can be found in Multimedia Appendix 1.

We performed a repeated measures, one-way analysis of variance (ANOVA) with Geisser-Greenhouse correction and a Tukey's multiple comparisons test with individual variances computed for each comparison. This was performed to assess the effect of adverse conditions in scenario two and three in comparison to the nonobstructed view in scenario one. All calculations were performed using GraphPad Prism 8.1.2 (GraphPad Software Inc, San Diego).

XSL•FO
RenderX

Textbox 1. Selection of medical consumables.

- Disposable bag valve mask
- Ampoule
- AuraOnce laryngeal mask
- Berotec inhalator
- Hand disinfection bottle
- Documentation sheet
- Boxed Dressings
- Packaged Gauze bandage
- Unpackaged Gauze bandages
- Gelafundin infusion solution
- Intravenous access orange
- Tube set for infusion solutions
- Intravenous access grey
- Braun sterile syringe
- Green Molinea protective pad
- White Protective pad
- Oxygen mask
- Oxygen tubing for mask
- Infusion solution Sterofundin
- Empty scenario (reference)

Results

Principal Findings

We trained the system in the context of a real ICU, taking special lighting conditions and other circumstances into account. We randomly chose a total of twenty common medical consumables from diverse categories with various sizes and formats to train the system on-site (Textbox 1). An empty scenario where no materials were present was provided to the system as a reference.

Setup of the System and Training

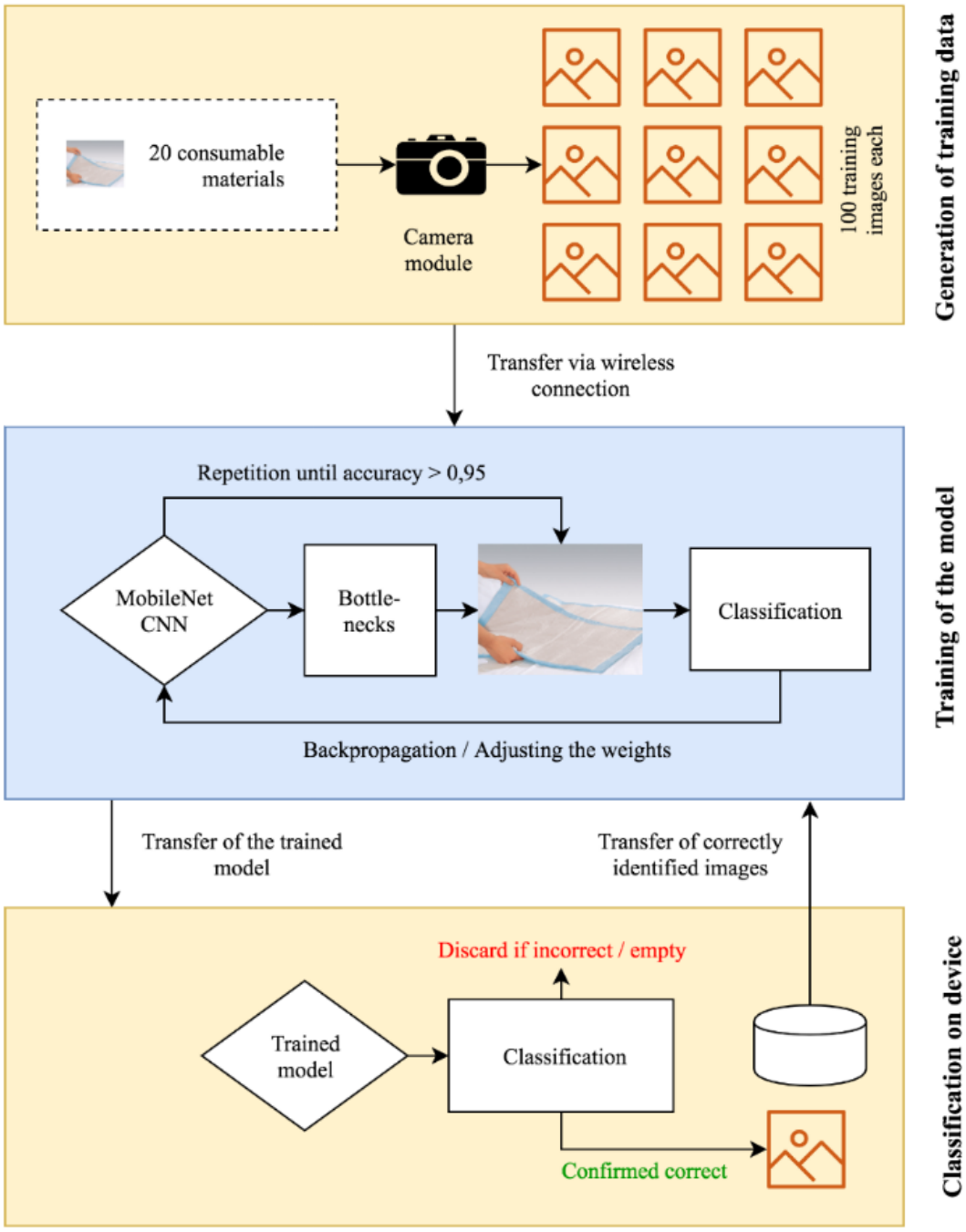
Figure 3 shows the overall setup of Consumabot. The initial training was carried out using a newly developed data generation script, generating a series of 100 images of each medical consumable to be recognized. A total of 2000 images were generated. For training, we used 1800/2000 (90%) images, and 200 images were randomly picked from this training set for validation. Finally, the remaining 200/2000 (10%) were picked for testing. We ran the system for 500 epochs, or training steps (training, validation, and testing), each epoch consisting of 100 randomly chosen images per item.

The top layer of the CNN received a 1001-dimensional vector as input for each image, and we trained a softmax layer on top of this vector representation [22]. Assuming the softmax layer contains N labels, this corresponds to learning of 1001 × N model parameters corresponding to the learned weights and biases.

For choosing the appropriate network, we took the recognition accuracy of different convolutional neural networks into account. For this purpose, we used the top-1 score [23]. Briefly, in this process, the predicted class multinomial distribution ( ) is obtained and compared to the appearance of the top classification as the target label (having the highest probability). The top-1 score is then computed as the times a predicted label matched the target label, divided by the number of data points evaluated. Selecting the correct model also needs to take the computational requirements into account, as the computational power of the recognition module is limited. Comparing MobileNet, AlexNet, GoogleNet and VGG16, we decided to apply a MobileNet, a class of efficient models for mobile and embedded vision apps, as a compromise between low requirements for computational power and high accuracy in image classification [24].

XSL•FO
RenderX

Figure 3. Overall setup of the Consumabot system. CNN: convolutional neural network.

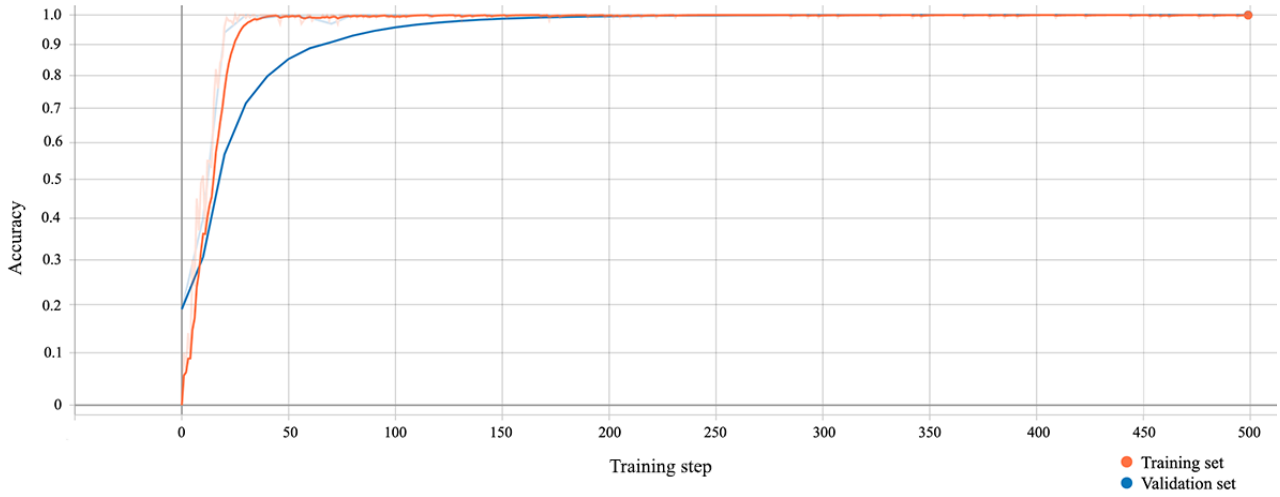


Evaluation of Training Setup and Machine Learning Scheme

Figure 4 shows the performance of the learning algorithm on the training and validation datasets. We observed a >99% accuracy of the model prediction after 60 training steps and 150 validation steps, as defined by the relation of true positives over the total number of occurrences. The prediction accuracy approached asymptotically and remained there during the training phase, consequently showing a high prediction accuracy. Of note, there was no sign of overfitting [25], a phenomenon indicated by an increase or constancy in training accuracy while there is an observed decrease in validation accuracy. Thus, since our model did not show any indication of over-learning, the generalization of the output of the learned model is applicable. The smoothing linear filter is explained in Multimedia Appendix 2.

XSL•FO
RenderX

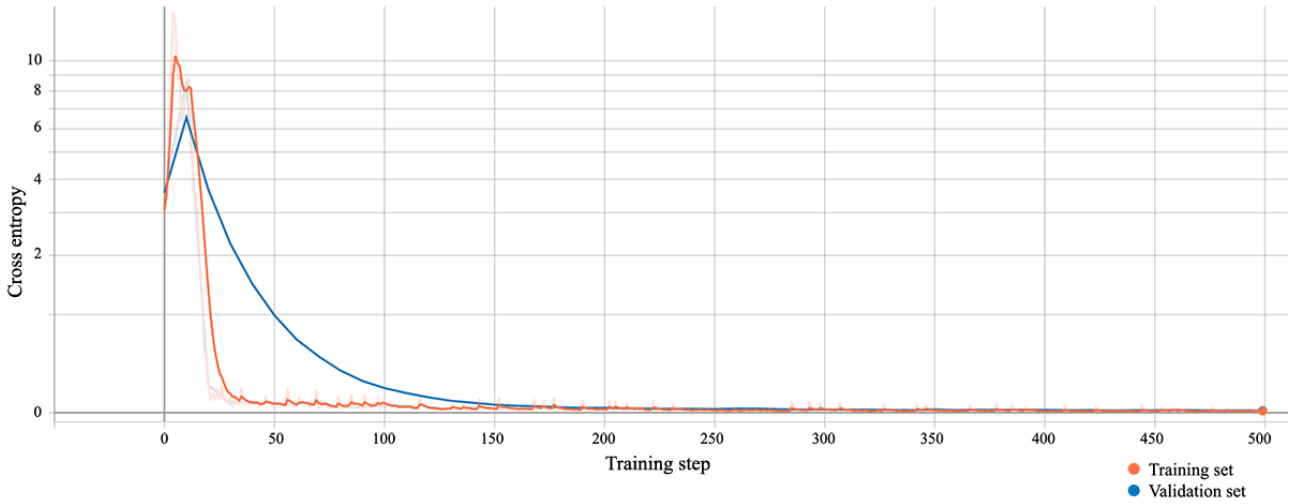
Figure 4. Accuracy of the model. The orange line represents the accuracy of correctly classified consumable material within the training set, while the blue line represents the accuracy of correctly classified consumable material within the validation set. A smoothing weight of 0.7 was applied and nonsmoothed curves are shown in pale orange and blue.



Like all common machine learning techniques, the used model also requires a cost function that has to be minimized. We used cross-entropy as a cost function [26]:

We applied this function to quantify the difference between the two probability distributions of the training and the validation set. As shown in Figure 5, our model revealed a desirable cross-entropy <0.03, with asymptotic stability after approximately 170 iteration steps in the validation set and 100 steps in the training set (Figure 5). As for Figure 4, the smoothing linear filter is explained in Multimedia Appendix 2.

Figure 5. Cross-entropy of the model during training. The orange line represents the entropy of the training set, while the blue line represents the entropy of the validation set. A smoothing weight of 0.7 was applied, and nonsmoothed curves are shown in pale orange and blue.



The fully trained model (graph visualization, retrained graph, label list) is provided in Multimedia Appendices 3 and 4.

On-Site Study

We then performed the on-site study at an ICU at the University Hospital Aachen, taking the particular lighting conditions and adverse circumstances due to clinical workflow into account. Each of the 20 consumable materials were presented ten times to the detection unit and were classified as correct if they appeared correctly on the screen with the highest associated probability (Table 1). The data generation for an entirely new consumable or for retraining, if a previously trained consumable material significantly changed its outer appearance, took approximately 100 seconds (1 second per picture, 100 pictures). We simulated adverse visual conditions as described in the methods section. For comparability reasons, no human feedback was included into the model training, and classification was based only on the first 100 training images. In Table 1, accuracy is provided in fractions of 1, as in a 0.7 recognition accuracy represents 70% (7/10) correct predictions.

XSL•FO
RenderX

Table 1. Top-1 recognition accuracy in the three scenarios.

Consumable material	Noncovered	Partially covered	Multiple materials
Bag valve mask	1	0.8	0.9
Ampoule	0.8	0.6	0.6
AuraOnce laryngeal mask	0.9	0.8	0.8
Berotec inhalator	0.7	0.8	0.7
Hand disinfection bottle	0.9	0.7	0.9
Documentation sheet	1	0.7	0.9
Boxed Dressings	0.8	0.7	0.7
Packaged Gauze bandage	0.8	0.7	0.8
Unpackaged Gauze bandage	1	0.9	0.8
Gelafundin infusion solution	0.8	0.7	0.8
Intravenous access orange	0.6	0.4	0.5
Tube set for infusion solutions	0.9	0.7	0.9
Intravenous access grey	0.8	0.6	0.7
Sterile syringe	0.9	0.7	0.8
Molinea protective pad green	0.8	0.8	0.7
White Protective pad	0.9	0.8	0.8
Oxygen mask	0.7	0.5	0.8
Oxygen tubing	0.9	0.6	0.9
Infusion solution Sterofundin	0.8	0.7	0.9
Empty scenario (reference)	1	1	0.8

In nonobstructed visual conditions, the model showed a good recognition performance, with a mean recognition accuracy of 0.85 (SD 0.11). Materials with large surface areas and many distinguishable visual features (eg, a disposable bag valve mask [mean recognition accuracy 1.0] or sterile syringes [mean recognition accuracy 0.9]) had particularly good detection rates. For materials only distinguishable by color (eg, intravenous [IV] accesses in different colors) Consumabot showed lower recognition accuracies for the grey IV access (mean 0.8) and the orange IV access (mean 0.6) (Figure 6, Table 1).

In a scenario where the surface area of the material was 50% covered, the system showed a lower, although still acceptable, mean recognition accuracy of 0.71 (SD 0.13). This was particularly true for materials with a small surface or with less distinguishable features (eg, for an oxygen tube), where the recognition accuracy dropped by 0.3 between when it was uncovered (mean 0.9) to when it was covered (mean 0.6).

Assessing the performance of the system in a scenario with multiple elements present in the scene resulted in a mean recognition accuracy of 0.78 (SD 0.11). For small elements compared to the secondary material present in a scene, this mostly resulted in a drop in recognition accuracy (eg, for a medication ampoule in the noncovered scenario [mean 0.8] versus a multiple element scenario [mean 0.6]).

We also performed an ANOVA with repeated measures to quantify the effect of the different scenarios, representing adverse real-world circumstances. Results of the ANOVA and other statistical analyses are given in Tables 2-4.

XSL•FO
RenderX

Figure 6. Results of the usability study in the context of a real ICU, real-world top-1 recognition accuracy of twenty sample materials. Top-1 recognition accuracy is provided in fractions of 1. Consumable materials: 1. AmbuBag (Disposable bag valve mask), 2. Ampoule, 3. AuraOnce laryngeal mask, 4. Berotec inhalator, 5. Hand disinfection bottle, 6. Documentation sheet, 7. Boxed Dressings, 8. Packaged gauze bandage, 9. Unpackaged gauze bandages, 10. Gelafundin infusion solution, 11. Intravenous access orange, 12. Tube set for infusion solutions, 13. Intravenous access grey, 14. Braun sterile syringe, 15. Molinea protective pad green, 16. White protective pad, 17. Oxygen Mask, 18. Oxygen tubing for mask, 19. Infusion solution Sterofundin, 20. Empty scenario (reference). ICU: intensive care unit.

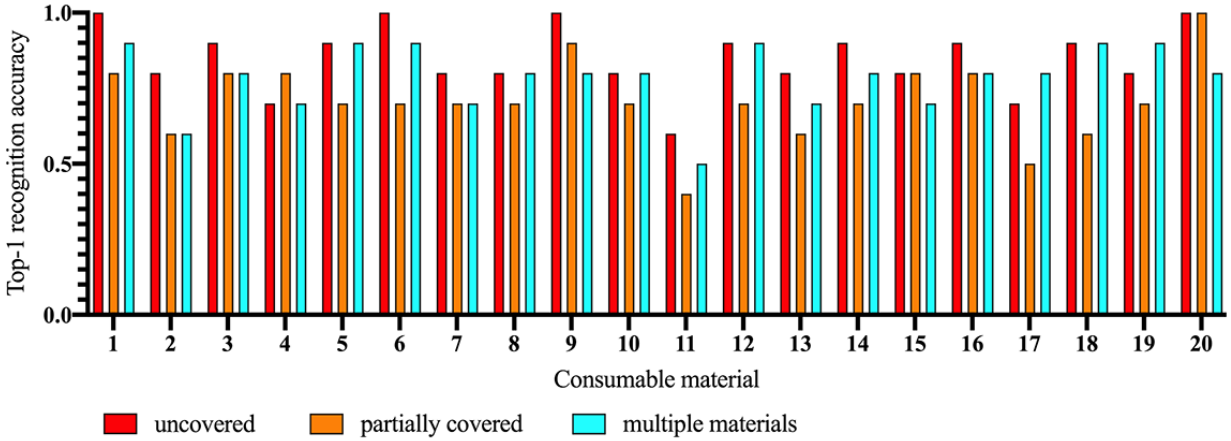


Table 2. Summary of the repeated measures ANOVA.

Test	F	*P* value	R^2	Geisser-Greenhouse's epsilon
ANOVA[a] summary	16.2	<.001	0.46	0.76
Matching effectiveness	4.89	<.001	0.57	—[b]

[a]ANOVA: analysis of variance

[b]Not applicable.

Table 3. Results of the Tukey's multiple comparisons test.

Comparisons	Mean difference	95% CI	Adjusted *P* value
Noncovered versus covered	0.14	0.08-0.2	<.001
Noncovered versus multiple	0.07	0.02-0.11	.001
Covered versus multiple	–0.075	–0.15 to 0.003	.06

Table 4. Detailed results of the repeated measures ANOVA.

Measures	Sum of squares	Degrees of freedom	Mean squares	F (DFn[a], DFd[b])	*P* value
Treatment (between columns)	0.20	2	0.1	F (1.5,29)=16	<.001
Individual (between rows)	0.56	19	0.03	F (19, 38)=4.9	<.001
Residual (random)	0.23	38	0.01	—[c]	—
Total	0.99	59	—	—	—

[a]DFn: degrees of freedom numerator

[b]DFd: degrees of freedom denominator

[c]Not applicable.

Independence of the observations among the groups, no sphericity, and a normal distribution were assumed for the analysis. The results of the performed ANOVA showed significant differences between the groups (F=16.2; *P*<.001; R^2=0.46), and *post hoc* analyses with Tukey's multiple comparisons test showed a significant difference between the noncovered cohort and the partially covered cohort (*P*=.001; 95% CI 0.08-0.2). Further significant differences between the noncovered scenario and the scenario with multiple consumables were also observed (*P*=.001; 95% CI 0.02-0.11), however, the differences between the noncovered group and multiple consumables were not statistically significant (*P*=.06).

XSL•FO
RenderX

Discussion

In this work we developed and evaluated Consumabot, a novel contactless visual recognition system for tracking medical consumable materials in ICUs using a deep learning approach on a distributed client-server architecture. In our proof-of-concept study in the context of a real ICU environment, we observed a high classificatory performance of the system for a selection of medical consumables, thus confirming its wide applicability in a real-world hospital setting.

Building on the foundation of fundamental mathematical research and technical progress, machine learning technologies today have the potential to drive ICUs towards a more sustainable, data-driven environment. In particular, contributions from other scientific disciplines, such as biology and engineering, have led to significant breakthroughs in quality and availability of neural networks, thus forming the backbone of Consumabot. The development of software for processing complex visual information is no longer a task requiring specialized hardware and software, as even the training of a complex neural network without specialist knowledge is possible now. This enables researchers and medical professionals alike to expand the use of artificial intelligence beyond today's commercial applications, such as in the fields of natural language processing [27,28], or intention, or pattern analysis [29], within constantly growing data volumes.

In this work we demonstrated the feasibility of the application of an on-device, deep learning–based computational platform for optical material recognition in the context of an ICU. Using a convolutional neural network infrastructure, the system Consumabot consistently achieved good results in the classification of consumables and thus is a feasible way to directly recognize and register medical consumables to a hospital's EHR system. Choosing a transfer learning technique based on MobileNet assured a fast training time while keeping steadily high recognition rates, achieving an optimal compromise of high accuracy and low computational requirements while maintaining a moderate model size. Using an optical recognition approach takes the specific conditions of the ICU into account, such as the need for a low maintenance, hygienic, contact-free solution. The use of MobileNet allowed us to apply Consumabot to the inexpensive, computationally weak, Raspberry Pi platform, while maintaining acceptable recognition speed. This confirmed the feasibility of use of the Raspberry Pi platform in healthcare, as described in multiple earlier works [11,30,31]. The upcoming increase in computational power of single-board computers could make the distributed client-server structure of the system unnecessary, as training could take place directly on the recognition units. This will enable its direct use in environments where no network connectivity is available (eg, rural areas), potentially facilitating scientific research in less developed medical infrastructure and healthcare systems. Thus, we believe that Consumabot will ultimately enable hospitals to reduce costs associated with consumable materials and consequently let them spend their resources on higher quality care (eg, by employing additional medical personnel).

Nevertheless, the conducted on-site study showed potential for optimization, particularly for standard medical consumables (eg, venous accesses of different sizes) since they did not show fully satisfactory recognition rates. This occurred if the distinguishing features were not clearly visible, partially covered, or multiple consumables were present in the scene. To solve this problem, a user training course with a note that identifying features must be clearly presented is recommended. Further, the model performance is likely to increase with added training data during daily use in the ICU due to the implemented feedback mechanism of Consumabot. Further development of the software and assessment of its performance with larger sets of medical consumables is desirable. The presentation of an object unknown to the detection unit results in a prediction with low confidence. In our prototype we display this confidence factor (see Multimedia Appendix 1) to the user. When using Consumabot in scenarios where many objects are unknown to the system (eg, if the system will only be used for detection of certain objects), the software should be adapted to only display predictions above a certain confidence threshold. In addition, the performance of multiple detection units running in parallel needs to be assessed as potential conflicts in the distribution of the model could occur.

Overall, though, the results were satisfactory enough to promote the further use and development of Consumabot in practice and research. The system fulfilled the requirements for recognizing materials without explicit labelling while maintaining the standard for quality and hygiene of ICUs. The system will make retrospective data analysis (eg, in the field of machine learning) considerably easier and enable time-critical research with direct correlation between action and reaction. The prototype's capabilities could potentially be enhanced by the integration of visual, multi-object detection algorithms, thus enabling it to detect a multitude of objects in parallel. Further, the need for tactile manual confirmation could be reduced by the integration of a microphone array to enable voice commands. The full source code of the detection unit, the pretrained model, and the training script have been released under the open source license Apache Version 2.0, January 2004 [32], and detailed assembly instructions have been released with the manuscript to encourage and enable other researchers to contribute to the development of the system and assess usability and feasibility in other use cases without increasing the financial burden of ICU patients [33].

Acknowledgments

This work has been funded by the European Institute of Innovation & Technology (EIT 19549).

XSL•FO
RenderX

Authors' Contributions

AP conceived the idea and developed the software and hardware prototype, and AH, AS, AP, GD, GM, LF, and LM interpreted the data. AP and LM wrote the manuscript and OS and LM edited the paper. OS, LM, and GM oversaw this project and provided ongoing feedback. All authors read and approved the final submitted manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demonstration video of the full Consumabot setup.
[MP4 File (MP4 Video), 50288 KB - medinform_v7i4e14806_app1.mp4]

Multimedia Appendix 2
Applied smoothing algorithm in Python.
[PDF File (Adobe PDF File), 7 KB - medinform_v7i4e14806_app2.pdf]

Multimedia Appendix 3
Retrained Tensorflow graph for materials used in experiment.
[ZIP File (Zip Archive), 4974 KB - medinform_v7i4e14806_app3.zip]

Multimedia Appendix 4
Labels for Tensorflow model.
[PDF File (Adobe PDF File), 116 KB - medinform_v7i4e14806_app4.pdf]

References

1. Negrini D, Sheppard L, Mills GH, Jacobs P, Rapoport J, Bourne RS, et al. International Programme for Resource Use in Critical Care (IPOC)--a methodology and initial results of cost and provision in four European countries. Acta Anaesthesiol Scand 2006 Jan;50(1):72-79. [doi: 10.1111/j.1399-6576.2006.00901.x] [Medline: 16451154]
2. Talmor D, Shapiro N, Greenberg D, Stone PW, Neumann PJ. When is critical care medicine cost-effective? A systematic review of the cost-effectiveness literature. Crit Care Med 2006 Nov;34(11):2738-2747. [doi: 10.1097/01.CCM.0000241159.18620.AB] [Medline: 16957636]
3. Ammar W, Khalife J, El-Jardali F, Romanos J, Harb H, Hamadeh G, et al. Hospital accreditation, reimbursement and case mix: links and insights for contractual systems. BMC Health Serv Res 2013 Dec 05;13:505 [FREE Full text] [doi: 10.1186/1472-6963-13-505] [Medline: 24308304]
4. Zucco L, Webb C. Improving the documentation of the daily review of patients in general intensive care. BMJ Qual Improv Rep 2014;3(1) [FREE Full text] [doi: 10.1136/bmjquality.u539.w496] [Medline: 26732891]
5. Wellner B, Grand J, Canzone E, Coarr M, Brady PW, Simmons J, et al. Predicting Unplanned Transfers to the Intensive Care Unit: A Machine Learning Approach Leveraging Diverse Clinical Elements. JMIR Med Inform 2017 Nov 22;5(4):e45 [FREE Full text] [doi: 10.2196/medinform.8680] [Medline: 29167089]
6. Wicks AM, Visich JK, Li S. Radio frequency identification applications in hospital environments. Hosp Top 2006;84(3):3-8. [doi: 10.3200/HTPS.84.3.3-9] [Medline: 16913301]
7. Kuhn T, Basch P, Barr M, Yackel T, Medical Informatics Committee of the American College of Physicians. Clinical documentation in the 21st century: executive summary of a policy position paper from the American College of Physicians. Ann Intern Med 2015 Feb 17;162(4):301-303. [doi: 10.7326/M14-2128] [Medline: 25581028]
8. Raspberry Pi. 2019. The Raspberry Pi Foundation URL: https://www.raspberrypi.org [accessed 2019-05-06] [WebCite Cache ID 78bq93AXV]
9. Gupta M, Patchava V, Menezes V. Healthcare based on IoT using Raspberry Pi. 2015 Oct 08 Presented at: Int Conf Green Comput Internet Things ICGCIoT; 2015; Noida, India p. 796. [doi: 10.1109/icgciot.2015.7380571]
10. Kumar R, Rajasekaran M. An IoT based patient monitoring system using raspberry Pi. 2016 Jan 07 Presented at: Int Conf Comput Technol Intell Data Eng ICCTIDE16; 2016; Kovilpatti, India p. 1. [doi: 10.1109/icctide.2016.7725378]
11. Fernandes CO, Lucena CJPD. A Software Framework for Remote Patient Monitoring by Using Multi-Agent Systems Support. JMIR Med Inform 2017 Mar 27;5(1):e9 [FREE Full text] [doi: 10.2196/medinform.6693] [Medline: 28347973]
12. Debian Website. Debian - The Universal Operating System URL: https://www.debian.org/index.en.html [accessed 2019-05-24] [WebCite Cache ID 78bqAQFtO]
13. Ciresan D, Meier U, Masci J, Gambardella L, Schmidhuber J. Flexible, High Performance Convolutional Neural Networks for Image Classification. In: IJCAI'11 Proceedings of the Twenty-Second international joint conference on Artificial

XSL•FO
RenderX

Intelligence - Volume Volume Two. 2011 Jul 16 Presented at: Twenty-Second international joint conference on Artificial Intelligence; 2011; Barcelona, Catalonia, Spain p. 1237-1242. [doi: 10.5591/978-1-57735-516-8/IJCAI11-210]
14. Brinker TJ, Hekler A, Utikal JS, Grabe N, Schadendorf D, Klode J, et al. Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review. J Med Internet Res 2018 Oct 17;20(10):e11936 [FREE Full text] [doi: 10.2196/11936] [Medline: 30333097]
15. Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ. editors. Adv Neural Inf Process Syst 25 Curran Associates, Inc; 2012. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf [accessed 2019-05-24] [WebCite Cache ID 78bqEfb9t]
16. Zhang Y, Allem J, Unger JB, Boley Cruz T. Automated Identification of Hookahs (Waterpipes) on Instagram: An Application in Feature Extraction Using Convolutional Neural Network and Support Vector Machine Classification. J Med Internet Res 2018 Nov 21;20(11):e10513 [FREE Full text] [doi: 10.2196/10513] [Medline: 30452385]
17. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A System for Large-Scale Machine Learning. In: OSDI'16 Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation. 2016 Nov 02 Presented at: 12th USENIX conference on Operating Systems Design and Implementation; November 02 - 04, 2016; Savannah, GA, USA p. 265-283. [doi: 10.1145/3190508.3190551]
18. Warden P. Pete Ward Blog. 2016. TensorFlow for Poets URL: https://petewarden.com/2016/02/28/tensorflow-for-poets/ [accessed 2019-05-09] [WebCite Cache ID 78bqkDu7K]
19. Torrey L, Shavlik J. Transfer Learning. Handbook of Research on Machine Learning Applications 2010;242:264. [doi: 10.4018/978-1-60566-766-9.ch011]
20. Sandler M, Howard A, Zhu M, Zhmoginov A. Cornell University. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks URL: http://arxiv.org/abs/1801.04381 [accessed 2019-05-09]
21. Hecht-Nielsen R. Theory of the Backpropagation Neural Network. In: Proceedings of the International Joint Conference on Neural Networks.: Neural Netw Percept Internet Academic Press; 1989 Presented at: International Joint Conference on Neural Networks; 1989; Washington, DC, USA. [doi: 10.1016/b978-0-12-741252-8.50010-8]
22. Liu W, Wen Y, Yu Z, Yang M. Large-margin Softmax Loss for Convolutional Neural Networks. In: ICML'16 Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. 2016 Jun 19 Presented at: International Conference on Machine Learning; 2016; New York City, USA.
23. Canziani A, Paszke A, Culurciello E. Cornell University. 2016 May 24. An Analysis of Deep Neural Network Models for Practical Applications URL: http://arxiv.org/abs/1605.07678 [accessed 2019-05-16]
24. Howard A, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. Cornell University. 2017 Apr 16. MobileNetsfficient Convolutional Neural Networks for Mobile Vision Applications URL: http://arxiv.org/abs/1704.04861 [accessed 2019-05-06]
25. Hawkins DM. The problem of overfitting. J Chem Inf Comput Sci 2004 Jan;44(1):1-12. [doi: 10.1021/ci0342472] [Medline: 14741005]
26. Rubinstein R, Kroese D. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. In: The Cross-Entropy Method. New York: Springer Science+Business Media; 2013.
27. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;11(5):392-402 [FREE Full text] [doi: 10.1197/jamia.M1552] [Medline: 15187068]
28. Denny JC, Choma NN, Peterson JF, Miller RA, Bastarache L, Li M, et al. Natural language processing improves identification of colorectal cancer testing in the electronic medical record. Med Decis Making 2012;32(1):188-197. [doi: 10.1177/0272989X11400418] [Medline: 21393557]
29. Rajkomar A, Oren E, Chen K, Dai A, Hajaj N, Liu P, et al. Cornell University. 2018 Jan 24. Scalable and accurate deep learning for electronic health records URL: http://arxiv.org/abs/1801.07860 [accessed 2018-02-02]
30. Ling Y, Ter Meer LP, Yumak Z, Veltkamp RC. Usability Test of Exercise Games Designed for Rehabilitation of Elderly Patients After Hip Replacement Surgery: Pilot Study. JMIR Serious Games 2017 Oct 12;5(4):e19 [FREE Full text] [doi: 10.2196/games.7969] [Medline: 29025696]
31. Jaiswal K, Sobhanayak S, Mohanta B, Jena D. IoT-cloud based framework for patients data collection in smart healthcare system using raspberry-pi. In: 2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA). 2017 Nov 21 Presented at: International Conference on Electrical and Computing Technologies and Applications (ICECTA); 2017; Ras Al Khaimah, United Arab Emirates. [doi: 10.1109/icecta.2017.8251967]
32. Apache License, Version 2. 0 Internet URL: https://www.apache.org/licenses/LICENSE-2.0 [accessed 2019-05-24] [WebCite Cache ID 78bqpaGmL]
33. Peine A. 2019. Bedside medical image recognition with TensorFlow and Raspberry Pi: arnepeine/consumabot Internet URL: https://github.com/arnepeine/consumabot [accessed 2019-05-24] [WebCite Cache ID 78bqqh9z7]

Abbreviations

ANOVA: analysis of variance
CNN: convolutional neural network

XSL•FO
RenderX

EHR: electronic health record
ICU: intensive care unit
IoT: Internet of Things
IPOC: International Programme for Resource Use in Critical Care
IV: intravenous
RFID: radio-frequency identification
USD: United States Dollar

Edited by G Eysenbach; submitted 25.05.19; peer-reviewed by A Davoudi, T Yulong; comments to author 19.06.19; revised version received 18.07.19; accepted 13.08.19; published 10.10.19.

Please cite as:
Peine A, Hallawa A, Schöffski O, Dartmann G, Fazlic LB, Schmeink A, Marx G, Martin L
A Deep Learning Approach for Managing Medical Consumable Materials in Intensive Care Units via Convolutional Neural Networks: Technical Proof-of-Concept Study
JMIR Med Inform 2019;7(4):e14806
URL: http://medinform.jmir.org/2019/4/e14806/
doi:10.2196/14806
PMID:31603430

©Arne Peine, Ahmed Hallawa, Oliver Schöffski, Guido Dartmann, Lejla Begic Fazlic, Anke Schmeink, Gernot Marx, Lukas Martin. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 10.10.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Opportunities and Challenges of Telehealth in Remote Communities: Case Study of the Yukon Telehealth System

Emily Seto[1,2*], PhD, PEng; Dallas Smith[3], BSW, MHSc; Matt Jacques[4], BA, PhD; Plinio Pelegrini Morita[1,2,5*], PhD, PEng

[1]Institute of Health Policy, Management, and Evaluation, University of Toronto, Toronto, ON, Canada

[2]Centre for Global eHealth Innovation, Techna Institute, University Health Network, Toronto, ON, Canada

[3]Health System Improvement and Transformation, Department of Health and Social Services, Government of Yukon, Whitehorse, YT, Canada

[4]Government Internal Audit Services, Executive Council Office, Government of Yukon, Whitehorse, YT, Canada

[5]School of Public Health and Health Systems, Faculty of Applied Health Sciences, University of Waterloo, Waterloo, ON, Canada

[*]these authors contributed equally

Corresponding Author:
Plinio Pelegrini Morita, PhD, PEng
School of Public Health and Health Systems
Faculty of Applied Health Sciences
University of Waterloo
200 University Avenue West
Waterloo, ON, N2L 3G1
Canada
Phone: 1 5198884567 ext 31372
Email: plinio.morita@uwaterloo.ca

Abstract

Background: Telehealth has been shown to improve access to health care and to reduce costs to the patient and health care system, especially for patients living in rural settings. However, unique challenges arise when implementing telehealth in remote communities.

Objective: The study aimed to explore the current use, challenges, and opportunities of the Yukon Telehealth System. The lessons learned from this study were used to determine important factors to consider when attempting to advance and expand telehealth programs in remote communities.

Methods: A mixed methods approach was used to evaluate the Yukon Telehealth System and to determine possible future advances. Quantitative data were obtained through usage logs. Web-based questionnaires were administered to nurses in each of the 14 Yukon community health centers outside of Whitehorse and patients who had used telehealth. Qualitative data included focus groups and semistructured interviews with 36 telehealth stakeholders.

Results: Since 2008, there has been a consistent number of telehealth sessions of about 1000 per year, with clinical care as the main use (69.06% [759/1099] of all sessions in 2015). From the questionnaire (11 community nurses and 10 patients) and the interview data, there was a consensus among the clinicians and patients that the system provided timely access and cost savings from reduced travel. However, they believed that it was underutilized, and the equipment was outdated. The following 4 factors were identified, which should be considered when trying to advance and expand a telehealth program: (1) patient and clinician buy-in: past telehealth experiences (eg, negative clinician experiences with outdated technology) should be considered when advancing the system. Expansion of services in orthopedics, dermatology, and psychiatry were found to be particularly feasible and beneficial in Yukon; (2) workflow: the use and scheduling of telehealth should be streamlined and automated as much as possible to reduce dependencies on the single Yukon telehealth coordinator; (3) access to telehealth technology: clinicians and patients should have easy access to up-to-date telehealth technology. The use of consumer products, such as mobile technology, should be leveraged as appropriate; and (4) infrastructure: the required human resources and technology need to be established when expanding and advancing telehealth.

Conclusions: While clinicians and patients had generally positive perceptions of the Yukon Telehealth System, there was consensus that it was underutilized. Many opportunities exist to expand the types of telehealth services and the number of telehealth sessions, including the expansion of services in several new specialty areas, updating telehealth equipment to streamline workflows

XSL•FO
RenderX

and increase convenience and uptake, and integrating novel technologies. The identified barriers and recommendations from this evaluation can be applied to the development and expansion of telehealth in other remote communities to realize telehealth's potential for providing efficient, safe, convenient, and cost-effective care delivery.

(JMIR Med Inform 2019;7(4):e11353) doi:10.2196/11353

KEYWORDS

health care systems; telemedicine; remote consultation; Yukon territory; telehealth; program evaluation; medical informatics

Introduction

Background

Telehealth services have become an integral part of health care in many jurisdictions within Canada and internationally [1-6]. Studies have found that telehealth can improve access to health services, improve health outcomes, reduce costs, and increase educational opportunities [7]. In particular, access to telehealth services for citizens in remote or underserved areas enables access to health care and programs that may otherwise be unavailable and reduces wait time, costs to the health care system, and personal expenses related to the patient's travel to reach health care services in urban centers [8,9]. Previous studies have found that telehealth increased access to health services, enhanced educational opportunities and social support, and improved health outcomes, quality of care, quality of life, and cost-effectiveness [7].

Telehealth provides substantial opportunities to improve health outcomes and service delivery, while reducing costs in regions such as Yukon, Canada, that have remote communities, centralization of health care services within a few cities, and significant reliance on out-of-territory clinical specialists [10,11]. However, the same remote nature of Yukon's communities outside of Whitehorse, with populations between 50 and 2500, provides some of the main challenges for the use and wide-scale deployment of telehealth technology.

The existing Yukon Telehealth System is used to serve the 38,450 inhabitants of Yukon (accurate as of 2017) [12], and is used for clinical care, clinician education, and administration. It comprises mobile telehealth units that are mainly used for clinical care and desktop telehealth software that is used for educational and administrative purposes. Each of the 14 community health centers has a single telehealth unit, and additional telehealth units are located in major centers such as Whitehorse. The system is managed by a single telehealth coordinator. Her duties include scheduling, initiating the scheduled telehealth sessions, technical support, and general oversight of the Yukon Telehealth System. Patients travel to one of the community health centers or other sites with telehealth units to participate in the scheduled telehealth sessions.

Objectives

The objective of this evaluation was to understand the current use, challenges, and opportunities of the current Yukon Department of Health and Social Services Telehealth System. The evaluation was initiated as part of the improvement initiatives focused on Mental Health, Addictions, and Chronic Conditions Support Program areas and aimed at providing a better understanding of the current state of the system that will enable the Yukon government to explore options to expand and advance the current system. The opportunities and barriers identified, as well as the provided recommendations, can be used to help guide other telehealth programs for remote communities to increase adoption and promote expansion of telehealth services.

Methods

Study Design Overview

A mixed methods approach was used to evaluate the Yukon Telehealth System and to determine the possible future advances to their existing infrastructure and services, combining interviews, focus groups, and site visits, with additional quantitative metrics using questionnaires. The Clinical Adoption Framework, which includes macroconstructs, mesoconstructs, and microconstructs that can influence the implementation and successful use of health care technologies, was used to guide the evaluation [13]. Data were collected between April 17, 2016, and August 2, 2016. The evaluation was aimed at answering the following 2 research questions:

1. What are the perceptions and use of the current Yukon Telehealth System?
2. What are the challenges and opportunities to improving the Yukon Telehealth System?

Quantitative Data

To determine the number and types of telehealth sessions that have occurred, usage logs were obtained and analyzed. Information on each telehealth session arranged by the telehealth coordinator was regularly logged in a comprehensive Excel (Microsoft Corporation) log file. Every record entered on the log file was time stamped and tagged using structured labels for the (1) type of call (administrative, educational, and clinical care), (2) location of the call initiation, (3) call duration, (4) location of external members (external to the territory), and (5) Yukon sites involved in the call. Other indicators such as more granular types of calls were logged using unstructured labels. The data were structured in a tabular format and logged using binary variables for each indicator described above.

These data were analyzed to determine usage patterns over the years, including the purposes of the telehealth sessions, number of sessions in different categories, and the location of the sessions. These usage logs were validated by comparing them with the data generated by the telehealth platform (Cisco Telepresence Management Suite) and with the telehealth billing data.

XSL•FO
RenderX

Community nurses and patients were asked to complete questionnaires regarding their use and perceptions of the telehealth system. A community nurse from each of the 14 Yukon communities, outside of Whitehorse with a telehealth unit, was sent an email with a link asking them to complete the Web-based questionnaire. During a regular visit to the community health center, patients who had previously used telehealth services were asked by the community nurses and clerks if they would be willing to voluntarily complete either a paper copy or Web-based version of a questionnaire. Completed paper copies of the questionnaires were mailed back to the evaluation team using a prestamped and addressed envelope. Both community nurses and patient questionnaires included a section enabling free-text comments. The questionnaire data were analyzed using descriptive statistics. Further statistical analysis was not possible owing to the small sample size of the responses to the questionnaires.

Qualitative Data

Individual semistructured interviews were conducted with stakeholders and users and 3 focus groups to gain an understanding of the current use, satisfaction, and perceived challenges and opportunities of the Yukon Telehealth System. Face-to-face interviews were conducted during site visits to Whitehorse, Carcross, and Dawson City. In total, 23 stakeholders participated in the face-to-face focus groups/individual interviews. The interview guide was developed to explore each relevant construct of the Clinical Adoption Framework [13].

In addition to the onsite interviews and focus groups, telephone interviews were conducted with (1) 4 community nurses, 1 each from Dawson City (population 1860), Watson Lake (population 1550), Beaver Creek (population 100), and Faro (population 400); (2) 9 physicians who provide services in Yukon (dermatologists, orthopedic surgeons, ophthalmologists, psychiatrists, and the chief of medical staff); (3) the manager of telehealth core services in British Columbia; and (4) 2 members of the Ontario Telemedicine Network. The outpatient services office manager at Whitehorse General Hospital referred the specialists for the interview as they were deemed as physicians who may be interested in using the telehealth services or who were already providing services in Yukon. The specialists were all out-of-territory health care providers as they flew periodically into Yukon to provide clinical services but were not residents of Yukon. In addition, a face-to-face interview was conducted with a previous manager of the telehealth services in Ontario. In total, 40 stakeholders of the Yukon Telehealth System and telehealth experts were consulted during this evaluation. A summary of the stakeholders interviewed is presented in Table 1.

Table 1. Study participants organized by methods and by groups used for data collection.

Method, roles or groups	Sample size, n
Face-to-face focus groups and interviews	
Yukon telehealth users	23 (community nurses, physicians, and administrators)
External telehealth specialists	1
Phone interviews	
Yukon specialists (from other provinces)	9 (dermatologists, orthopedic surgeons, ophthalmologists, psychiatrists, and the chief of medical staff)
Community nurses	4 (Dawson City, Watson Lake, Beaver Creek, and Faro)
Managers	1
External telehealth specialists	2

Furthermore, 2 researchers (ES and PPM) were present for all interviews and focus groups. Extensive notes were taken by the researcher who was not the main facilitator of the interview/focus group. All interviews and focus group sessions were also audio recorded but were not transcribed. A thematic analysis [14] was conducted, whereby the 2 researchers discussed the findings from each of the interviews/focus groups, immediately after each session, with the aid of the notes that were taken to determine emerging themes. The themes were added to the list of generated themes after each session. As the interview guide used for each of the interviews/focus groups followed the constructs of the Clinical Adoption Framework, the qualitative data largely followed the same format which facilitated thematic analysis. Any discrepancies in the interpretation of the interviews/focus groups were discussed until consensus was reached. After all the interviews/focus groups were completed, the 2 researchers convened to finalize the derived themes. The recordings were reviewed as necessary to refresh the memory of the researchers and to extract relevant quotes. Finally, the themes were presented and discussed with 2 other members of the study team (DS and MJ) who worked in the Health and Social Services of the Yukon government.

The quantitative and qualitative findings were triangulated through discussions among all authors to provide a comprehensive understanding of the current state of the system and the potential for an improved future Yukon Telehealth System.

Results

Current Use of the Yukon Telehealth System

The Yukon Telehealth System was first deployed in 2006 and has had no substantial upgrades since that time. Each of the 14 community health centers (Beaver Creek, Carmacks, Dawson City, Destruction Bay, Haines Junction, Mayo, Pelly Crossing,

XSL•FO
RenderX

Old Crow, Watson Lake, Teslin, Carcross, Ross River, Faro, and Whitehorse sites) has a single mobile telehealth unit; additional telehealth units are located in major centers such as Whitehorse and Dawson City. Other stationary telehealth units are mounted on the walls of boardrooms and meeting rooms in major centers (Whitehorse, Dawson City, etc). A complementary telehealth desktop application was installed on the computers used by the nurses in charge in the communities to attend meetings remotely and for peer consultations, while they were at their workstations. The desktop application was not used for clinical care with patients.

The Yukon Telehealth System was being used for 3 major purposes: (1) clinical care, (2) clinician education, and (3) administration. There has been a consistent number of telehealth sessions of approximately 1000 sessions per year since 2008, with the main use of the telehealth system being clinical care. The total number of calls for each year for the 3 categories (administration, educational, and clinical care), along with the total number of billed consultations in the Yukon for each year are presented in Table 2. The same data are presented in a graphical form in Figure 1. On an average, 4.61% (5345/115,867) of the specialist consultations were delivered through telehealth.

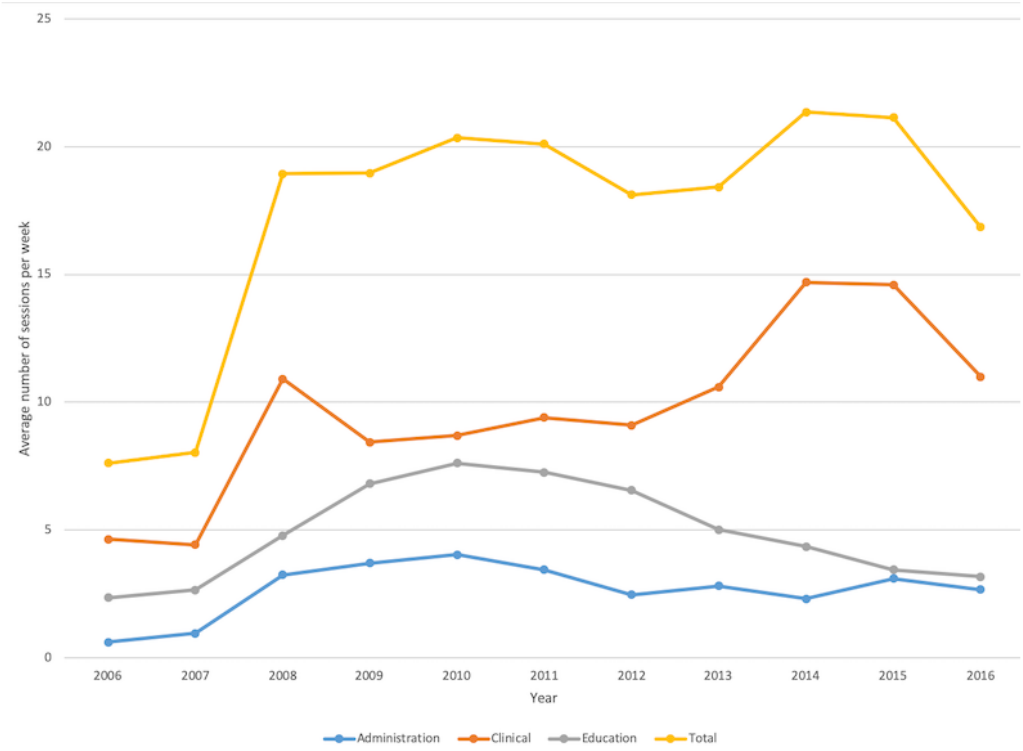
The specific use of the telehealth system was recorded by the telehealth coordinator on an Excel log file. A total of 676 unique labels existed on the dataset, with some labels only logging a couple of calls over the years. The top 10 specific reasons for the telehealth sessions are displayed in Table 3. The information presented in Table 3 provides the number of calls for each type of telehealth session, in addition to the corresponding percentage of yearly calls associated with that specific use. The top 10 uses of the telehealth platform shown in the table account for 40% to 50% of all telehealth calls for each year. Owing to the complexity of the log file and the lack of a systematic labeling structure, further analyses were not possible.

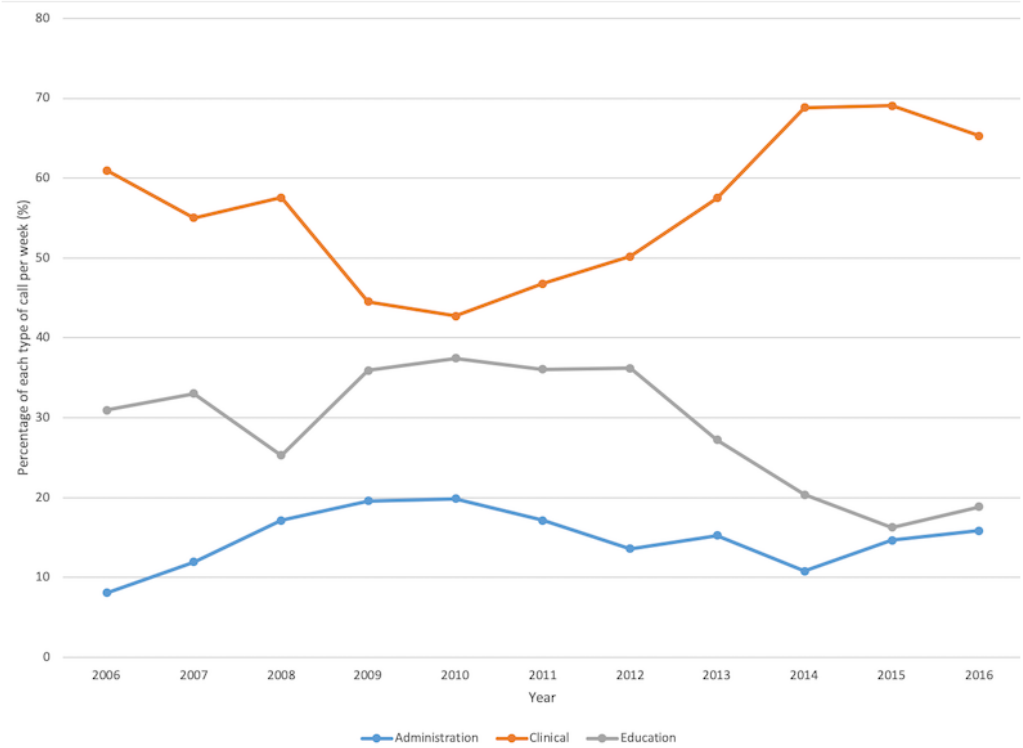
Table 2. The total number of calls per year for 3 different purposes (administrative, educational, and clinical care).

Year	Telehealth calls, n			Billed Consultations, n (% via telehealth)
	Administration	Clinical	Education	
2006	24	181	92	8975 (2.02)
2007	50	230	138	9681 (2.38)
2008	169	567	249	9453 (6.00)
2009	193	439	354	9120 (4.81)
2010	210	452	396	10,246 (4.41)
2011	179	489	377	10,619 (4.60)
2012	128	473	341	10,499 (4.51)
2013	146	551	261	11,595 (4.75)
2014	120	764	226	12,152 (6.29)
2015	161	759	179	11,708 (6.48)
2016	107	440	127	11,819 (3.72)
Total	1487	5345	2740	115,867 (4.161)

XSL•FO
RenderX

Figure 1. Telehealth usage by the number of sessions per week and the percentage of calls per week, collected from the log file generated by the telehealth coordinator and organized by the different types of calls.





XSL•FO
RenderX

Table 3. Specific use of the telehealth platform (top 10). Numbers presented in the table indicate the total number of sessions logged with that specific label and the percentage of that type of session for each year.

Reason for session	Year										
	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Clinical care, n (%)											
Diabetic education follow-up with patients	20 (7)	59 (13)	144 (15)	105 (11)	92 (9)	68 (7)	71 (8)	74 (8)	59 (5)	61 (6)	26 (4)
AA[a] meetings	0 (0)	1 (0)	22 (2)	41 (4)	46 (4)	45 (4)	51 (5)	46 (5)	50 (5)	51 (5)	35 (5)
Counseling (mental health patient/professional)	1 (0)	6 (1)	72 (7)	103 (10)	35 (3)	42 (4)	23 (2)	21 (2)	21 (2)	7 (1)	9 (1)
Occupational stress injury	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	16 (2)	1 (0)	53 (6)	106 (10)	97 (9)	55 (8)
Doctor's appointment (mental health)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	61 (6)	88 (8)	80 (7)	38 (6)
Interview (general interviews)	8 (3)	11 (2)	27 (3)	48 (5)	45 (4)	48 (5)	30 (3)	23 (2)	13 (1)	6 (1)	0 (0)
Cancer patient appointment	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	6 (1)	21 (2)	49 (5)	65 (6)	74 (7)	43 (6)
Surgeon's clinic (virtual surgeon's appointment with patients in community)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	90 (8)	76 (7)	62 (9)
Educational, n (%)											
Rural mental health supervision	0 (0)	2 (0)	40 (4)	20 (2)	57 (5)	29 (3)	14 (1)	18 (2)	22 (2)	17 (2)	9 (1)
Administrative, n (%)											
General meetings	16 (5)	35 (8)	91 (9)	83 (8)	107 (10)	82 (8)	52 (6)	76 (8)	63 (6)	85 (8)	34 (5)

[a]AA: Alcoholics Anonymous.

Perceptions of Clinicians Residing in Yukon

Clinicians residing in Yukon were generally satisfied with the telehealth system. In particular, users of the telehealth system cited the quality of the telehealth coordinator's work and commitment to the operation of the platform as the key factors in their satisfaction with the system. The clinicians residing in Yukon who were interviewed believed that telehealth had several benefits to the health care system, clinicians, and patients, including the following:

- Saving patients' time and money by reducing travel to urban centers and hospitals (ie, patients would not have to take as much time off from work to attend consultations).
- Saving government funds by not having to pay for the patients' travel expenses to go to urban centers for consultations that could have been delivered through telehealth.
- Improving the patients' quality of care by providing more timely and convenient access to clinical care.
- Preventing isolation as it was reassuring to clients to see the clinician's face when isolated in remote communities. Prevention of isolation was also cited as a benefit of using telehealth to connect social worker staff as isolation is one of the main contributors to burning out.
- Preventing unnecessary medevac cases and, therefore, reducing the need for nurses to leave the community health centers (nurses travel with patients to hospital in medevac cases).
- Multiple family members and care providers can be included in the same telehealth session, improving the overall awareness of the patients' conditions and care plans.
- Enabling community members to participate in programs not offered in their local community, such as Alcoholics Anonymous (AA) meetings.

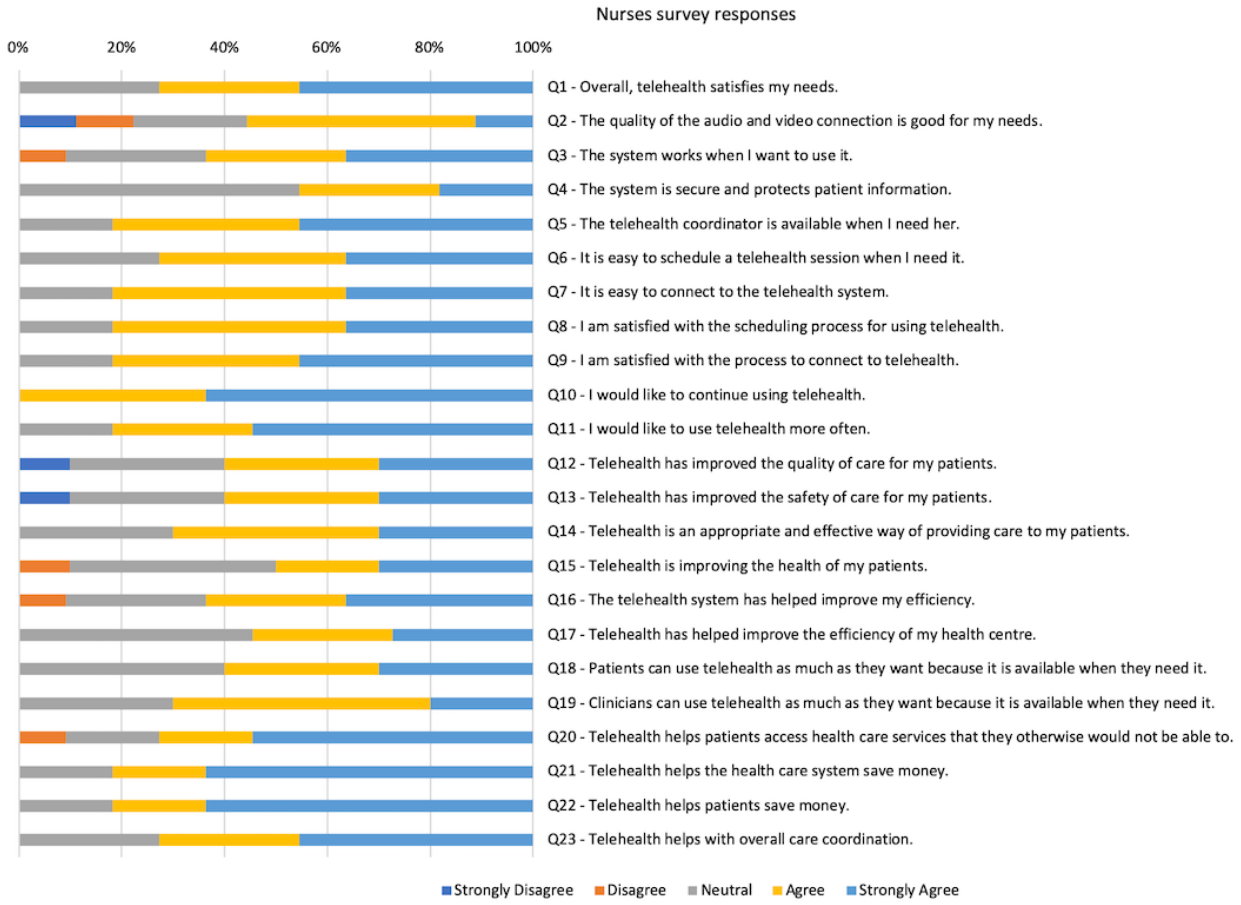
The questionnaire responses provided by the community nurses (11 nurses completed a questionnaire out of the 14 nurses that were invited to participate) supported the information gathered at the interviews in that the responses were generally neutral to satisfactory with the telehealth system (see Figure 2). However, the interviewed clinicians identified several limitations of the telehealth system, including the following:

- Suboptimal, outdated, and complex technology, combined with complex workflows to access and use the technology, resulting in the underutilization of the infrastructure for clinical purposes. A clinician commented, "I hope we could use it more, but the technology is not there yet." Another clinician stated, "The system seems old and clunky."
- Dependency on a single telehealth coordinator resulted in the loss of service quality and associated system knowledge when that individual was unavailable (vacation, sick leave, retirement, etc), even considering the existence of a backup person. One of the interviewed clinicians stated, "If we didn't have (the telehealth coordinator), we would not have a program."
- Nonexclusive telehealth rooms (rooms shared for meetings/boardrooms) resulted in telehealth units not being

XSL•FO
RenderX

available when needed, in addition to not providing a comfortable and conducive environment for the patient.
- Inadequate training of users of the telehealth technology led to an overreliance on the telehealth coordinator.
- Not all sessions or specialties are suitable for telehealth, as some may still require a face-to-face meeting for a comprehensive assessment of the patients' health.
- Limitations in the availability of telehealth units, lack of access of telehealth on their own desktops, and inconvenient setup process resulted in out-of-territory physicians spending more time to see a telehealth patient compared with an in-person visit by a patient to their office.
- Complexity and being unaware of the territories' billing process limited the willingness of the out-of-territory providers to use telehealth.

Figure 2. Results from the community nurse questionnaire. A total of 11 out of 14 nurses completed the questionnaire.



Out-of-Territory Specialists' Perceptions

The out-of-territory providers that were interviewed fell under 1 of the 4 specialties: dermatology, psychiatry, orthopedics, and ophthalmology. Each of the 4 specialties had different perspectives on the potential benefits of telehealth and different ideas on how to expand the use of the platform, as outlined below:

- Dermatology: The out-of-territory providers mentioned that dermatology was an area that could greatly benefit from the use of telehealth in terms of store-and-forward and live consultations in conjunction with inspection cameras.
- Psychiatry: Telehealth could be extremely useful to connect to Yukon clients between the psychologist's visits to Whitehorse, and it could also prevent clients from driving long distances to come for a consultation in Whitehorse. The benefits of telehealth for group sessions, such as AA meetings, were also highlighted.
- Orthopedics: Telehealth was viewed as a potential way for orthopedic surgeons to assess more Yukon citizens, to triage patients (ie, prioritize urgent cases) before a surgical consultation, and for postsurgical follow-ups. The presence of a physiotherapist with the client would enable the surgeon to assess whether the client should be scheduled for a face-to-face consultation in Whitehorse.
- Ophthalmology: A more systematic and secure method of sending ophthalmology images may be of use (store-and-forward), particularly for any future screening programs for diabetic clients. It was also believed that nurses trained to dilate the pupil could take the images locally and send the images along with relevant client information (eg, blood pressure, hemoglobin A1c, and blood glucose levels) for review. It was also suggested that the camera on mobile phones, particularly with an adapter lens, would be sufficient to take images for screening sessions that were noncritical. However, telehealth sessions between clients and ophthalmologists did not seem to be of benefit.

Although most specialists stated that they perceived the potential benefit of telehealth, 2 themes of barriers specific to the

XSL•FO
RenderX

specialists emerged from the interviews. The first barrier was that the specialists were already too busy and there was a lack of incentives to use telehealth. The specialists commented that they already had full schedules with wait lists for their clients in their home location. Therefore, there were not a lot of incentives (financial and otherwise) to start seeing more Yukon clients with telehealth. An additional complication was revealed by one orthopedic surgeon who discussed that he was only allowed to perform a certain number of surgeries on joints per year. The number of joint surgeries he performs in Yukon would get deducted from his quota of clients in his home province. The second barrier was with regard to the challenges with scheduling telehealth consultations. Many specialists commented on the difficulty of scheduling telehealth sessions between face-to-face consultations because of the timing (ie, clients being late or not showing up for telehealth sessions, time required to connect via telehealth, normal backlog of in-person clients, etc).

Patients' Perceptions

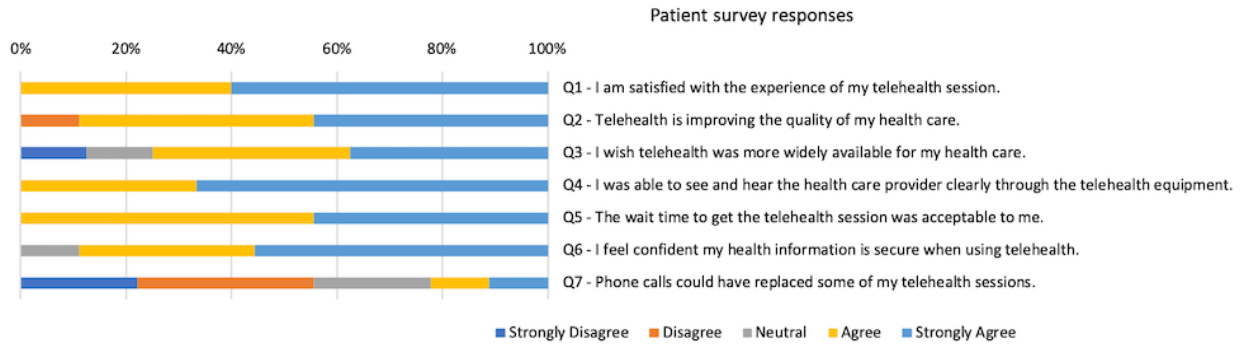
The patients' perspective was collected through patient questionnaires and indirect information provided by the clinicians. A total of 10 patients completed the questionnaire. The results from the patient survey are provided in Figure 3. Collecting additional data through patient interviews was considered infeasible by the evaluation team owing to the infrequent use of telehealth in each community and the remoteness of the communities likely leading to challenges in recruitment. In addition, the small population size of the remote communities could present challenges in assuring patient confidentiality during the interviews.

Patients perceived telehealth to be extremely important to the quality of their care and wanted it to be more widely available. In general, the patients were satisfied with their telehealth experiences, including the quality of the sessions, security, and wait times. Some quotes from patients who provided comments on the questionnaire, support the perceived benefits from telehealth:

> *I only have used telehealth for an AA meeting, but it is extremely significant to my recovery.* [Patient]
>
> *Excellent service for the communities.* [Patient]
>
> *I appreciate the convenience of going to the Health Centre versus driving an hour to Whitehorse for doctors' appointments.* [Patient]

The interviewed clinicians believed that the clients would want to use telehealth more often but were not aware that it was available. Instead, patients would often ask if it was possible to use Skype (Microsoft Corporation). Clinicians believed that many patients would find the reduction of travel to be the main benefit of telehealth, especially as many lived in remote areas and the weather in Yukon can be a barrier to traveling. By not having to travel to their health care appointments, clients would not need to take as much time away from work, which can be especially important depending on their jobs (eg, storekeepers and farmers). However, several clinicians mentioned that some clients would want to continue traveling to Whitehorse for their appointments as they are reimbursed for their travel and hotel costs, and that it is an opportunity for them to get chores (eg, shopping and visiting friends and family) accomplished while they are in Whitehorse for their consultations.

Figure 3. Results from the patient questionnaire. A total of 10 patients completed the questionnaire used in this study.



Discussion

Principal Findings

This evaluation explored the current use and perceptions of the Yukon Telehealth System, as well as the opportunities and challenges to improving the system. Although the clinicians' and patients' experiences with telehealth have been generally positive, there was a consensus that telehealth services were underutilized, which is in alignment with other studies in the field [11]. This qualitative finding was supported by the observed plateau in the number of telehealth sessions since 2008, which can be seen in Figure 1. Several factors could potentially explain the occurrence of such a plateau, which include (1) the system reaching the maximum capacity to handle calls, (2) limited capacity of the telehealth staff to handle more calls, (3) limited access to telehealth-equipped locations to initiate and receive calls, (4) limited interest by physicians and practitioners to use the telehealth system, or (5) limited awareness about the telehealth services and capabilities. Such results indicate that the limitations are likely at the system level, providing several opportunities for improvement as discussed below. However, since 2012 there has been an increase in the use of telehealth for clinical purposes and a decrease in the use for educational purposes. There was no direct evidence on the reasons for this shift in the use of the Yukon Telehealth System, but it could be because of an increase in the desire to use telehealth for clinical purposes owing to an improved perception of its clinical benefits; however, this resulted in less availability of the system for educational purposes.

XSL•FO
RenderX

The main contribution of this case study is the determination of 4 overarching factors related to sustainability, quality improvement, and scalability that should be considered when attempting to advance and expand telehealth systems in remote communities such as in Yukon. These factors can be easily overlooked or not properly addressed, which can lead to the stagnation and underuse of telehealth programs, as was the situation in this case study. This case study provides specific examples of how these factors have impacted the growth and adoption of telehealth in a remote setting. These factors are generally aligned with the findings from other telehealth studies in similar environments [11,15]. Each of these 4 factors is discussed below with recommendations on how to address the identified issues. See Table 4 for a summary of the factors and recommendations.

Table 4. Overarching factors and recommendations for the expansion of telehealth systems.

Factor	Implications if not addressed	Recommendations
Patient and clinician buy-in	• Underutilization of the telehealth program owing to lack of interest, resulting in wasted resources.	• Leverage existing patient and clinician buy-in to focus on the specific applications of telehealth. • Consider population-specific social drivers and goals. • Capitalize on the existing clinical interest to identify clinical champions.
Workflow	• Telehealth sessions may take more time than face-to-face consultations, resulting in clinician frustration and decision to stop using telehealth. • Poor patient satisfaction with telehealth owing to scheduling delays.	• Ensure that scheduling and initiation to telehealth services are quick and easy (ideally directly between the provider and patient).
Access to telehealth technology	• Clinicians spending time relocating to another room or not having a suitable time slot for the telehealth session, leading to frustration and decision to stop using telehealth. • Telehealth being inaccessible to patients as they cannot physically get to the location of the telehealth site.	• Enable clinicians to provide telehealth services from their own offices with desktop solutions instead of relocating to other rooms. • In case there is a separate telehealth room, ensure that it is accessible at all times with a priority for telehealth use. • Provide options for patients to access telehealth from their own homes.
Infrastructure	• Lack of appropriate human resources and technological infrastructure can result in telehealth services being unavailable if staff are away (eg, become sick) or the technology has a point of failure.	• Ensure redundancy of telehealth staff (ie, do not solely rely on a single telehealth coordinator). • Ensure that the telehealth coordinator has the time and resources for quality improvement initiatives. • Develop detailed training and maintenance plans. • Consider multiple points of access to telehealth services for patients, such as through consumer mobile devices.

Patient and Clinician Buy-In

The overall perception of patients and clinicians, both from interviews and questionnaires, indicated perceived value to using telehealth in Yukon, which include reduced need to mobilize patients while creating opportunities to connect patients and physicians on a more regular basis. The desire to expand telehealth services was voiced by both clinicians and patients.

However, program evaluations should also consider population-specific social drivers and goals. For example, some patients may prefer to travel for clinical visits as it is an opportunity for paid travel to Whitehorse to perform other errands in the city. Other challenges also include past negative experiences with telehealth (eg, past difficulty to use or access telehealth services, connectivity issues, and scheduling conflicts), which directly influence their willingness to use the system in the future.

A major opportunity to expand the use of telehealth services is through clinical champions who have indicated particular interest in using telehealth, as demonstrated by Wade et al [16]. Our interviews found strong interest from champions in Yukon in the areas of orthopedics, dermatology, and psychiatry, highlighting a relevance for both physical (orthopedics and dermatology) and social (psychiatry) interactions with patients. These specialties have had recent significant developments in telehealth [17-20], which indicate that expansion in these areas in Yukon may be particularly beneficial. A study of future developments of telehealth in Western Australia found that their top 4 most needed telehealth services were wound care, emergency, psychiatry, and ophthalmology [21], which somewhat differed from our case study. This points to the potential differences in targeting applications of telehealth for implementation, depending on the current opportunities and clinical buy-in of the jurisdiction.

Workflow

One of the barriers identified by the clinicians was the current cumbersome workflow related to scheduling calls. Telehealth users expect the scheduling process to be seamless and easy,

XSL•FO
RenderX

similar to scheduling calls via other current communication channels, such as through the phone or Skype. The availability of consumer technology in the market that provides patients with a better experience is a significant driver for improvements in other services [22]. In the case of the Yukon Telehealth System, clinicians expect to be able to schedule a call directly with the patient, without having to go through telehealth coordinators. Anecdotal information collected during this study from clinicians that currently use telehealth in Yukon also indicate that patients share the same feeling and would benefit from the opportunity to take calls from the convenience of their own home, as explored by DelliFraine and Dansky [23] and Bensink et al [24].

This evaluation found that clinicians perceived that owing to complexities in the workflow and the limitations of the telehealth system, some telehealth calls require more time than a regular in-person consultation session, which is consistent with the findings in the literature [11]. To deliver a positive user experience to clinicians, an ideal telehealth platform should enable clinicians to initiate and receive calls from their own office and connect directly to the patient, minimizing uncertainties introduced by a cumbersome workflow and allowing users to initiate their own sessions. Although specific schedules for telehealth were identified as a condition for telehealth implementation in previous studies [11], our study found that establishing an easy scheduling procedure was a key factor to increase adoption.

For a telehealth system to be successful and support expansion, an improved and direct scheduling system should be implemented to deliver a more streamlined workflow. This scheduling system should allow direct and automatic scheduling by patients and clinicians to schedule their own sessions with a telehealth coordinator who can provide oversight for conflicts and prioritization.

Access to Telehealth Technology

In line with the workflow difficulties, getting physical access to telehealth units (rooms too far from their main workplace or rooms inaccessible at the time of the call) was also described as a significant issue. The Yukon telehealth equipment was usually in rooms that were used for multiple activities, and clinicians commented that on several instances, they were not able to use the technology as the space was booked for meetings or face-to-face consultations. Consequently, the clinicians' daily activities were disrupted if they had to relocate for telehealth sessions, which was compounded by cases in which patients did not show up for their telehealth session.

The solutions to some of these recurring issues are complementary to those identified in the workflow section, where the telehealth systems should provide expanded desktop-based telehealth services that would enable clinicians to make and receive calls directly from their offices and provide dedicated telehealth space for the telehealth units, when more specialized equipment is necessary (cameras, vitals sensors, etc). Numerous technologies in the market can provide secure, desktop-based telehealth services such as Jabber (Cisco Systems) and Skype for Business, among others. The use of desktop-based technology would enable physicians to schedule multiple sequential sessions directly from their office, which would minimize the impact on their face-to-face consultations and workload [25].

Patients have also shown a strong interest in being able to minimize the number of visits to their local community health center to receive telehealth, as a trip could be a hazardous endeavor in winter months in Northern Canada. The ability to connect via telehealth directly from home using their own home devices, such as mobile phones and tablets, would provide a significant improved experience for these patients and potentially reduce issues related to mobilizing sick patients to community health centers. Similar home-based services have been widely presented by other authors in the field, showcasing the widespread benefit of enabling patients to attend their visits from the comfort of their own home [26-31].

Infrastructure

The fourth factor relates to the infrastructure (human resources and technology) necessary to operate a telehealth system. The evaluation identified an understaffed telehealth team, with overdependence on a single telehealth coordinator. A telehealth assistant or a second coordinator could be employed to add redundancy for the current telehealth coordinator, improving the overall service quality for telehealth users. The automation of currently manual procedures (eg, scheduling and telehealth session initiation) could free up time for the telehealth coordinator to conduct continuous quality improvement initiatives that would improve the overall experience of telehealth users.

Auxiliary services that are often not considered when implementing a telehealth service include training and maintenance staff to support the telehealth units that are in remote communities. Owing to a high turnover rate of staff in remote locations, a training plan is particularly essential. Training and maintenance are especially important with regard to the Yukon Telehealth System because of its outdated equipment, which has components that can no longer be procured if broken.

The ideal shift in technology should include a combination of more modern telehealth units, telehealth desktop clients, and consumer mobile devices to be used based on the application and needs of the users. The use of consumer mobile devices for telehealth is a relatively new concept and a potential area of future research. An integrated, multiplatform system would deliver a much better experience to patients and clinicians, potentially increasing the usage of telehealth in Yukon and expanding the accessibility of the telehealth service to underserved groups. Such an integrated solution can potentially include modules for scheduling sessions, initiation of sessions, call tracking, and quality improvement.

Limitations

The limitations of this evaluation include an underrepresentation of the patient perspective owing to a low patient questionnaire response rate (10 patients completed the questionnaires) and the inability to interview patients as the potential burden was deemed to be too high. A more comprehensive evaluation of the patients' perspectives could have provided more

XSL•FO
RenderX

patient-driven issues for this evaluation. In addition, only 11 community nurses returned a completed questionnaire. However, considering that there were only 14 community nurses in total, the response rate was relatively high. Furthermore, the out-of-territory perspective from clinicians also had limited representation, as this study had access to only 4 types of out-of-territory specialists (dermatology, psychiatry, orthopedics, and ophthalmology), which corresponds to the clinical specialties that were already delivering telehealth or face-to-face care to Yukon citizens by out-of-territory specialists. Finally, the clinicians who agreed to be interviewed or to participate in focus groups may have been more interested or had a more positive opinion of telehealth than the clinicians who did not participate in the study.

Conclusions

This evaluation found that there are significant opportunities to improve and expand the Yukon Telehealth System, which has plateaued in the number of telehealth consultations since 2008. These opportunities include the expansion of services in several new specialty areas, updating telehealth equipment to streamline workflows and increase convenience and uptake, and integrating novel technologies such as telemonitoring, education tools, and online programs. This expansion would be facilitated by the current general positive perceptions of the Yukon Telehealth System by both patients and clinicians. The factors that have been historically challenging to expansion and should be considered while the system evolves include patient and clinician buy-in, workflow, access to telehealth technology, and infrastructure with regard to human resources and technology. These factors and the lessons learned from this case study can be valuable considerations for the development of new telehealth programs in remote communities and for programs that may have plateaued in use.

Acknowledgments

This study was funded by the Yukon Department of Health and Social Services. The authors would like to thank all the participants from the interviews and surveys who volunteered their time to enable the team to identify opportunities for improvement. This manuscript was approved for publication by the Yukon Department of Health and Social Services.

Conflicts of Interest

None declared.

References

1. Dorsey ER, Topol EJ. State of telehealth. N Engl J Med 2016 Jul 14;375(2):154-161. [doi: 10.1056/NEJMra1601705] [Medline: 27410924]
2. Chi N, Demiris G. A systematic review of telehealth tools and interventions to support family caregivers. J Telemed Telecare 2015 Jan;21(1):37-44 [FREE Full text] [doi: 10.1177/1357633X14562734] [Medline: 25475220]
3. Kvedar J, Coye MJ, Everett W. Connected health: a review of technologies and strategies to improve patient care with telemedicine and telehealth. Health Aff (Millwood) 2014 Feb;33(2):194-199. [doi: 10.1377/hlthaff.2013.0992] [Medline: 24493760]
4. Cox A, Lucas G, Marcu A, Piano M, Grosvenor W, Mold F, et al. Cancer survivors' experience with telehealth: a systematic review and thematic synthesis. J Med Internet Res 2017 Jan 9;19(1):e11 [FREE Full text] [doi: 10.2196/jmir.6575] [Medline: 28069561]
5. Bradford NK, Caffery LJ, Smith AC. Telehealth services in rural and remote Australia: a systematic review of models of care and factors influencing success and sustainability. Rural Remote Health 2016;16(4):4268 [FREE Full text] [Medline: 27817199]
6. Franscico K, Archer N. The impact of telehealth on mental healthcare in Canada. Am J Med Res 2016;3(2):59-83. [doi: 10.22381/AJMR3220163]
7. Jennett PA, Hall LA, Hailey D, Ohinmaa A, Anderson C, Thomas R, et al. The socio-economic impact of telehealth: a systematic review. J Telemed Telecare 2003;9(6):311-320. [doi: 10.1258/135763303771005207] [Medline: 14680514]
8. Geffen M, Gordon D, Chien E. Canada Health Infoway: Digital Health in Canada. 2011. Telehealth Benefits and Adoption: Connecting People and Providers Across Canada URL: https://www.infoway-inforoute.ca/en/component/edocman/resources/reports/333-telehealth-benefits-and-adoption-connecting-people-and-providers-full [accessed 2019-09-12]
9. Speyer R, Denman D, Wilkes-Gillan S, Chen Y, Bogaardt H, Kim J, et al. Effects of telehealth by allied health professionals and nurses in rural and remote areas: a systematic review and meta-analysis. J Rehabil Med 2018 Feb 28;50(3):225-235 [FREE Full text] [doi: 10.2340/16501977-2297] [Medline: 29257195]
10. Moffatt JJ, Eley DS. The reported benefits of telehealth for rural Australians. Aust Health Rev 2010 Aug;34(3):276-281. [doi: 10.1071/AH09794] [Medline: 20797357]
11. Gagnon MP, Duplantie J, Fortin JP, Landry R. Implementing telehealth to support medical practice in rural/remote regions: what are the conditions for success? Implement Sci 2006 Aug 24;1:18 [FREE Full text] [doi: 10.1186/1748-5908-1-18] [Medline: 16930484]

XSL•FO
RenderX

12. Canadian Institute for Health Information. 2019. Your Health System URL: http://yourhealthsystem.cihi.ca/hsp/indepth?lang=en&_ga=2.125232970.1344406657.1550247246-1101653056.1459967413#/theme/C99003/2/N4IgKgFgpgtlDCAXATgGxALlAYwPatQEMAHAZygBNNQAGGgNkxQFcoAaEOgFieVY7oBGXqwC-4oAAA [accessed 2019-09-12]
13. Lau F, Price M, Keshavjee K. From benefits evaluation to clinical adoption: making sense of health information system success in Canada. Healthc Q 2011;14(1):39-45. [doi: 10.12927/hcq.2011.22157] [Medline: 21301238]
14. Denzin NK, Lincoln YS, editors. The Sage Handbook Of Qualitative Research. Thousand Oaks, CA: Sage Publications; 2013.
15. Joseph V, West RM, Shickle D, Keen J, Clamp S. Key challenges in the development and implementation of telehealth projects. J Telemed Telecare 2011;17(2):71-77. [doi: 10.1258/jtt.2010.100315] [Medline: 21097563]
16. Wade V, Eliott J. The role of the champion in telehealth service development: a qualitative analysis. J Telemed Telecare 2012 Dec;18(8):490-492. [doi: 10.1258/jtt.2012.gth115] [Medline: 23209264]
17. Tensen E, van der Heijden JP, Jaspers MW, Witkamp L. Two decades of teledermatology: current status and integration in national healthcare systems. Curr Dermatol Rep 2016;5:96-104 [FREE Full text] [doi: 10.1007/s13671-016-0136-7] [Medline: 27182461]
18. Hubley S, Lynch SB, Schneck C, Thomas M, Shore J. Review of key telepsychiatry outcomes. World J Psychiatry 2016 Jun 22;6(2):269-282 [FREE Full text] [doi: 10.5498/wjp.v6.i2.269] [Medline: 27354970]
19. Shah K. Use of telemedicine for initial consultations in elective orthopaedics—results from a large volume centre. Foot Ankle Orthop 2017 Sep 18;2(3):2473011417S0003. [doi: 10.1177/2473011417S000366]
20. Nami N, Massone C, Rubegni P, Cevenini G, Fimiani M, Hofmann-Wellenhof R. Concordance and time estimation of store-and-forward mobile teledermatology compared to classical face-to-face consultation. Acta Derm Venereol 2015 Jan;95(1):35-39 [FREE Full text] [doi: 10.2340/00015555-1876] [Medline: 24889827]
21. Bahaadini K, Yogesan K, Wootton R. Health staff priorities for the future development of telehealth in Western Australia. Rural Remote Health 2009;9(3):1164 [FREE Full text] [Medline: 19663540]
22. Morita PP, Cafazzo JA. Challenges and paradoxes of human factors in health technology design. JMIR Hum Factors 2016 Mar 1;3(1):e11 [FREE Full text] [doi: 10.2196/humanfactors.4653] [Medline: 27025862]
23. Dellifraine JL, Dansky KH. Home-based telehealth: a review and meta-analysis. J Telemed Telecare 2008;14(2):62-66. [doi: 10.1258/jtt.2007.070709] [Medline: 18348749]
24. Bensink M, Hailey D, Wootton R. A systematic review of successes and failures in home telehealth: preliminary results. J Telemed Telecare 2016 Dec 2;12(3_suppl):8-16. [doi: 10.1258/135763306779380174]
25. Radhakrishnan K, Xie B, Jacelon CS. Unsustainable home telehealth: a Texas qualitative study. Gerontologist 2016 Oct;56(5):830-840. [doi: 10.1093/geront/gnv050] [Medline: 26035878]
26. Greenhalgh T, Shaw S, Wherton J, Vijayaraghavan S, Morris J, Bhattacharya S, et al. Real-world implementation of video outpatient consultations at macro, meso, and micro levels: mixed-method study. J Med Internet Res 2018 Apr 17;20(4):e150 [FREE Full text] [doi: 10.2196/jmir.9897] [Medline: 29625956]
27. Powell RE, Stone D, Hollander JE. Patient and health system experience with implementation of an enterprise-wide telehealth scheduled video visit program: mixed-methods study. JMIR Med Inform 2018 Feb 13;6(1):e10 [FREE Full text] [doi: 10.2196/medinform.8479] [Medline: 29439947]
28. Hanlon P, Daines L, Campbell C, McKinstry B, Weller D, Pinnock H. Telehealth interventions to support self-management of long-term conditions: a systematic metareview of diabetes, heart failure, asthma, chronic obstructive pulmonary disease, and cancer. J Med Internet Res 2017 May 17;19(5):e172 [FREE Full text] [doi: 10.2196/jmir.6688] [Medline: 28526671]
29. Taylor A, Morris G, Pech J, Rechter S, Carati C, Kidd MR. Home telehealth video conferencing: perceptions and performance. JMIR Mhealth Uhealth 2015 Sep 17;3(3):e90 [FREE Full text] [doi: 10.2196/mhealth.4666] [Medline: 26381104]
30. Sturesson L, Groth K. Effects of the digital transformation: qualitative study on the disturbances and limitations of using video visits in outpatient care. J Med Internet Res 2018 Jun 27;20(6):e221 [FREE Full text] [doi: 10.2196/jmir.9866] [Medline: 29950290]
31. Agboola SO, Ju W, Elfiky A, Kvedar JC, Jethwani K. The effect of technology-based interventions on pain, depression, and quality of life in patients with cancer: a systematic review of randomized controlled trials. J Med Internet Res 2015 Mar 13;17(3):e65 [FREE Full text] [doi: 10.2196/jmir.4009] [Medline: 25793945]

Abbreviations

AA: Alcoholics Anonymous

XSL•FO
RenderX

Edited by C Lovis; submitted 20.08.18; peer-reviewed by K Groth, B McKinstry, M Sharma; comments to author 27.12.18; revised version received 20.04.19; accepted 28.08.19; published 01.11.19.

Please cite as:
Seto E, Smith D, Jacques M, Morita PP
Opportunities and Challenges of Telehealth in Remote Communities: Case Study of the Yukon Telehealth System
JMIR Med Inform 2019;7(4):e11353
URL: http://medinform.jmir.org/2019/4/e11353/
doi:10.2196/11353
PMID:31682581

©Emily Seto, Dallas Smith, Matt Jacques, Plinio Pelegrini Morita. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 01.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation

Patrick Wu[1,2*], BS; Aliya Gifford[1*], PhD; Xiangrui Meng[3*], PhD; Xue Li[3], PhD; Harry Campbell[3], MD; Tim Varley[4], BSc; Juan Zhao[1], PhD; Robert Carroll[1], PhD; Lisa Bastarache[1], MS; Joshua C Denny[1,5], MD, MS; Evropi Theodoratou[3,6], PhD; Wei-Qi Wei[1], MD, PhD

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

[2]Medical Scientist Training Program, Vanderbilt University School of Medicine, Nashville, TN, United States

[3]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, United Kingdom

[4]Public Health and Intelligence Strategic Business Unit, National Services Scotland, Edinburgh, United Kingdom

[5]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States

[6]Edinburgh Cancer Research Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Wei-Qi Wei, MD, PhD
Department of Biomedical Informatics
Vanderbilt University Medical Center
2525 West End Ave, Suite 1500
Nashville, TN, 37203
United States
Phone: 1 615 343 1956
Email: wei-qi.wei@vumc.org

Abstract

Background: The phecode system was built upon the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) for phenome-wide association studies (PheWAS) using the electronic health record (EHR).

Objective: The goal of this paper was to develop and perform an initial evaluation of maps from the International Classification of Diseases, 10th Revision (ICD-10) and the International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM) codes to phecodes.

Methods: We mapped ICD-10 and ICD-10-CM codes to phecodes using a number of methods and resources, such as concept relationships and explicit mappings from the Centers for Medicare & Medicaid Services, the Unified Medical Language System, Observational Health Data Sciences and Informatics, Systematized Nomenclature of Medicine-Clinical Terms, and the National Library of Medicine. We assessed the coverage of the maps in two databases: Vanderbilt University Medical Center (VUMC) using ICD-10-CM and the UK Biobank (UKBB) using ICD-10. We assessed the fidelity of the ICD-10-CM map in comparison to the gold-standard ICD-9-CM phecode map by investigating phenotype reproducibility and conducting a PheWAS.

Results: We mapped >75% of ICD-10 and ICD-10-CM codes to phecodes. Of the unique codes observed in the UKBB (ICD-10) and VUMC (ICD-10-CM) cohorts, >90% were mapped to phecodes. We observed 70-75% reproducibility for chronic diseases and <10% for an acute disease for phenotypes sourced from the ICD-10-CM phecode map. Using the ICD-9-CM and ICD-10-CM maps, we conducted a PheWAS with a Lipoprotein(a) genetic variant, rs10455872, which replicated two known genotype-phenotype associations with similar effect sizes: coronary atherosclerosis (ICD-9-CM: P<.001; odds ratio (OR) 1.60 [95% CI 1.43-1.80] vs ICD-10-CM: P<.001; OR 1.60 [95% CI 1.43-1.80]) and chronic ischemic heart disease (ICD-9-CM: P<.001; OR 1.56 [95% CI 1.35-1.79] vs ICD-10-CM: P<.001; OR 1.47 [95% CI 1.22-1.77]).

Conclusions: This study introduces the beta versions of ICD-10 and ICD-10-CM to phecode maps that enable researchers to leverage accumulated ICD-10 and ICD-10-CM data for PheWAS in the EHR.

(JMIR Med Inform 2019;7(4):e14325) doi:10.2196/14325

XSL•FO
RenderX

KEYWORDS

electronic health record; genome-wide association study; phenome-wide association study; phenotyping; medical informatics applications; data science

Introduction

Background

Electronic health records (EHRs) have become a powerful resource for biomedical research in the last decade, and many studies based on EHR data have used International Classification of Diseases (ICD) codes [1]. When linked to DNA biobanks, healthcare information in EHRs can be a tool to help discover genetic associations by using billing codes in phenotyping algorithms. The phenome-wide association study (PheWAS) paradigm was introduced in 2010 as an approach that scans across a range of phenotypes, similar to what is done for the genome in genome-wide association studies. Studies using PheWAS have replicated hundreds of known genotype-phenotype associations and discovered dozens of new ones [2-12]. The initial version of phecodes consisted of 733 custom groups of ICD Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis codes. The most recent iteration of phecodes consists of 1866 hierarchical phenotype codes that map to 15,558 ICD-9-CM codes [13,14]. However, many health systems and international groups use the International Classification of Diseases, 10th Revision (ICD-10) or the International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM) codes [15], therefore necessitating a new phecode map.

Transition from ICD-9 to ICD-10

In 1979, the World Health Organization (WHO) developed ICD-9 to track mortality and morbidity. To improve its application to clinical billing, the United States National Center for Health Statistics (NCHS) modified ICD-9 codes to create ICD-9-CM, whose end-of-life date was scheduled around the year 2000 but was delayed until October 2015 [15]. In 1990, the WHO developed ICD-10 [16], which the NCHS used to create ICD-10-CM to replace ICD-9-CM.

Moving from ICD-9-CM to ICD-10-CM led to major structural changes in the coding system. First, the structure moved from a broadly numeric-based system in ICD-9-CM (eg, 474.11 for “Hypertrophy of tonsils alone”) to an alphanumeric system in ICD-10-CM (eg, J35.1 for the same condition). Second, ICD-10-CM contains much more granular information than ICD-9-CM, as seen with the approximately tenfold increase in the number of diabetes-related codes in ICD-10-CM. ICD-10-CM also differs from ICD-9-CM in terms of semantics and organization [15,17].

Compared to ICD-10, ICD-10-CM has even more codes and granularity. While the 2018AA Unified Medical Language System (UMLS) [18] contains 94,201 unique ICD-10-CM codes, it has 12,027 unique ICD-10 codes after exclusion of range codes (eg, ICD-10-CM A00-A09). Further, there are ICD-10 codes that do not exist in ICD-10-CM, and vice versa, like ICD-10 A16.9 “Respiratory tuberculosis unspecified, without mention of bacteriological or histological confirmation”, which has no ICD-10-CM equivalent.

Prior Work

To develop the original phecode system, one or more related ICD-9-CM codes were combined into distinct diseases or traits. For example, three depression-related ICD-9-CM codes, 311, 296.31, and 296.2, were condensed to phecode 296.2 “Depression”. With the help of clinical experts in disparate domains, such as cardiology and oncology, we have iteratively updated the phecode groupings [19].

The phecode scheme is unique because it has built-in exclusion criteria to prevent contamination by cases in the control cohort. This is an important feature, as case contamination of control groups decreases the statistical power for finding genotype-phenotype associations [20]. For each disease phenotype, we defined exclusion criteria by using our clinical knowledge and by consulting physician specialists.

An example for how users can use phecode exclusion criteria is illustrated by a type 2 diabetes study using EHRs. To define cases of type 2 diabetes, users include patients with ICD codes that map to phecode 250.2 “Type 2 diabetes”. To create the control cohort, they only include patients without phenotypes in the “Diabetes” group, which is comprised of phecodes in the range of 249-250.99. This prevents contamination of the control group by patients with diseases such as “Type 1 diabetes” (phecode 250.1) and “Secondary diabetes mellitus” (phecode 249). Excluded patients also include those with signs and symptoms commonly associated with type 2 diabetes, such as “Abnormal glucose” (phecode 250.4), which may indicate someone who has not yet been diagnosed with diabetes.

Though the phecode system is effective at replicating and identifying novel genotype-phenotype associations, PheWAS have largely been limited to using ICD-9-CM codes. A few studies have mapped ICD-10 codes to phecodes by converting ICD-10 to ICD-9-CM, and then mapping the converted ICD-9-CM codes to phecodes [3,10]. However, these studies limited their mappings to ICD-10 (non-CM) codes, did not provide a map to translate ICD-10-CM codes to phecodes, and did not evaluate the accuracy of these maps.

Study Goals

In this study, we developed and evaluated maps of ICD-10 and ICD-10-CM codes to phecodes. The primary aims of this study were to create an initial beta map to perform PheWAS using ICD-10 and ICD-10-CM codes and to focus the analyses on PheWAS-relevant codes. Our goal was to demonstrate that researchers should expect similar results from the ICD-10-CM phecode map compared to the gold-standard ICD-9-CM map. To accomplish this goal, we investigated phecode coverage, phenotype reproducibility, and the results from a PheWAS.

XSL•FO
RenderX

Methods

Databases

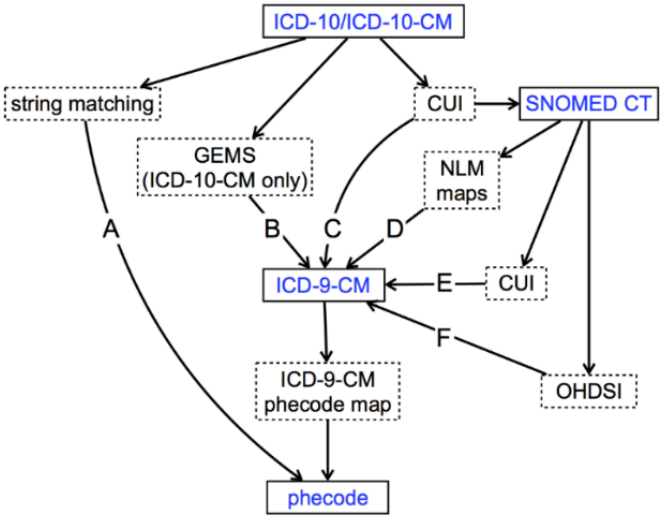
In this study, we used data obtained from the Vanderbilt University Medical Center (VUMC) and UK Biobank (UKBB) databases. The VUMC EHR contains clinical information derived from the medical records of >3 million unique individuals. The UKBB is a prospective longitudinal cohort study designed to investigate the genetic and environmental determinants of diseases in UK adults. Between 2006-2010, the study recruited >500,000 men and women aged 40-69 years. Participants consented to allow their data to be linked to their medical records. EHR records from the UKBB were obtained under an approved data request application (ID:10775).

We used VUMC data with >2.5 years of ICD-10-CM data (October 10, 2015 to June 1, 2017) for inpatient and outpatient encounters. Comparatively, we used UKBB data with >2 decades of ICD-10 data [21] (April 1, 1995 to March 31, 2015) for only inpatient encounters.

Mapping ICD-10-CM and ICD-10 Codes to Phecodes

We extracted ICD-10-CM codes from the 2018AA release of the UMLS [18] and used several automated methods to translate ICD-10-CM diagnosis codes to phecodes (Figure 1). We mapped 515 ICD-10-CM codes directly to phecodes by matching code descriptions regardless of capitalization (eg, ICD-10-CM H52.4 “Presbyopia” to phecode 367.4 “Presbyopia”). We mapped 82,287 ICD-10-CM codes indirectly to phecodes using the existing ICD-9-CM phecode map [14]. To convert ICD-10-CM codes indirectly to phecodes, we used General Equivalence Mappings (GEMS) provided by the Centers for Medicare & Medicaid Services that map ICD-10-CM to ICD-9-CM and vice versa [22]. We included both equivalent and nonequivalent GEMS mappings (ie, where the approximate flag was either 0 or 1). As an example of this indirect approach, to map ICD-10-CM E11.9 “Type 2 diabetes mellitus without complications” to phecode 250.2 “Type 2 diabetes,” we mapped ICD-10-CM E11.9 to ICD-9-CM 250.0 “Diabetes mellitus without mention of complication” to phecode 250.2.

Figure 1. Mapping strategy for ICD-10 (non-CM) and ICD-10-CM diagnosis codes to phecodes. We mapped ICD-10-CM codes directly by matching code descriptions (path A) or indirectly to phecodes, using a number of manually validated mapping resources (paths B, C, D, E, and F). In path D, we used NLM’s SNOMED CT to create ICD-9-CM one-to-one and many-to-one maps [23]. To map ICD-9-CM codes to phecodes, we applied Phecode Map 1.2 with ICD-9 Codes (ICD-9-CM phecode map) [14]. Boxes with solid lines indicate clinical terminologies, and those with dashed lines describe the resources and mapping methods used. ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification; CUI: Concept Unique Identifier; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms; GEMS: General Equivalence Mappings; NLM: National Library of Medicine; ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification; OHDSI: Observational Health Data Sciences and Informatics.



Since the GEMS do not provide ICD-9-CM mappings for all ICD-10-CM codes [17], we complemented this approach with UMLS semantic mapping [24], Observational Health Data Sciences and Informatics (OHDSI) concept relationships [25,26], and National Library of Medicine (NLM) maps [23]. In this approach to indirect mapping, we first mapped ICD-10-CM codes to Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) through UMLS Concept Unique Identifier (CUI) equivalents, which were then converted to ICD-9-CM through either UMLS CUI equivalents [18,24], OHDSI [25], or NLM maps [23]. For example, we mapped ICD-10-CM L01.00 “Impetigo, unspecified” to CUI C0021099 to SNOMED CT 48277006 to OHDSI Concept ID 140480 to OHDSI Concept ID 44832600 to ICD-9-CM 684 and finally to phecode 686.2 “Impetigo”.

There were two general instances when an ICD-10-CM code mapped to more than one phecode. First, some ICD-10-CM codes mapped to both a parent phecode and one of its child phecodes that was lower in the hierarchy. To maintain the granular meanings of ICD-10-CM codes, we only kept the mappings to child phecodes, a decision that we could make due to the hierarchical structure of phecodes. For example, ICD-10-CM I10 “Essential (primary) hypertension” was mapped to phecodes 401 “Hypertension” and 401.1 “Essential hypertension”, but we only kept the mapping to phecode 401.1. Second, we kept all the mappings for ICD-10-CM codes that

XSL•FO
RenderX

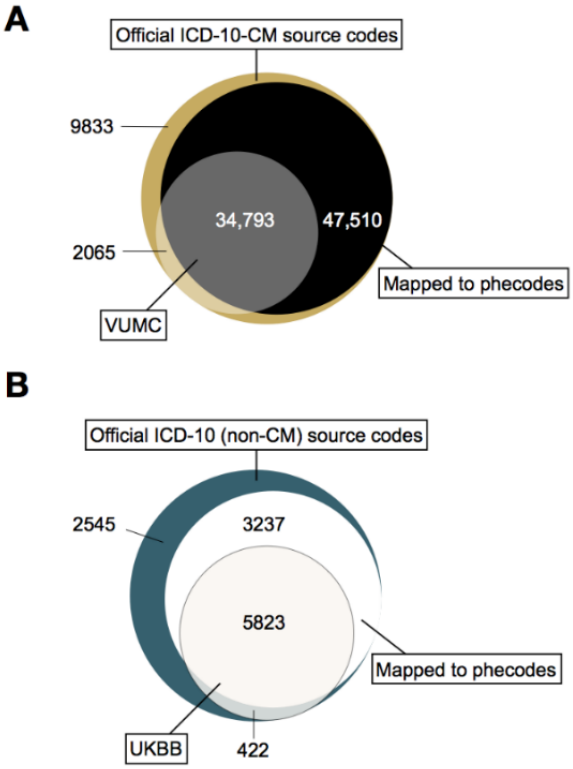
were translated to phecodes that were not in the same family. This can be seen in the mapping of ICD-10-CM D57.812 "Other sickle-cell disorders with splenic sequestration" to phecodes 282.5 "Sickle cell anemia" and 289.5 "Diseases of spleen". This latter association created a polyhierarchical nature to phecodes that did not previously exist.

To map ICD-10 (non-CM) codes to phecodes, we also used ICD-10 codes from the 2018AA UMLS [18]. ICD-10 codes were mapped to phecodes in a similar manner to ICD-10-CM, but since a GEMS to translate ICD-10 to ICD-9-CM was not available, we used only string matching and previously manually reviewed resources from the UMLS [24], NLM [23], and OHDSI [25,26].

Evaluation of Phecode Coverage of ICD-10 and ICD-10-CM in UKBB and VUMC

To evaluate the phecode coverage of ICD-10 and ICD-10-CM source codes in UKBB and VUMC, respectively, we calculated the number of source codes in the 2018AA UMLS, the number of source codes mapped to phecodes, and the number of mapped and unmapped source codes that were used in the two EHRs (Figure 2). To identify potential limitations of our automated mapping approach, two authors with clinical training (PW, WQW) manually reviewed all the unmapped ICD-10 and ICD-10-CM codes that were used at UKBB and VUMC, respectively.

Figure 2. Counts of distinct ICD-10-CM source codes at VUMC and ICD-10 (non-CM) source codes in UKBB. (A) Number of unique ICD-10-CM codes in each category. For example, there were 34,793 unique codes (grey section) that were in the official ICD-10-CM system, observed in the VUMC dataset, and mapped to phecodes. (B) Number of unique ICD-10 codes in each category. For example, there were 5823 unique codes (off-white section) that were in the official ICD-10 system, observed in the UKBB dataset, and mapped to phecodes. ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification; VUMC: Vanderbilt University Medical Center; ICD-10: International Classification of Diseases, 10th Revision; UKBB: UK Biobank.



Comparison of Phenotypes Generated from the ICD-10-CM Phecode Map

We aimed to provide evidence that the ICD-10-CM phecode map resulted in phenotypes like those sourced from the ICD-9-CM phecode map. First, we selected 357,728 patients in the VUMC EHR who had ≥1 ICD-9-CM and ≥1 ICD-10-CM codes in two 18-month windows. We selected windows to occur prior to and after VUMC's transition to ICD-10-CM. To reduce potential confounders, we left a 6-month buffer after ICD-9-CM was replaced with ICD-10-CM. Further, the ICD-10-CM observation window ended before VUMC switched from its locally developed EHR [27] to the Epic system. This created two windows ranging from January 1, 2014 to June 30, 2015 for ICD-9-CM, and January 1, 2016 to June 30, 2017 for ICD-10-CM (Figure 3). The final cohort consisted of 55.10% (197,109/357,728) females with mean age of 45 (SD 25) years old. From the two observation periods, we extracted all ICD-9-CM and ICD-10-CM codes for each patient. We then mapped these codes to phecodes using the ICD-9-CM phecode [14] and ICD-10-CM phecode maps.

We used the patient cohort to test our hypothesis that the ICD-10-CM phecode map created phenotype definitions that were comparable to those generated using the gold-standard ICD-9-CM phecode map. For this analysis, we used four common chronic diseases (Hypertension, Hyperlipidemia, Type 1 Diabetes, and Type 2 Diabetes) and chose one acute disease (Intestinal infection) as a negative control. We expected that a

XSL•FO
RenderX

large majority of the chronic disease patients and a small minority of the acute disease patients from the ICD-9-CM era would reproduce the same phenotypes during the ICD-10-CM era. We defined the phenotype cases as follows: Hypertension with phecodes 401.* (* means one or more digits or a period); Hyperlipidemia, phecodes 272.*; Type 1 diabetes, phecodes 250.1*; Type 2 diabetes, phecodes 250.2*; Intestinal infection, phecodes 008.*.

For each phenotype, we reported the number of ICD-9-CM cases and the number of those individuals who were also ICD-10-CM cases. To identify the possible reasons for individuals who were not identified as phenotype cases in the ICD-10-CM period, two authors with clinical training (PW, WQW) manually reviewed the EHRs of ten randomly selected patients from each chronic disease group, except Type 1 diabetes, for a total of thirty patients.

Figure 3. Timeline of the two 18-month periods from which ICD-9-CM and ICD-10-CM codes from VUMC were analyzed. The cohort of 357,728 patients had at least one ICD-9-CM and one ICD-10-CM code in the respective 18-month windows. ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification; ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification.

Comparative PheWAS Analysis of a Lipoprotein(a) Single-Nucleotide Polymorphism

To evaluate the accuracy of the ICD-10-CM phecode map, we performed two PheWASs on a Lipoprotein(a) (LPA) genetic variant (rs10455872) using mapped phecodes from ICD-9-CM and ICD-10-CM. The LPA single-nucleotide polymorphism (SNP) is associated with increased risks of developing hyperlipidemia and cardiovascular diseases [28-30].

We used data from BioVU, the deidentified DNA biobank at VUMC, to conduct the PheWAS [31]. We identified 13,900 adults (56.9% female; mean 59 [SD 15] years old in 2014), who had rs10455872 genotyped and at least one ICD-9-CM and ICD-10-CM code in their respective time windows. For rs10455872, we observed 86.7% AA, 12.8% AG, and 0.5% GG. We used 1632 phecodes that overlapped in the time windows for PheWAS using the R PheWAS package [13] with binary logistic regression, adjusting for age, sex, and race.

Results

Phecode Coverage of ICD-10-CM and ICD-10 in VUMC and UKBB

Of all possible ICD-10-CM codes [18], 82,303 (87.37%) mapped to at least one phecode, with 7881 (8.37%) mapping to >1 phecode. For example, ICD-10-CM I25.708 “Atherosclerosis of coronary artery bypass graft(s), unspecified, with other forms of angina pectoris” mapped to phecodes 411.3 “Angina pectoris” and 411.4 “Coronary atherosclerosis”. Of all possible ICD-10 codes, 9060 (75.33%) mapped to at least one phecode, and 289 (2.40%) mapped to >1 phecode. For example, ICD-10 code B21.1 “HIV disease resulting in Burkitt lymphoma” mapped to phecodes 071.1 “HIV infection, symptomatic” and 202.2 “Non-Hodgkins lymphoma”.

Among the 36,858 ICD-10-CM codes used at VUMC, 34,793 (94.40%) codes were mapped to phecodes. Of the 6245 ICD-10 codes used in the UKBB, 5823 (93.24%) codes mapped to phecodes (Table 1, Figure 2). Considering all the instances of ICD-10-CM and ICD-10 codes used at each site, we generated a total count of unique codes grouped by patient, date, and those codes that mapped to phecodes (Table 1). Among the total number of codes used, the vast majority of ICD-10-CM (17,658,470/19,682,697; 89.72%) and ICD-10 (4,279,544/5,114,363; 83.68%) codes were mapped to phecodes.

XSL•FO
RenderX

Table 1. ICD-10-CM and ICD-10 codes data summary.

	ICD-10-CM[a] (VUMC[b])	ICD-10[c] (UKBB[d])
Official classification systems		
Unique codes, n	94,201	12,027
Unique codes mapped, n (%)	82,303 (87.37)	9,060 (75.33)
Official codes used in cohorts		
Unique codes, n	36,858	6,245
Unique codes mapped, n (%)	34,793 (94.40)	5,823 (93.24)
Total patients (with ICD-10-CM or ICD-10 codes), n	651,649	391,181
Total instances of all ICD[e] codes, n	19,682,697	5,114,363
Instances mapped to phecodes, n (%)	17,658,470 (89.72)	4,279,544 (83.68)

[a]ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification

[b]VUMC: Vanderbilt University Medical Center

[c]ICD-10: International Classification of Diseases, 10th Revision

[d]UKBB: UK Biobank

[e]ICD: International Classification of Diseases

Analysis of Unmapped ICD-10 and ICD-10-CM Codes

Many of the unmapped ICD-10 codes used in the UKBB dataset represented medical concepts related to personal (ie, past medical history) or family history of disease. For ICD-10-CM, removing codes used at VUMC that we expected to be unmapped (ie, local or supplementary classification codes) left 2065 ICD-10-CM codes that did not map to a phecode. After excluding 1395 codes (eg, X, Y, and Z codes) indicating nonbiological disease phenotypes, 670 codes remained, the majority of which represented either external causes of morbidity or factors influencing health status and contact with health services. All the remaining unmapped ICD-10-CM codes in this cohort had <200 unique individuals (ie, <0.1% of the cohort), and the majority of the ICD-10-CM codes with >10 unique individuals were phenotypes that are most likely due to nongenetic factors. For example, 287 (59.2%) of the unmapped ICD-10-CM codes represented external causes of morbidity, such as assault and injuries due to motor vehicle accidents.

Reproducibility Analysis of the ICD-10-CM Phecode map

In the defined 18-month time windows, a cohort of 357,728 patients had both ICD-9-CM and ICD-10-CM codes (Figure 3). For the chronic diseases, 70-75% of individuals with the relevant phecodes in the ICD-9-CM observation period also had the same phecodes of interest during the ICD-10-CM period. On the contrary, for the reproducibility analysis with an acute disease we observed that <10% of individuals who had phecodes 008.* (Intestinal infection) in the ICD-9-CM period also had the same phecodes in the ICD-10-CM period (Table 2).

To identify the reasons that may explain why some patients were not identified as cases for the phenotype of interest during the ICD-10-CM period, we manually reviewed their medical records. A total of 30 patients were selected for review, 10 each from the Hypertension, Hyperlipidemia, and Type 2 diabetes cohorts (see Multimedia Appendix 1). We found that none of the patients had a relevant ICD-10-CM code for the phenotype being studied in the 18-month observation period. Reasons for patients not being ICD-10-CM cases included: patients were labeled with the relevant ICD-10-CM code(s) outside of the short ICD-10-CM observation window (8 patients), patients had <2 visits at VUMC during the ICD-10-CM period or were only seen by physician specialists (10 patients; eg, a patient with hypertension was only seen by their neurologist during the ICD-10-CM period), and patients were inconsistently diagnosed (2 people; eg, patient with Type 1 diabetes given Type 2 diabetes ICD-9-CM code). No cases were missed due to errors in the ICD-10-CM phecode map.

XSL•FO
RenderX

Table 2. ICD-10-CM phecode map reproducibility analysis.

Phenotype	Phecodes[a]	ICD-9-CM[b] cases (n)	ICD-10-CM[c] case|ICD-9-CM case[d], n (%)
Hypertension	401.*	65,216	49,468 (75.85)
Hyperlipidemia	272.*	51,187	36,187 (70.7)
Type 1 diabetes	250.1*	5782	4412 (76.31)
Type 2 diabetes	250.2*	25,077	19,066 (76.03)
Intestinal infection	008.*	3410	273 (8.01)

[a]In the phecode column, * means ≥1 digits or a period (eg, phecode 401.*=phecodes 401, 401.1, 401.3, 401.22, 401.21, or 401.2)

[b]ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification

[c]ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification

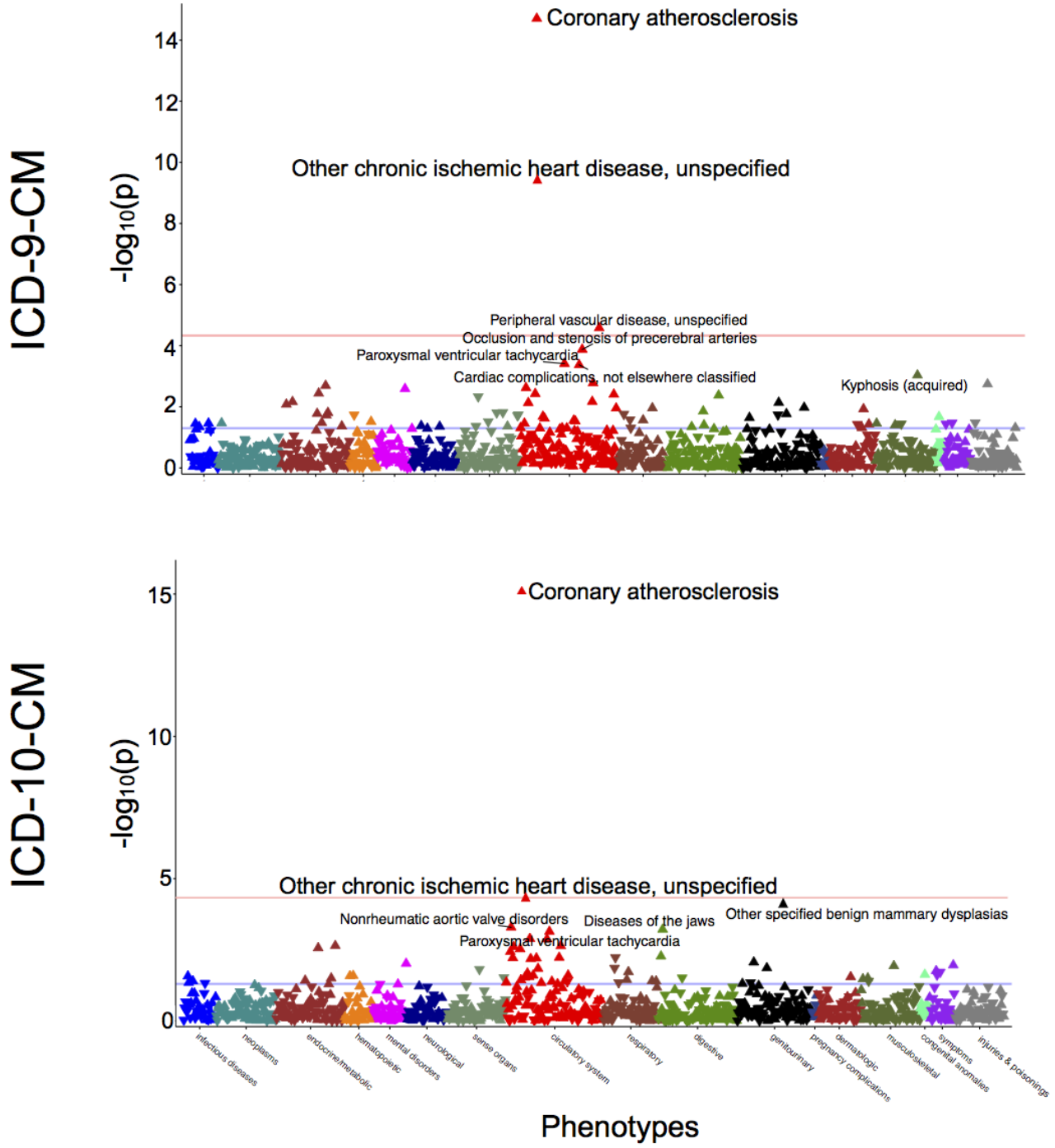
[d]In the last column, "ICD-10-CM case|ICD-9-CM case" indicates patients who were cases for the phenotype of interest during the ICD-9-CM period who were also ICD-10-CM cases

Comparative PheWAS Analysis of the Lipoprotein(a) SNP, rs10455872

To further evaluate the ICD-10-CM phecode map, we performed and compared the results of PheWAS analyses for rs10455872. One PheWAS was conducted using the ICD-9-CM map and another was conducted using the ICD-10-CM map. Both analyses replicated previous findings with similar effect sizes: coronary atherosclerosis (ICD-9-CM: $P<.001$; odds ratio [OR] 1.60 [95% CI 1.43-1.80] vs ICD-10-CM: $P<.001$, OR 1.60 [95% CI 1.43-1.80]) and chronic ischemic heart disease (ICD-9-CM: $P<.001$; OR 1.56, [95% CI 1.35-1.79] vs ICD-10-CM: $P<.001$, OR 1.47 [95% CI 1.22-1.77]) (Figure 4).

XSL•FO
RenderX

Figure 4. Comparative PheWAS of lipoprotein(a) genetic variant, rs10455872. "Coronary atherosclerosis" (phecode 411.4) and "Other chronic ischemic heart disease" (phecode 411.8) were top hits associated with rs10455872 in a PheWAS analysis conducted using ICD-9-CM (top) and ICD-10-CM (bottom) phecode maps. Analyses were adjusted for age, sex, and race. PheWAS: phenome-wide association studies; ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification; ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification.



Discussion

Maps of ICD-10 and ICD-10-CM Codes to Phecodes have High Coverage and Yield Similar Results

In this study, we described the process of mapping ICD-10 and ICD-10-CM codes to phecodes and evaluated the results of the new maps in two databases. These results show that the majority of the ICD-10 and ICD-10-CM codes used in EHRs were mapped to phecodes. Our analyses suggest that researchers can expect that phenotypes sourced using the ICD-10-CM phecode map will be like those sourced from the gold-standard ICD-9-CM phecode map. As the use of ICD-10 and ICD-10-CM codes increases, so does the need for convenient and reliable methods of aggregating codes to represent clinically meaningful phenotypes.

Since the introduction of phecodes, many studies have demonstrated the value of aggregating ICD-9-CM codes for genetic association studies. These maps will allow biomedical

XSL•FO
RenderX

researchers to leverage clinical data represented by ICD-10 and ICD-10-CM codes for their large-scale PheWAS using EHRs. They will also allow researchers to combine phenotypes as phecodes mapped from ICD-9- and ICD-10-based coding systems, thereby increasing the size of their patient cohorts and statistical power of their studies. The maps are available from the PheWAS Resources page [14] and are incorporated in the PheWAS R package, version 0.99.5-2 [13,32].

ICD-10 and ICD-10-CM Codes not Mapped to Phecodes

Analysis of the unmapped ICD-10 codes demonstrates a possible area of expansion for phecodes. The ICD-10 phecode map did not include medical concepts representing personal history or family history of disease.

We observed that a majority of the unmapped ICD-10-CM codes represented concepts that we did not expect to have phecode equivalents. Most of the codes were from ICD-10-CM chapters 20, “External causes of morbidity” and 21, “Factors influencing health status and contact with health services”. Codes from chapter 19, “Injury, poisoning, and certain other consequences of external causes” also made up a large proportion of unmapped codes, such as ICD-10-CM T38.3X6A, “Underdosing of insulin and oral hypoglycemic [antidiabetic] drugs, initial encounter”. We did not expect ICD-10-CM T38.3X6A to map to a phecode, as it is an encounter code that is not relevant to PheWAS. Three-digit codes that are not frequently used for reimbursement purposes, such as ICD-10-CM I67, “Other cerebrovascular diseases”, also made up many unmapped codes. A few potential clinically meaningful phenotypes, such as ICD-10-CM O04.6, “Delayed or excessive hemorrhage following [induced] termination of pregnancy”, were unmapped and represent areas of potential expansion for phecodes.

ICD-10-CM Phecode Map Phenotype Reproducibility Analysis

In general, our analysis suggests that in most of the cases in which phenotypes are not reproduced in the ICD-10-CM observation period, they are not due to errors in the ICD-10-CM phecode map.. This study’s reproducibility analysis (Table 2) demonstrates that most patients (70-75%) with phecodes of four chronic diseases sourced from ICD-9-CM codes were also phenotype cases in the ICD-10-CM era. In comparison, when the same experiment is repeated for an acute disease (Intestinal infection), a minority (<10%) of patients had the same phenotype in the ICD-10-CM period.

Using the ICD-9-CM and ICD-10-CM maps, PheWAS found significant genetic associations with similar effect sizes for coronary atherosclerosis and chronic ischemic heart disease (Figure 4). Results of this analysis provide additional support for the accuracy of the ICD-10-CM map when compared to the gold-standard ICD-9-CM phecode map.

PheWAS Using ICD-10 Phecode Map

Two published studies have used the ICD-10 phecode map to identify genotype-phenotype associations using UKBB data. Zhou et al used the map to demonstrate a method that adjusts for case-control imbalances in a large genome-wide PheWAS [33], and Li et al used the same map to estimate the causal effects of elevated serum uric acid across the phenome [12].

Utilization of Phecodes Outside of PheWAS

In addition to being employed for PheWAS, phecodes have been used to answer a range of questions in biomedicine. Phecodes have been used to identify features in radiographic images that are associated with disease phenotypes [34] and used in machine learning models to improve cardiovascular disease prediction [35]. In a recent study to understand public opinion about diseases, Huang et al identified articles about diseases and mapped them to phecodes [36]. Motivated by the difficulties in automatically translating diagnosis codes from EHRs, Shi et al used phecodes to map ICD-9-CM diagnosis codes from one health system to another [37]. Phecodes have also been applied to identify conditions for aggregation in phenotype risk scores, much as SNPs are aggregated as a genetic risk score to identify Mendelian diseases and determine pathogenicity of genetic variants [38].

Related Work

The Clinical Classification Software (CCS) is another maintained system for aggregating ICD codes into clinically meaningful phenotypes. CCS was originally developed by the Agency for Healthcare Research and Quality (AHRQ) to cluster ICD-9-CM diagnosis and procedure codes to a smaller number of clinically meaningful categories [39]. CCS has been used for many purposes, such as measuring outcomes [40] and predicting future health care usage [41]. In a previous study, we showed that phecodes were better aligned with diseases mentioned in clinical practice and that were relevant to genomic studies than CCS for ICD-9-CM (CCS9) codes [20]. We found that phecodes outperform CCS9 codes, in part because CCS9 was not as granular as phecodes. Since CCS for ICD-10-CM (CCS10) is of similar granularity as CCS9 (283 versus 285 disease groups) [42], we believe that the phecode map would likely still better represent clinically meaningful phenotypes in genetic research.

Limitations

This study has limitations. First, only 84.14% (1570/1866) of phecodes are mapped to at least one ICD-10 code. This may be due in part to the automated strategy that we used to map ICD-10 to ICD-9-CM. Second, the VUMC data are from a single site, thereby making it difficult to generalize the results of our accuracy studies (eg, phenotype reproducibility analysis and LPA SNP PheWAS) to patient cohorts in other EHRs. Third, we have not yet manually reviewed all the mappings in these beta phecode maps, and our assumptions that the manually reviewed resources (eg, NLM and OHDSI) are highly accurate could have affected the accuracy of the new phecode maps. For example, in the 2009 ICD-10-CM to ICD-9-CM GEMS, >90% of the mappings were approximate (ie, nonequivalent) [15]. For this study’s purposes, we aimed to maximize phecode coverage of ICD source codes and thus included both equivalent and nonequivalent 2018 GEMS translations, which could have decreased mapping performance.

Fourth, our automated approach to map >80,000 ICD-10-CM and >9000 ICD-10 codes to phecodes with minimal human engineering could have decreased the accuracy of the final maps.

XSL•FO
RenderX

Hripcsak et al [43] recently evaluated the effects of translating ICD-9-CM codes to SNOMED CT codes on the creation of patient cohorts. In general, they found that mapping source billing codes to a standard clinical vocabulary (eg, ICD-9-CM to SNOMED CT) did not greatly affect cohort selection. Their findings suggested that optimized domain knowledge–engineered mappings outperformed simple automated translations between clinical vocabularies. Using four phenotype concept sets, they showed that automated mappings resulted in errors of up to 10% and that domain-knowledge engineered mappings had errors of <0.5%. Other studies have also found that mapping performance is generally better with smaller value sets [17]. To create a more comprehensive and accurate map between ICD-9-CM and ICD-10-CM, future mapping studies could consider using an iterative forward and backward mapping approach using GEMS [17].

Future Directions

Currently, if an ICD-10 or ICD-10-CM code maps to ≥2 unlinked phecodes, we keep all the mappings. In subsequent studies, it will be important to further scrutinize these mappings to ensure accuracy through manual review. As new ICD-10-CM codes are released, we plan to assess their relevance to clinical practice and genetic research and decide whether we should translate them to phecodes. We intend to address the unmapped source codes (eg, ICD-10-CM E78.41 "Elevated Lipoprotein(a)") by potentially expanding the phecode system, and to systematically evaluate the mappings with input from users.

Conclusions

In this paper, we introduced our work on mapping ICD-10 and ICD-10-CM codes to phecodes. We provide initial beta maps with high coverage of EHR data in two large databases. Results from this study suggested that the ICD-10-CM phecode map created phenotypes similar to those generated by the ICD-9-CM phecode map. These mappings will enable researchers to leverage accumulated ICD-10 and ICD-10-CM data in the EHR for large PheWAS.

Acknowledgments

The project was supported by NIH grant R01 LM 010685, R01 HL133786, T32 GM007347, T15 LM007450, P50 GM115305, and AHA Scientist Development Grant 16SDG27490014. The dataset used in the analyses described were obtained from Vanderbilt University Medical Center's BioVU, which is supported by institutional funding and by the Vanderbilt CTSA grant ULTR000445 from NCATS/NIH. This research was also conducted using the UK Biobank Resource under Application Number 10775. The work conducted in Edinburgh was supported by funding for the infrastructure and staffing of the Edinburgh CRUK Cancer Research Centre. ET is supported by a CRUK Career Development Fellowship (C31250/ A22804). XM and XL are supported by the China Scholarship Council Studentships. We thank those individuals who manually reviewed the various maps that we used in this study [18,22-25]. We also thank the peer-reviewers who provided feedback for this manuscript.

Authors' Contributions

PW, AG, JCD, and WQW contributed to the design of the studies. PW, AG, XM, XL, HC, ET, TV, JZ, JCD, and WQW analyzed the data. PW and AG were responsible for the literature review. AG, XM, XL, and ET retrieved the raw data. PW, AG, RC, LB, JCD, ET, and WQW interpreted the data. PW, AG, JCD, and WQW drafted the initial manuscript. PW, AG, JCD, and WQW were involved in the creation and design of figures and tables. All authors revised the document and gave final approval for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
ICD-10-CM reproducibility analysis, manual chart review results.
[XLSX File (Microsoft Excel File), 10 KB - medinform_v7i4e14325_app1.xlsx]

References

1. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc 2016 Nov;23(6):1046-1052. [doi: 10.1093/jamia/ocv202] [Medline: 27026615]
2. Gamazon ER, Segrè AV, van de Bunt M, Wen X, Xi HS, Hormozdiari F, GTEx Consortium, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat Genet 2018 Jul;50(7):956-967 [FREE Full text] [doi: 10.1038/s41588-018-0154-4] [Medline: 29955180]
3. Nielsen JB, Thorolfsdottir RB, Fritsche LG, Zhou W, Skov MW, Graham SE, et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. Nat Genet 2018 Sep;50(9):1234-1239 [FREE Full text] [doi: 10.1038/s41588-018-0171-3] [Medline: 30061737]

XSL•FO
RenderX

4. Simonti CN, Vernot B, Bastarache L, Bottinger E, Carrell DS, Chisholm RL, et al. The phenotypic legacy of admixture between modern humans and Neandertals. Science 2016 Feb 12;351(6274):737-741 [FREE Full text] [doi: 10.1126/science.aad2149] [Medline: 26912863]
5. Diogo D, Bastarache L, Liao KP, Graham RR, Fulton RS, Greenberg JD, et al. TYK2 protein-coding variants protect against rheumatoid arthritis and autoimmunity, with no evidence of major pleiotropic effects on non-autoimmune complex traits. PLoS One 2015;10(4):e0122271 [FREE Full text] [doi: 10.1371/journal.pone.0122271] [Medline: 25849893]
6. Rastegar-Mojarad M, Ye Z, Kolesar JM, Hebbring SJ, Lin SM. Opportunities for drug repositioning from phenome-wide association studies. Nat Biotechnol 2015 Apr;33(4):342-345. [doi: 10.1038/nbt.3183] [Medline: 25850054]
7. Millard LAC, Davies NM, Timpson NJ, Tilling K, Flach PA, Davey Smith G. MR-PheWAS: hypothesis prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization. Sci Rep 2015 Nov 16;5:16645 [FREE Full text] [doi: 10.1038/srep16645] [Medline: 26568383]
8. Ehm MG, Aponte JL, Chiano MN, Yerges-Armstrong LM, Johnson T, Barker JN, et al. Phenome-wide association study using research participants' self-reported data provides insight into the Th17 and IL-17 pathway. PLoS One 2017;12(11):e0186405 [FREE Full text] [doi: 10.1371/journal.pone.0186405] [Medline: 29091937]
9. Liu J, Ye Z, Mayer JG, Hoch BA, Green C, Rolak L, et al. Phenome-wide association study maps new diseases to the human major histocompatibility complex region. J Med Genet 2016 Oct;53(10):681-689 [FREE Full text] [doi: 10.1136/jmedgenet-2016-103867] [Medline: 27287392]
10. Neuraz A, Chouchana L, Malamut G, Le Beller C, Roche D, Beaune P, et al. Phenome-wide association studies on a quantitative trait: application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics. PLoS Comput Biol 2013;9(12):e1003405 [FREE Full text] [doi: 10.1371/journal.pcbi.1003405] [Medline: 24385893]
11. Doshi-Velez F, Ge Y, Kohane I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. Pediatrics 2014 Jan;133(1):e54-e63 [FREE Full text] [doi: 10.1542/peds.2013-0819] [Medline: 24323995]
12. Li X, Meng X, Spiliopoulou A, Timofeeva M, Wei W, Gifford A, et al. MR-PheWAS: exploring the causal effect of SUA level on multiple disease outcomes by using genetic instruments in UK Biobank. Ann Rheum Dis 2018 Jul;77(7):1039-1047 [FREE Full text] [doi: 10.1136/annrheumdis-2017-212534] [Medline: 29437585]
13. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics 2014 Aug 15;30(16):2375-2376 [FREE Full text] [doi: 10.1093/bioinformatics/btu197] [Medline: 24733291]
14. PheWAS Catalog. Phecode Map 1.2 with ICD-9 Codes. URL: https://phewascatalog.org/phecodes [accessed 2019-07-14]
15. Steindel SJ. International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. J Am Med Inform Assoc 2010;17(3):274-282 [FREE Full text] [doi: 10.1136/jamia.2009.001230] [Medline: 20442144]
16. Topaz M, Shafran-Topaz L, Bowles KH. ICD-9 to ICD-10: evolution, revolution, and current debates in the United States. Perspect Health Inf Manag 2013;10:1d [FREE Full text] [Medline: 23805064]
17. Fung KW, Richesson R, Smerek M, Pereira KC, Green BB, Patkar A, et al. Preparing for the ICD-10-CM Transition: Automated Methods for Translating ICD Codes in Clinical Phenotype Definitions. EGEMS (Wash DC) 2016;4(1):1211 [FREE Full text] [doi: 10.13063/2327-9214.1211] [Medline: 27195309]
18. Wilder V. NLM Technical Bulletin.: US National Library of Medicine; 2018 May. UMLS 2018AA Release Available. URL: https://www.nlm.nih.gov/pubs/techbull/mj18/mj18_umls_2018aa_release.html [accessed 2019-07-16]
19. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol 2013 Dec;31(12):1102-1110 [FREE Full text] [doi: 10.1038/nbt.2749] [Medline: 24270849]
20. Wei W, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, et al. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLoS One 2017;12(7):e0175508 [FREE Full text] [doi: 10.1371/journal.pone.0175508] [Medline: 28686612]
21. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015 Mar;12(3):e1001779 [FREE Full text] [doi: 10.1371/journal.pmed.1001779] [Medline: 25826379]
22. Centers for Medicare & Medicaid Services. CMS.gov. 2017 Aug 11. 2018 ICD-10-CM and GEMs. URL: https://www.cms.gov/Medicare/Coding/ICD10/2018-ICD-10-CM-and-GEMs.html [accessed 2019-07-05]
23. US National Library of Medicine. 2018 Jan 26. SNOMED CT to ICD-9-CM Rule Based Mapping to Support Reimbursement. URL: https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd9cm_reimburse.html [accessed 2019-04-06]
24. Fung KW, Bodenreider O. Utilizing the UMLS for semantic mapping between terminologies. AMIA Annu Symp Proc 2005:266-270 [FREE Full text] [Medline: 16779043]
25. Observational Health Data Sciences and Informatics. 2016 Jun 04. ICD9CM. URL: https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:icd9cm [accessed 2019-07-17]
26. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]

XSL•FO
RenderX

27. Giuse DA. Supporting communication in an integrated patient record system. AMIA Annu Symp Proc 2003:1065 [FREE Full text] [Medline: 14728568]
28. Nordestgaard BG, Chapman MJ, Ray K, Borén J, Andreotti F, Watts GF, European Atherosclerosis Society Consensus Panel. Lipoprotein(a) as a cardiovascular risk factor: current status. Eur Heart J 2010 Dec;31(23):2844-2853 [FREE Full text] [doi: 10.1093/eurheartj/ehq386] [Medline: 20965889]
29. Zhao J, Feng Q, Wu P, Warner JL, Denny JC, Wei W. Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of Lipoprotein(a) (LPA). PLoS One 2019;14(2):e0212112 [FREE Full text] [doi: 10.1371/journal.pone.0212112] [Medline: 30759150]
30. Wei W, Li X, Feng Q, Kubo M, Kullo IJ, Peissig PL, et al. LPA Variants Are Associated With Residual Cardiovascular Risk in Patients Receiving Statins. Circulation 2018 Oct 23;138(17):1839-1849. [doi: 10.1161/CIRCULATIONAHA.117.031356] [Medline: 29703846]
31. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 2008 Sep;84(3):362-369 [FREE Full text] [doi: 10.1038/clpt.2008.89] [Medline: 18500243]
32. PheWAS. GitHub. PheWAS R Package. URL: https://github.com/PheWAS/PheWAS [accessed 2019-10-04]
33. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet 2018 Sep;50(9):1335-1341 [FREE Full text] [doi: 10.1038/s41588-018-0184-y] [Medline: 30104761]
34. Chaganti S, Mawn LA, Kang H, Egan J, Resnick SM, Beason-Held LL, et al. Electronic Medical Record Context Signatures Improve Diagnostic Classification Using Medical Image Computing. IEEE J Biomed Health Inform 2019 Sep;23(5):2052-2062. [doi: 10.1109/JBHI.2018.2890084] [Medline: 30602428]
35. Zhao J, Feng Q, Wu P, Lupu RA, Wilke RA, Wells QS, et al. Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. Sci Rep 2019 Jan 24;9(1):717. [doi: 10.1038/s41598-018-36745-x] [Medline: 30679510]
36. Huang M, ElTayeby O, Zolnoori M, Yao L. Public Opinions Toward Diseases: Infodemiological Study on News Media Data. J Med Internet Res 2018 May 08;20(5):e10047 [FREE Full text] [doi: 10.2196/10047] [Medline: 29739741]
37. Shi X, Li X, Cai T. Spherical Regression under Mismatch Corruption with Application to Automated Knowledge Translation. arXiv preprint 2019 Sep 04 [FREE Full text]
38. Bastarache L, Hughey JJ, Hebbring S, Marlo J, Zhao W, Ho WT, et al. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science 2018 Mar 16;359(6381):1233-1239 [FREE Full text] [doi: 10.1126/science.aal4043] [Medline: 29590070]
39. Agency for Healthcare Research and Quality. Healthcare Cost and Utilization Project (HCUP). 2012 Jan. HCUP CCS Fact Sheet. URL: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccsfactsheet.jsp [accessed 2019-07-15]
40. Sabbatini AK, Kocher KE, Basu A, Hsia RY. In-Hospital Outcomes and Costs Among Patients Hospitalized During a Return Visit to the Emergency Department. JAMA 2016 Feb 16;315(7):663-671. [doi: 10.1001/jama.2016.0649] [Medline: 26881369]
41. Hu Z, Hao S, Jin B, Shin AY, Zhu C, Huang M, et al. Online Prediction of Health Care Utilization in the Next Six Months Based on Electronic Health Record Information: A Cohort and Validation Study. J Med Internet Res 2015 Sep 22;17(9):e219 [FREE Full text] [doi: 10.2196/jmir.4976] [Medline: 26395541]
42. Agency for Healthcare Research and Quality. Healthcare Cost and Utilization Project (HCUP). 2019 Sep. Clinical Classifications Software (CCS) for ICD-10-PCS (beta version). URL: https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp [accessed 2019-07-05]
43. Hripcsak G, Levine ME, Shang N, Ryan PB. Effect of vocabulary mapping for conditions on phenotype cohorts. J Am Med Inform Assoc 2018 Dec 01;25(12):1618-1625 [FREE Full text] [doi: 10.1093/jamia/ocy124] [Medline: 30395248]

Abbreviations

AHRQ: Agency for Healthcare Research and Quality
CCS: Clinical Classification Software
CUI: Concept Unique Identifier
EHR: electronic health record
GEMS: General Equivalence Mappings
ICD: International Classification of Diseases
ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification
ICD-10: International Classification of Diseases, 10th Revision
ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification
LPA: lipoprotein(a)
NCHS: National Center for Health Statistics
NLM: National Library of Medicine

XSL•FO
RenderX

OHDSI: Observational Health Data Sciences and Informatics
OR: odds ratio
PheWAS: phenome-wide association studies
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
SNP: single nucleotide polymorphism
UKBB: UK Biobank
UMLS: Unified Medical Language System
VUMC: Vanderbilt University Medical Center
WHO: World Health Organization

Edited by G Eysenbach; submitted 09.04.19; peer-reviewed by I Perros, V Curcin; comments to author 02.07.19; revised version received 03.08.19; accepted 24.09.19; published 29.11.19.

Please cite as:
Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, Zhao J, Carroll R, Bastarache L, Denny JC, Theodoratou E, Wei WQ
Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation
JMIR Med Inform 2019;7(4):e14325
URL: http://medinform.jmir.org/2019/4/e14325/
doi:10.2196/14325
PMID:31553307

©Patrick Wu, Aliya Gifford, Xiangrui Meng, Xue Li, Harry Campbell, Tim Varley, Juan Zhao, Robert Carroll, Lisa Bastarache, Joshua C Denny, Evropi Theodoratou, Wei-Qi Wei. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 29.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

How Online Reviews and Services Affect Physician Outpatient Visits: Content Analysis of Evidence From Two Online Health Care Communities

Wei Lu[1], MA; Hong Wu[1], PhD

School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Corresponding Author:
Hong Wu, PhD
School of Medicine and Health Management
Tongji Medical College
Huazhong University of Science and Technology
13 Hangkong Road
Qiaokou District, Hubei Province
Wuhan, 430060
China
Phone: 86 132 7794 2186
Fax: 86 027 83692727
Email: wuhong634214924@163.com

Abstract

Background: Online healthcare communities are changing the ways of physician-patient communication and how patients choose outpatient care physicians. Although a majority of empirical work has examined the role of online reviews in consumer decisions, less research has been done in health care, and endogeneity of online reviews has not been fully considered. Moreover, the important factor of physician online services has been neglected in patient decisions.

Objective: In this paper, we addressed the endogeneity of online reviews and examined the impact of online reviews and services on outpatient visits based on theories of reviews and channel effects.

Methods: We used a difference-in-difference approach to account for physician- and website-specific effects by collecting information from 474 physician homepages on two online health care communities.

Results: We found that the number of reviews was more effective in influencing patient decisions compared with the overall review rating. An improvement in reviews leads to a relative increase in physician outpatient visits on that website. There are channel effects in health care: online services complement offline services (outpatient care appointments). Results further indicate that online services moderate the relationship between online reviews and physician outpatient visits.

Conclusions: This study investigated the effect of reviews and channel effects in health care by conducting a difference-in-difference analysis on two online health care communities. Our findings provide basic research on online health care communities.

(JMIR Med Inform 2019;7(4):e16185) doi:10.2196/16185

KEYWORDS

online health care communities; online reviews; online services; outpatient care; channel effect; patient choice

Introduction

Background

Patients often face uncertainty regarding the quality of physician services such as medical quality and bedside manner and often lack channels to access that information [1]. Information disclosure of medical quality is mainly based on the hospital/nursing home/organization level. However, patients are increasingly concerned about health care quality at the physician level. Information asymmetries between patients and physicians are extensive. Traditionally, patients relied on social networks to learn this information, such as peer recommendations. With the growing popularity of Web 2.0 technologies, online health care communities provide a useful channel for people to get physician information and have

XSL•FO
RenderX

become an integral part of their daily lives. In 2013, the number of adults who used the internet to search online for health care information was 59% in the United States [2]. More than 80% of patients search for health information before going to the doctor in China [3].

Online health care communities provide review forums, in which patients can share their disease information and treatment experience with other members of the community. In the absence of other channels to acquire information on physician medical quality, online review forums provide a potential opportunity for patients. Compared with traditional channels (eg, acquaintance recommendations), however, there has not been enough research into whether patients trust and refer to this information received online from strangers. Much effort has been dedicated to researching the health care quality of organizations such as hospitals and nursing homes [4,5], but less has been done at the individual physician level. Moreover, although quite a few studies have investigated the relationship between reviews and performance and generally get consistent results that higher reviews correlate with improved performance in other fields [6-8], the endogeneity of online reviews that may cause bias has not been fully considered in previous studies.

Today's organizations are continually adding new marketing channels through the internet to better serve their products and/or service receivers [9], and this phenomenon is also manifested in the health care industry. Online health care communities enable physicians to better help and serve patients by providing physicians with a variety of functions—for example, question and answer (written consultation) and telephone consultation services. With channel diversification, researchers try to find channel effects and channel choice [10,11]. Some researchers assert that the internet competes with traditional channels by decreasing transaction costs, such as search and monitoring [12,13]. For example, service receivers could find service providers in distant geographic markets who have lower prices, provide better service, offer higher quality products, or have products that better match their needs [12,13]. However, other researchers emphasize the importance of synergies between online and offline channels [14,15], demonstrating that the use of multiple channels tends to be more successful. Online channels have spillover effects, generating increased purchases in offline channels [16]. However, there are only a few studies that empirically explore the channel effect, especially in the health field [11].

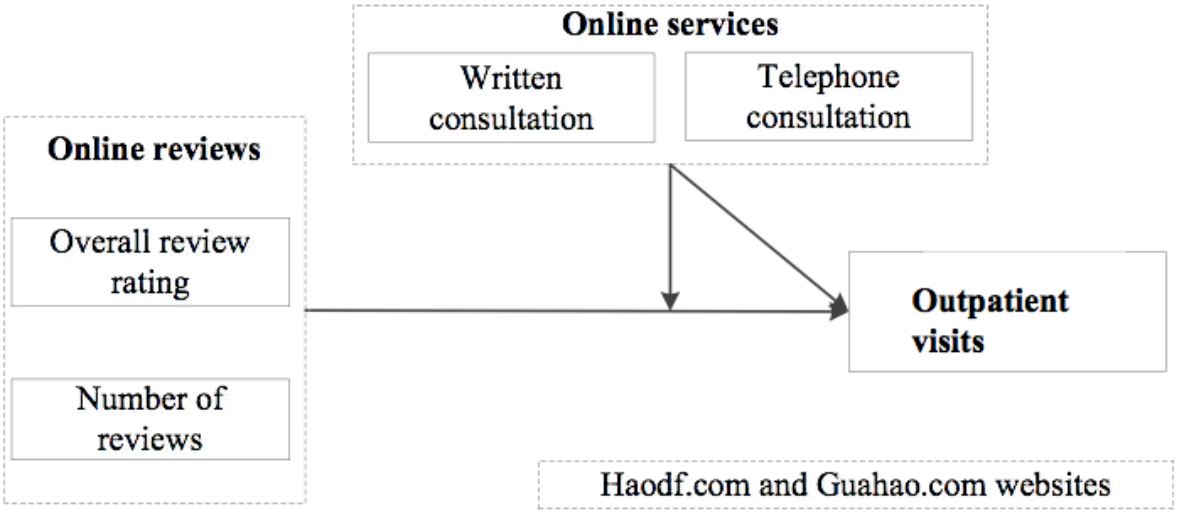
Both online reviews and online services provide information sources for patients. Online reviews help patients get information about the treatment experience from others, and patients can use online services to get their own experience by communicating with physicians directly. When physicians choose to provide online services, this may decrease patient dependence on online reviews. With the development of online channels, efforts to examine whether there are moderating effects of online services on the relationship between reviews and offline service (ie, outpatient visits) becomes necessary. The specific research questions addressed in this paper are as follows:

- RQ1: How do online reviews impact physician outpatient visits?
- RQ2: How do physician services provided via online channels impact their outpatient visits?
- RQ3: How does the relationship between online reviews and outpatient visits change relative to physician online services?

To answer these questions and solve endogeneity issues in the empirical estimation, we used a dataset of 474 physicians from two leading online health care communities, Haodf [17] and Guahao [18], to construct measures of each physician's outpatient visits and used a difference-in-difference approach to account for physician- and website-specific effects. Both Haodf and Guahao allow patients to post reviews on their platforms. In addition to outpatient visits, physicians on Haodf can provide online services (written and telephone consultations) for patients. Guahao only provides outpatient appointments. We use the overall review rating and number of reviews to measure the quality of reviews, both of which have been used in prior studies and are considered to be useful [1,19]. A difference-in-difference approach similar to that used in Chevalier and Mayzlin [20] is used in our paper: we measured reviews and number of outpatient visits for each physician who works at both Haodf and Guahao over three time points, and we examined whether a change in overall review rating and number of reviews over time for a physician on one website relative to the other predicts a change in subsequent outpatient visits of that physician on one website relative to the other. By using this approach, we were able to control for possible effects of unobserved physician characteristics on both reviews and outpatient visits. Moreover, by focusing on the differences across websites over time, we controlled for the unobserved website fixed effects at the two websites that may have affected both reviews and outpatient visits, such as website design, different patient populations, and patient preference. Figure 1 shows the conceptual model of this study. The hypotheses, presented below, were established according to the relationships expressed in the model in Figure 1.

XSL•FO
RenderX

Figure 1. Conceptual model.



Online Health Care Communities

With the growing popularity of Web 2.0 technologies such as blogs, tweets, podcasts, and wikis, many health care organizations and professionals are embracing social media. The use of social network software, with its ability to enrich the connection between patients and the rest of the medical industry, has been dubbed Health 2.0 [21], and the number of organizations adopting Health 2.0 is growing. Many online communities have been developed by patient organizations, providers, and nonprofit organizations in recent years, making it easier for patients to find health information [22,23]. Such online communities are virtual forums for patients to discuss their health concerns, share information about treatments and support, and communicate with physicians, an example of which is PatientsLikeMe [24].

Researchers have started to investigate the benefits of online health care communities for physicians [25,26] and patients [27]. Xiao et al [28] examined the factors that influence patients' online health care information searches and found that perceived health status could affect patients' online health care search frequency and diversity. Privacy concerns, trust, and information sensitivity are factors that have an impact on people's decisions about whether to place their health information online [29].

As a result of the limitations of existing health services, online communities in China have emerged in recent years. China has the world's largest population and thus represents a huge resource-consumption country. China's large population generates a variety of unique needs relating to medical services, therefore exhibiting unique behaviors within online health care communities. Health intimately concerns everyone, and with the emergence of online health care communities, patients have more channels to get physician information and physicians have more choices in the ways of helping patients. Based on the existing literature, we have found few studies exploring the causal effects of online reviews and services on outpatient visits. Our study aims to fill these gaps.

Endogeneity of Online Reviews

Several factors cause the endogeneity of online reviews. First, whether to post reviews is self-selected. Existing studies suggest that consumers are motivated to engage in posting reviews for different reasons (altruistic, product involvement, self-enhancement purposes, anxiety-reducing, vengeful, and advice-seeking, etc) [30,31]. This kind of volitional activity is likely subject to a variety of biases and social influences [32,33] that will cause estimation deviation if not considered. Second, product and service quality can be an underlying factor that drives both reviews and sales. Research has shown that consumers who are particularly satisfied or dissatisfied with the product or service quality will post feedback to let other knows [34]. Product and service quality is often hard to quantify or observe, especially in health care. Researchers face difficulties deciding whether high reviews or high product/service quality impacts high visits [35]. In this paper, we attempt to resolve the above endogeneity problem using a sophisticated econometric method.

Online Reviews and Outpatient Visits

Numerous empirical studies suggest reputation is one of the predominant factors in influencing seller performance [36,37] and consistently reveals that there is a close relationship between reviews and future visits. Online reviews can improve the interaction between consumers and sellers and decrease consumer risk, thereby increasing trust and cooperation on both sides [38,39]. Positive reviews can also positively impact product demands [7,20]. Reviews are increasingly believed to influence consumer behavior [40] and be more effective than traditional advertising [41].

Online health care communities are changing how patients choose physicians. The digitization of health care reviews makes it easy for patients to find physician treatment information, assists them in thoroughly evaluating physicians before making a choice, increases their trust in the physician, and decreases perceived risk [42,43]. However, how reviews impact patient

XSL•FO
RenderX

choice in health care has rarely been researched. In this paper, we investigated the role of online reviews in influencing patient decisions and thoroughly considered the endogeneity of online reviews. Previous studies have used different measurements about reviews, including the overall review rating [44-46] and the volume of online posting [47]. We hypothesized that both the overall review rating and number of reviews positively impact outpatient visits.

- H1a: An improvement in the overall review rating leads to an increase in outpatient visits on that website.
- H1b: An improvement in the number of reviews leads to an increase in outpatient visits on that website.

Channel Effects and Moderating Effects

With the emergence of online communities, more and more physicians adopt multiple channels to serve patients. Existing studies from other fields find both complementary and substitute effects between online and offline channels. From the complementary perspective, a number of studies suggest that the internet has a distinct influence on offline sales [16,48]. Many product and service receivers still rely on offline stores for the actual product or service purchase. Because the internet gains increasing importance for information collection [48], online channels may have spillover effects, generating increased purchases in offline channels [16]. These studies emphasize the theoretical advantages of integrating online services with existing physical channels. For example, a combinations of channels can be used to target different kinds of service receivers and offer different kinds of services cost effectively [49]. From a substitution perspective, researchers suggest that there may be substitution by advertisers between print, television, and radio advertising channels [50,51].

By analyzing the existing studies, we believe that studies that find substitute effects often focus on these highly standardized products. Using the example of a cup, a seller can sell it online or in the store, and the buyer gets the same thing regardless of the channel chosen. Some product categories compete because they can serve a similar defining purpose and thus may have similar potential customers [52,53]. However, for the health care industry, diagnoses often cannot be given to patients using online services; only suggestions can be given. Online channels cannot provide services that are identical to offline channels. If patients choose to get advice online, they have to accept the risks associated with the fact that the doctor cannot communicate with them face-to-face or look directly at the patient, listen to verbal cues, examine the patient physically, or even use the four diagnostic methods of traditional Chinese medicine. We hypothesized, however, that there is a complementary effect between online services and outpatient visits in health care.

Online health care communities can help patients access information about and contact physicians [54]. Through written and telephone consultation services, patients can engage with physicians before going to hospitals. Online communication helps patients to get to know the physician, thus reducing their uncertainty and sense of risk and enhancing their trust in the physician and increasing outpatient visits. Based on these insights, we hypothesized that the more online services that patients use, the higher the use of outpatient visits.

- H2a: A physician who provides written consultation services has higher totals of outpatient visits.
- H2b: A physician who provides telephone consultation services has higher totals of outpatient visits.

A physician providing online services can give patients more opportunity to evaluate the, which can enhance patient trust. Online service content is public to all users of online health communities, so these public communications give patients some insight into the physician's ability, including medical quality and bedside manner. Reviews are from patients who have finished an outpatient visit and can provide information to potential patients. If a physician provides online services, online service content offers a source of information for patients so they may be less dependent on reviews. If a patient communicates with a physician using online services before making an appointment for outpatient care, communication in advance can also decrease the uncertainty between physicians and patients.

During this channel extension process, consumer experiences with a seller in one channel may affect their perceptions and beliefs about the same seller in another channel [55]. The use of online services can reflect physician popularity and decrease the perceived risk of offline service, a similar effect of reviews, which are also described as a quality signal and can reduce perceived risk. Based on these considerations, we hypothesized that online services mitigate the relationship between reviews and outpatient visits.

- H3a: Increasing numbers of online written consultations by a physician mitigates the main effect between reviews and outpatient visits.
- H3b: Increasing numbers of online telephone consultations by a physician mitigates the main effect between reviews and outpatient visits.

Methods

Research Contexts

Our research contexts are Haodf and Guahao, two very popular and professionally regarded online health care communities in China that have established cooperative relationships with big companies such as Tencent, Sina, and Sohu.

Haodf was founded in 2006 and has become the most influential medical information and physician-patient interaction platform in China. On this platform, physicians can choose to offer online written consultations, telephone consultations, outpatient visits, or all of the above. Patients can search for generalized health information and/or ask physicians questions. Many unique attributes and services are available on Haodf to help patients make better and more accurate selections that suit their needs. Patients visit in increasing numbers and use this website to get help from physicians online. Haodf began to provide video consultation services in late 2016; however, only a very few patients use these services, so we did not consider them in our paper. Figure 2 shows a physician page on the Haodf website.

XSL•FO
RenderX

Figure 2. Haodf website.

Guahao was founded in 2010 and has become the leading online health care community for outpatient appointments specifically. Guahao was authorized by the China Health and Family Planning Committee in March 2010. With the help of Guahao, patients can make appointments easily, save valuable time, and increase efficiency. It has helped more than one hundred million people. Guahao began to provide online written consultations and video consultation services in September 2016. However, compared with outpatient appointments, the proportion of written and video consultation use is small. Our data were collected in 2014 when only outpatient care appointment service was provided on Guahao. Figure 3 shows a physician page on the Guahao website.

Figure 3. Guahao website.

XSL•FO
RenderX

While Haodf is designed to help patients find suitable physicians to provide written and telephone consultations and outpatient visits, Guahao aims to provide people with the most efficient and best medical treatment and only provides outpatient appointments. These two communities automatically create homepages for physicians and their hospitals based on a directory collected. Physicians can choose to manage their homepages and work on them. Both websites have a formal and comprehensive reputation mechanism, which is important for this study. Patients can post their treatment experiences after receiving outpatient services, which helps potential patients make better choices.

Sample and Data Collection

The homepage contains details of the physician, including their title (eg, chief physician, associate chief physician, attending physician), hospital that the physician belongs to, and the hospital's level (eg, level A, B and C; level A offers the highest level of care). More importantly, it shows text content of all treatment experiences (reviews) and number of patients treated by online consultation, telephone consultation, and outpatient visit. The website also calculates the overall review rating for each physician based on all reviews.

We developed a crawler to automatically download homepages of physicians and information about physicians from Haodf and Guahao. For Guahao, we crawled the active physicians who have added or modified outpatient information or individual information, and for Haodf, we crawled physicians who are active and provided outpatient visits. We completed the collection process for three periods (one week each in June, September, and December 2014). We used a difference-in-difference method to compare physician outpatient visits on Haodf and Guahao, so we needed to determine which physicians had a homepage on both websites. Our physician samples needed to be present on both websites during all sampling periods; after we matched physicians from two websites and three time points, the number of observations shrank. A total of 474 physicians were included in our research using this criterion. For each physician in our sample, we gathered their corresponding service and review information at each time point.

We collected the following data from each physician's homepage on both websites:

- All reviews for physician posted by patients until the day of our data collection, including the number of reviews for each physician and the overall review rating (on a scale of 0 to 10, 0 meaning very dissatisfied and 10 meaning very satisfied). The overall review rating reflects both medical quality and bedside manner of physician. Reviews on both websites come from patients who receives treatment at outpatient visits
- Number of outpatient visits for physician
- Number of online written and telephone consultations for physician (available on Haodf only)
- Date physician joined website (because length of time on the website can influence reviews, online services, and outpatient visits)

Variables and Models

Our empirical variables are shown in Table 1. Dependent variables are the number of outpatient visits on both websites, which are easily obtained from the physician homepages and represent the performance of physicians. We took the logarithmic value of the number of reviews, online services, and outpatient visits to stabilize the variance. The number of outpatient visits on each website is a function of a physician fixed effect (p_i), an online health care community website fixed effect (w_i), and other factors like the number of reviews. A physician fixed effect is related to factors such as age, education, gender, medical title of the physician, level of the hospital that the physician belongs to, and popularity of the physician. The online health care community website fixed effect is related to website design and patient preference.

We used *Houtpatient_care* and *Goutpatient_care* to denote the number of outpatient visits, *Hreview* and *Greview* to denote the number of patient reviews on Haodf (Figure 4) and Guahao (Figure 5), respectively (we allow Haodf reviews to influence patients on Guahao and Guahao reviews to influence patients on Haodf). Similarly, *Hrating* and *Grating* respectively represent the summary statistics of a physician's online reviews—the overall review rating. In addition, for Haodf, we consider two extra variables: *Hwritten_consultation* refers to the number of a physician's online written consultations and *Htelephone_consultation* denotes the number of online telephone consultations. The superscripts *H* and *G* refer to Haodf and Guahao, respectively.

XSL•FO
RenderX

Table 1. Variable description.

Variable and symbol	Explanation
Dependent variables	
Ln(Houtpatient_care)	Number of physician outpatient visits on Haodf (logarithmic form).
Ln(Goutpatient_care)	Number of physician outpatient visits on Guahao (logarithmic form).
Independent variables	
Hrating	Overall review rating of the physician on Haodf.
Grating	Overall review rating of the physician on Guahao.
Ln(Hreview)	Number of reviews on Haodf (logarithmic form).
Ln(Greview)	Number of reviews on Guahao (logarithmic form).
Ln(Hwritten_Consultation)	Number of physician online written consultations on Haodf (logarithmic form).
Ln(Htelephone_Consultation)	Number of physician online telephone consultations on Haodf (logarithmic form).
Moderating effects	
*Hrating*Ln(Hwritten_Consultation)*	Moderating effect of online written consultations on the relationship between reviews and outpatient visits.
*Grating*Ln(Hwritten_Consultation)*	same
*Ln(Hreview)*Ln(Hwritten_Consultation)*	same
*Ln(Greview)*Ln(Hwritten_Consultation)*	same
*Hrating*Ln(Htelephone_Consultation)*	Moderating effect of online telephone consultations on the relationship between reviews and outpatient visits.
*Grating*Ln(Htelephone_Consultation)*	same
*Ln(Hreview)*Ln(Htelephone_Consultation)*	same
*Ln(Greview)*Ln(Htelephone_Consultation)*	same
Control variables	
Htime	Opening date of physician homepage on Haodf.
Gtime	Opening date of physician homepage on Guahao.

Figure 4. Equation for Haodf.

$$\ln(Houtpatient_care_i) = \alpha_1^H Hrating_i + \alpha_2^H \ln(Hreviews_i) + \gamma_1^H Grating_i + \gamma_2^H \ln(Greviews_i) + \alpha_3^H \ln(Hwritten_consultation_i) + \alpha_4^H \ln(Htelephone_consultation_i) + w_i^H + p_i + \varepsilon_i^H \quad (1)$$

Figure 5. Equation for Guahao.

$$\ln(Goutpatient_care_i) = \alpha_1^G Grating_i + \alpha_2^G \ln(Greviews_i) + \gamma_1^G Hrating_i + \gamma_2^G \ln(Hreviews_i) + w_i^G + p_i + \varepsilon_i^G \quad (2)$$

We expect there are unobservable factors (fixed effects) that may affect the independent or dependent variable and cause a deviation of estimation if omitted. The physicians we collected are matched; for example, consider physician *i* in our

XSL•FO
RenderX

dataset—although they work on both websites and have homepages on both websites, they are exactly the same person and have exactly the same characteristics such as title, popularity, etc ($p_i^H=p_i^G$). We are able to control for the possible effect of unobserved physician characteristics on both reviews and outpatient visits and can eliminate physician fixed effects by differencing the data across websites (Figure 6).

Figure 6. Equation to eliminate physician fixed effects.

$$\begin{aligned}\ln(Houtpatient_care_i)-\ln(Goutpatient_care_i)&=\beta_1^H Hrating_i+\beta_2^H\ln(Hreviews_i)\\&+\beta_3^G Grating_i+\beta_4^G\ln(Greviews_i)+\beta_5^H\ln(Hwritten_consultation_i)\\&+\beta_6^H\ln(Htelephone_consultation_i)+w_i^H+w_i^G+\varepsilon_i\end{aligned}\tag{3}$$

For the online health care communities fixed effect, we first assume that both websites are virtually identical in terms of patient preference (ie, $w_i^H=w_i^G$), so we eliminate website fixed effects by differencing the data across websites (Figure 7).

Figure 7. Equation to eliminate online health care community fixed effects.

$$\begin{aligned}\ln(Houtpatient_care_i)-\ln(Goutpatient_care_i)&=\beta_1^H Hrating_i+\beta_2^H\ln(Hreviews_i)\\&+\beta_3^G Grating_i+\beta_4^G\ln(Greviews_i)+\beta_5^H\ln(Hwritten_consultation_i)\\&+\beta_6^H\ln(Htelephone_consultation_i)+\varepsilon_i\end{aligned}\tag{4}$$

However, if there are differences across the two websites (ie, $\mu_i^H\neq\mu_i^G$), we need to collect data for another time point and difference the data across the websites and time (Figure 8).

Figure 8. Equation to eliminate differences across the websites.

$$\begin{aligned}\Delta[\ln(Houtpatient_care_i)-\ln(Goutpatient_care_i)]&=\beta_1^H\Delta Hrating_i+\beta_2^H\Delta\ln(Hreviews_i)\\&+\beta_3^G\Delta Grating_i+\beta_4^G\Delta\ln(Greviews_i)+\beta_5^H\Delta\ln(Hwritten_consultation_i)\\&+\beta_6^H\Delta\ln(Htelephone_consultation_i)+\varepsilon_i\end{aligned}\tag{5}$$

All the above equations omit interaction terms. Accordingly, we add the moderating effects (Figures 7 and 8) in our empirical models. Formula expression is omitted to save space.

Results

Descriptive Statistics and Correlations

Tables 2 and 3 show the summary, description, and correlation of our variables. From Table 2, we can see there are obvious changes for all variables, which is helpful for empirical analysis. There are a few notable differences across the two websites that are apparent in Table 2. First, the mean of the difference between the number of outpatient visits on Haodf and Guahao is less than zero. This is consistent with the primary functions of the websites: Haodf provides many services for patients to choose, and its primary services are online services; Guahao specializes in providing outpatient care appointments. Second, Haodf has more reviews than Guahao. Third, the overall review rating is higher at Haodf; although again, they are overwhelmingly positive overall at both websites.

From Table 3, we can see the number of online written and telephone consultations positively impacts the difference between the number of outpatient visits of physicians on Haodf and Guahao. We can also see that the number of reviews on Haodf is positively related to the difference in the number of outpatient visits on the two websites. However, the overall review rating and number of reviews on Guahao are negatively related to the difference in outpatient visits on the two websites.

XSL•FO
RenderX

Table 2. Summary data.

Variable	Jun 2014 mean (standard error)		Sep 2014 mean (standard error)		June-December 2014 mean (standard error)	
	Haodf	Guahao	Haodf	Guahao	Haodf	Guahao
Lnrating	9.084 (2.560)	8.084 (2.072)	9.284 (2.220)	8.353 (2.141)	—	—
Lnreviews	3.450 (1.070)	2.893 (1.661)	3.483 (1.041)	3.103 (1.616)	—	—
Lnwritten consultation	6.363 (1.749)	—	6.485 (1.663)	—	—	—
Lntelephone consultation	1.781 (1.931)	—	1.987 (1.998)	—	—	—
LnHoutpatient_care-LnGoutpatiet_care	–1.380 (2.370)	—	—	—	–1.662 (2.209)	—

Table 3. Description and correlation.

Variable	1	*P* value	2	*P* value	3	*P* value	4	*P* value	5	*P* value	6	*P* value
1. LnHoutpatient_care-LnGoutpatiet_care	—	—	—	—	—	—	—	—	—	—	—	—
2. LnHwritten_consultation	0.323	.04	—	—	—	—	—	—	—	—	—	—
3. LnHtelephone_consultation	0.195	.03	0.453	.02	—	—	—	—	—	—	—	—
4. Hrating	0.037	.22	0.184	.01	0.051	.34	—	—	—	—	—	—
5. LnHreview	0.225	.03	0.577	.01	0.419	.02	0.278	.01	—	—	—	—
6. Grating	–0.316	.02	–0.060	.23	0.065	.21	–0.034	.32	0.034	.43	—	—
7. LnGreview	–0.660	.03	0.110	.07	0.102	.07	0.070	.44	0.222	.02	0.359	.03

Empirical Results

Results Without Considering the Website-Specific Fixed Effects

We used an ordinary least squares regression model for analysis using STATA (StataCorp LLC) software. We first assume there were no website-specific fixed effects and examined the model (Figure 7). Table 4 shows the estimation results. Column 1 presents the results of the control variables. As we chose physician *i*, who provides services on both websites, the physician individual characteristics did not need to be considered. We included the opening time (duration of use) for each physician in column 1. A longer time on the website may lead to having more patients and affect important variables in our model. Columns 2 and 3 introduce these independent variables. Both written and telephone consultations increase the difference in outpatient visits (written consultation: β=0.325, *P*<.001; telephone consultation: β=0.093, *P*=.005), and this suggests there are complementary effects between online services and outpatient visits. Physicians can use online services to attract more patients to have treatment in hospitals. The coefficient of the number of reviews on Haodf is positive and statistically significant (β=0.518, *P*<.001), suggesting that when reviews increase, visits on Haodf becomes larger. However, the ratings on Haodf do not significantly impact the difference in outpatient visits. The overall review rating on Haodf is 9.084, which is very high compared with the full mark (ie, 10). High overall review ratings make it more likely patients will discount the reviews and not use them for decision making. Again, when ratings and number of reviews rise on Guahao, the difference in outpatient visits decreases (ie, outpatient visit increases on Guahao relative to Haodf; rating: β=–0.066, *P*=.009; number of reviews: β=–1.037, *P*<.001). The absolute value of the coefficient of the number of reviews on Guahao is bigger than on Haodf, suggesting that difference in visits responds more to the number of reviews on Guahao than on Haodf. This is consistent with the main function of the two sites. Guahao only provides outpatient care appointments, and patients can only refer to the reviews from other patients to make choices. However, in addition to offline services, Haodf also provides online services, so there is more information for patients to make choices.

Column 4 in Table 4 includes the interaction effects. As the impact of ratings on Haodf is not significant, we only introduce the interaction terms of significant factors. Online services negatively moderate the relationship between the number of reviews on Haodf and difference in visits (written consultations and reviews: β=–0.101, *P*<.001; telephone consultations and reviews β=–0.011, *P*=.04). However, the moderating effects are not statistically significant for the number of reviews on Guahao. When a physician provides online services, the impact of ratings on Guahao on difference in visits declines (written consultations: β=0.019, *P*=.01; telephone consultation: β=0.033, *P*=.04). The adjusted R^2 is 64.2%; these variables explain the independent variable well.

XSL•FO
RenderX

Table 4. The effect of online services and reviews on outpatient visits (sample is the complete June 2014 sample. Dependent variable is the difference between the log outpatient visits on Haodf and the log outpatient visits on Guahao. Dependent variable is Ln(Houtpatient_carei)-Ln(Goutpatiet_carei).

Variable	Model 1 Coefficient (robust standard error)	*P* value	Model 2 Coefficient (robust standard error)	*P* value	Model 3 Coefficient (robust standard error)	*P* value	Model 4 Coefficient (robust standard error)	*P* value
HTime	0.289 (0.058)	.11	0.424 (0.055)	.11	0.036 (0.035)	.11	0.108 (0.065)	.04
GTime	–0.154 (0.107)	.12	–0.086 (0.102)	.12	0.102 (0.066)	.12	0.035 (0.035)	.12
LnHwritten_Consultation	—	—	0.397 (0.066)	<.001	0.325 (0.048)	<.001	–0.191 (0.197)	.11
LnHtelephone_Consultation	—	—	0.075 (0.060)	.10	0.093 (0.039)	.005	0.397 (0.219)	.04
Hrating	—	—	—	—	–0.028 (0.027)	.23	–0.003 (0.028)	.22
LnHreview	—	—	—	—	0.518 (0.081)	<.001	0.116 (0.244)	.03
Grating	—	—	—	—	–0.066 (0.034)	.009	–0.163 (0.143)	.04
LnGreview	—	—	—	—	–1.037 (0.043)	<.001	–1.119 (0.170)	<.001
*LnHwritten_Consultation*LnHreview*	—	—	—	—	—	—	–0.101 (0.039)	<.001
*LnHtelephone_Consulation*LnHreview*	—	—	—	—	—	—	–0.011 (0.039)	.04
*LnHwritten_Consultation*Grating*	—	—	—	—	—	—	0.019 (0.021)	.01
*LnHwritten_Consultation*LnGreview*	—	—	—	—	—	—	0.015 (0.028)	.45
*LnHtelephone_Consulation*Grating*	—	—	—	—	—	—	0.033 (0.206)	.04
*Lntelephone_Consulation*LnGreview*	—	—	—	—	—	—	–0.001 (0.026)	.31
Adjusted R^2	0.0002	—	0.101	—	0.634	—	0.642	—
N	474	—	474	—	474	—	474	—

Results With Considering the Website-Specific Fixed Effects

The websites have different characteristics, so omitting the website-specific fixed effects may bias the estimation results. In this section, we estimate the equation seen in Figure 8. The results are shown in Table 5.

The homepages for all 474 physicians on both websites existed during the second period. Columns 1 and 2 on Table 5 include the independent variables. The coefficients of the number of reviews are higher in magnitude than on Table 4, even though some are no longer significant. The impacts of ratings on both websites are not significant. This may be due to relatively little variance in the overall review rating over time. Most of the results of the previous section are replicated. Thus, there are complementary effects between online services and visits (written consultations: β=0.172, *P*=.03; telephone consultations: β=0.155, *P*<.001), and therefore hypotheses H2a and H2b are supported. An increase in the number of reviews on Haodf over time results in a higher number of visits to the physician on Haodf over time (β=0.588, *P*<.001); the same is true for the number of reviews on Guahao (β=–1.661, *P*<.001), supporting hypothesis H1b.

Column 3 on Table 5 shows the results of moderating effects. We only introduce the moderating effects of significant factors. The results are almost the same as we predicted (hypothesis H3b is partly supported), except the moderating effect between telephone consultations and number of reviews on Haodf is not significant. When a physician provides online written consultations, the impact of reviews declines (β=–0.829, *P*=.004). When a physician provides online services, the impact of reviews on Guahao for visits declines (written consultations: β=0.730, *P*=.03; telephone consultations: β=0.296, *P*=.009). The adjusted R^2 is 42.7%, which has declined compared with the same value on Table 4.

XSL•FO
RenderX

Table 5. The effect of changes in online services and reviews on changes in visits over 2 months (sample is the set of physicians who were available on both websites in June and December 2014. Reviews were collected in June and September 2014. Dependent variable is Δ[Ln(Houtpatient_carei)-Ln(Goutpatiet_carei)]).

Variable	Model 1 Coefficient (robust standard error)	*P* value	Model 2 Coefficient (robust standard error)	*P* value	Model 3 Coefficient (robust standard error)	*P* value
Δ*Hrating*	–0.001 (0.021)	.50	–0.001 (0.208)	.50	–0.003 (0.208)	.50
Δ*LnHreview*	0.591 (0.127)	<.001	0.588 (0.126)	<.001	0.655 (0.143)	<.001
Δ*Grating*	0.049 (0.049)	.12	0.043 (0.048)	.13	0.049 (0.048)	.13
Δ*LnGreview*	–1.643 (0.097)	—	–1.661 (0.096)	<.001	–1.827 (0.119)	<.001
Δ*LnHwritten_Consultation*	—	—	0.172 (0.097)	.03	0.175 (0.151)	.04
Δ*LnHtelephone_Consultation*	—	—	0.155 (0.058)	<.001	0.049 (0.073)	.03
Δ*LnHwritten_Consultation*Δ*LnHreview*	—	—	—	—	–0.829 (0.359)	.004
Δ*LnHtelephone_Consultation*Δ*LnHreview*	—	—	—	—	–0.057 (0.349)	.54
Δ*LnHwritten_Consultation*Δ*LnGreview*	—	—	—	—	0.730 (0.567)	.03
Δ*LnHtelephone_Consultation*Δ*LnGreview*	—	—	—	—	0.296 (0.156)	.009
Adjusted R^2	0.39	—	0.411	—	0.427	—
N	474	—	474	—	474	—

Robustness Check

We examine the robustness of our estimations in Tables 4 and 5. For Table 4, we repeat the specification of column 4, but we examine only the subsample of 400 physicians who have at least one review on each website. The results are shown in column 1 on Table 5 and are similar to those we presented previously. All signs of the coefficients are as we predicted. For Table 5, we only include physicians who have at least one review variable changed; we repeated the equation found in Figure 8 by using the subsample 371, and the results are shown in column 2 on Table 6. The results prove the robustness of our empirical results.

Table 6. Robustness check results (for column 1, sample is the subsample of physicians who had at least one review on both websites in June 2014, and dependent variable is Ln(Houtpatient_carei)-Ln(Goutpatiet_carei). For column 2, sample is the subsample of physicians who had new reviews posted on both websites between June and September 2014).

Variable	Model 1 Coefficient (robust standard error)	*P* value	Model 2 Coefficient (robust standard error)	*P* value
HTime	0.104 (0.064)	.01	—	—
GTime	0.035 (0.034)	.22	—	—
LnHwritten_Consultation	–0.035 (0.205)	.46	0.189 (0.020)	.01
LnHtelephone_Consultation	0.237 (0.204)	.03	0.067 (0.012)	.03
Hrating	0.008 (0.033)	.43	0.001 (0.002)	.43
LnHreview	0.219 (0.278)	.04	0.738 (0.201)	<.001
Grating	–0.023 (0.143)	.04	0.002 (0.024)	.06
LnGreview	–1.083 (0.143)	<.001	–0.205 (0.121)	<.001
*LnHwritten_Consultation*LnHreview*	–0.105 (0.044)	<.001	–0.988 (0.273)	.009
*LnHtelephone_Consulation*LnHreview*	–0.006 (0.038)	.03	–0.105 (0.023)	.05
*LnHwritten_Consultation*Grating*	0.004 (0.022)	.04	—	—
*LnHwritten_Consultation*LnGreview*	0.027 (0.032)	.32	0.870 (0.556)	.03
*LnHtelephone_Consulation*Grating*	0.012 (0.019)	.03	—	—
*Lntelephone_Consulation*LnGreview*	–0.015 (0.027)	.32	0.443 (0.154)	.03
Adjusted R^2	0.613	—	0.408	—
N	400	—	371	—

XSL•FO
RenderX

Discussion

Principal Findings

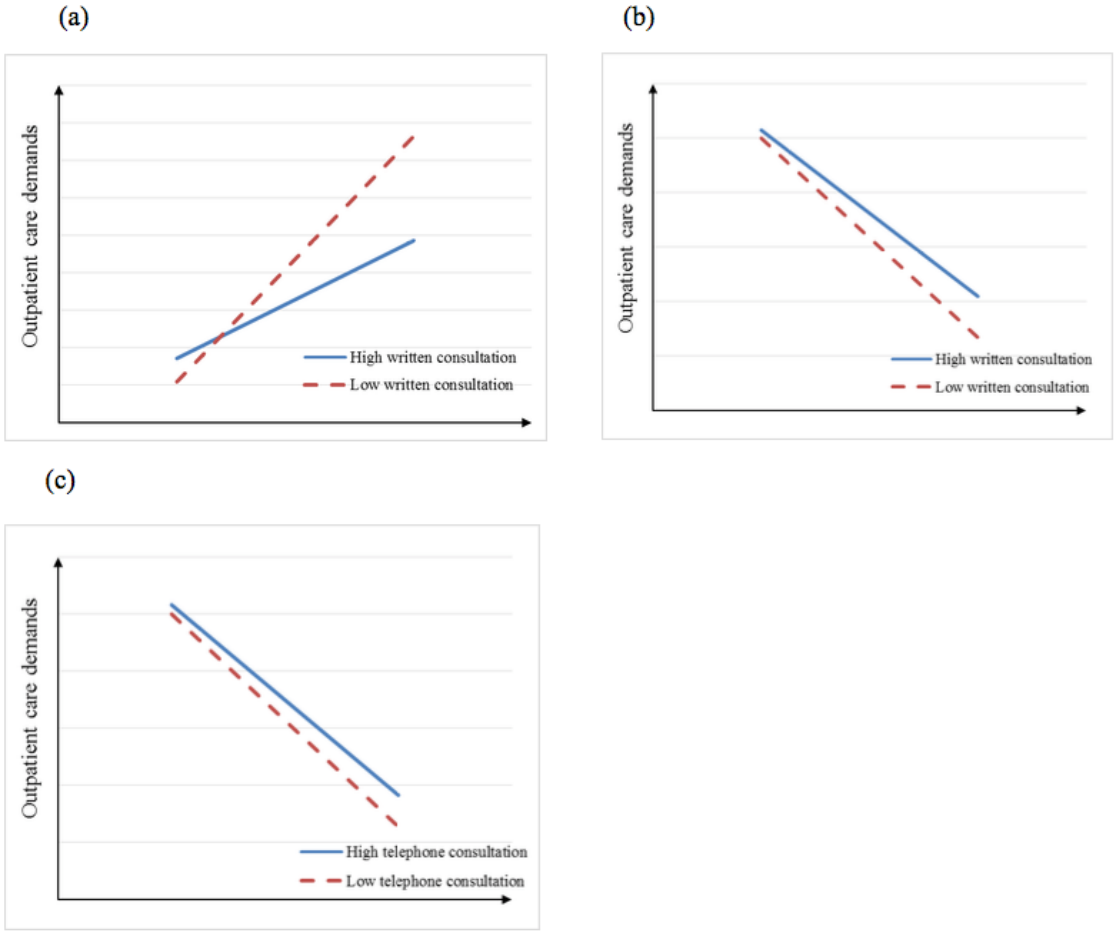
We studied the role of reviews in the health care industry and found that the number of reviews tended to have positive impacts on both websites. Our empirical results show that patients value the number of reviews more than the average rating when making decisions. The evidence suggests that physicians should try to improve their service quality and attitude to attract more patients to write reviews for them. Our regression estimates show that the relative visits of a physician across the two websites are related to the differences across the websites in the number of reviews.

When we used the equation found in Figure 8 to eliminate the physician- and website-specific fixed effects, the effects of overall review ratings for both websites were no longer significant. This finding differs from prior studies, which generally saw significant and positive effects of the overall review rating [44-46]. First, we found the overall review rating was very high on both websites—much higher than in other fields such as e-commerce. One possible explanation is that the health care industry in China is facing intense physician-patient conflicts [56]. Possible manipulations in reviews on websites may exist [57] such as deleting negative reviews. Another possible explanation is that many diseases (eg, chronic diseases) require follow-up, and patients may be afraid of being retaliated against by physicians they review poorly. From the data summary on Table 3, we show that the mean value of the overall review rating is over 9.0 on Haodf and 8.0 on Guahao, which may seem artificially inflated to patients. Second, we use a more sophisticated econometric method, difference-in-difference, to eliminate physician- and website-specific fixed effects, which may not have been fully addressed in other studies.

For channel effects, our results show that online services complement offline services (outpatient visits), which is consistent with our hypotheses. Online services can help patients get more information but cannot replace face-to-face service. By first having a written or telephone consultation with a physician, patients gain a basic understanding of their disease, and then they can see the physician in the hospital for further details.

For the interaction effects, our results show that online services mitigate the relationship between reviews and outpatient visits (Table 5). The interaction effects are illustrated in Figure 9. We only draw the moderating effects in Table 5. Haodf provides online written and telephone consultations. We show that these two kinds of services significantly affect outpatient visits in our empirical results, and this eliminates the impact of reviews of Haodf on patient choice.

Figure 9. Images (a) and (b) show the moderating effects of written consultations on the relationship between the number of reviews on Haodf/Guahao and outpatient visits. Image (c) shows the moderating effect of telephone consultations on the relationship between the number of reviews on Guahao and outpatient visits.



XSL•FO
RenderX

Limitations

This paper has several limitations. First, we do not explore the context of the reviews. We just study the overall review rating and number of reviews, and this could affect the impact of reviews. For example, some reviews contain more information, and these kinds of reviews may have more impact. Second, we studied only two contexts. This helped us improve the internal validity, but it may have also reduced the generalizability of our findings, and future research should validate our results in more service contexts. Future research can take more effective empirical methods to solve these limitations.

Implications of the Research

Our study has important theoretical and practical implications. For the theoretical implications, our study enriches the research on the role of reviews by investigating them in the health care field. Existing studies mainly research them in the marketplace. Moreover, we have addressed the endogeneity of reviews—self-selected problem and the impact of underlying factors (eg, service quality) in our paper. We use a difference-in-difference method to account for physician- and website-specific effects. Our paper has important theoretical implications for research in health care. Second, despite some studies indicating that there are channel effects in the marketplace [12,16], literature rarely uses empirical methods to validate claims. Our study is among the first to use real data to empirically examine the channel effect, especially in health care, which is a universally beneficial sector. The research contexts allow us to study the effects of two kinds of online channels on offline channels, and our results show that there are channel effects in health care. Third, our study contributes to existing theories of reviews and channel effects by hypothesizing and empirically testing the moderating influence of online channels on the relationship between online reviews and offline channels. In analyzing existing literature, we found that there were few studies combining them. Although some researchers have studied the importance of reviews [7,20], few studies consider their effect on the relationship between online and offline channels.

This paper also makes contributions to practice. First, multichannel use is on the rise, with practitioners seeking guidance on how to balance different channels. Our empirical results show that a multichannel strategy is helpful for physicians to access more patients. Therefore, we believe our analysis provides insights that are helpful to physicians as they consider implementing a multichannel strategy such as providing online consultation services to patients. Second, based on the empirical results, our study gives physicians suggestions to improve their reviews (both medical quality and bedside manner), such as learning more to improve medical skills. Moreover, we found that the overall review rating is not always effective in influencing patient decisions and recommend that physicians encourage and remind patients to write reviews for them. Third, our study highlights the importance of rethinking the nature of reviews in relation to multichannel strategies. Our study shows online services have a significant moderating effect on the relationship between reviews and outpatient visits. This result indicates that even if physicians have lower reviews, they can improve their career outcomes by working hard in an online health care community. Our study has proved that online health communities benefit not only patients but also physicians. These results can encourage physicians to attract more patients and achieve their career goals by participating in online communities.

Conclusions

A majority of empirical work has examined the role of online reviews in consumer decisions. However, less evidence has been found in health care, and endogeneity of online reviews has not been fully considered. Moreover, the important factor of physician online services has been neglected in patient decisions. To address this research gap, this study investigates the effect of reviews and channel effects in health care by conducting a difference-in-difference analysis on two online health care communities. Our empirical results show that compared with average rating, patients consider number of reviews more when making decisions. The evidence suggests that physicians should try to improve their service quality and attitude to attract more patients to write reviews for them. Our regression estimates show that the number of visits to a physician across the two websites is related to the differences across the websites in the number of reviews. Our findings provide basic research on online health care communities and have both theoretical and practical implications.

Conflicts of Interest

None declared.

References

1. Xu Y, Armony M, Ghose A. The effect of online reviews on physician demand: a structural model of patient choice. SSRN J 2016. [doi: 10.2139/ssrn.2778664]
2. Fox S, Duggan M. Health online 2013. Washington: Pew Internet and American Life Project; 2013 Jan 15. URL: https://www.pewinternet.org/2013/01/15/health-online-2013/ [accessed 2019-10-19]
3. Ifeng. Phoenix New Media. 2011 Jun 24. More Chinese patients search online before seeking medical treatment offline URL: http://news.ifeng.com/gundong/detail_2011_06/24/7232672_0.shtml [accessed 2019-10-19]
4. Harris K, Buntin M. Choosing a health care provider: the role of quality information. 2008. URL: https://pdfs.semanticscholar.org/dcfe/5b064f7160a72af7d038f8e7362cfcad7577.pdf [accessed 2019-10-19]

XSL•FO
RenderX

5. Jha N, Premarajan K, Nagesh S, Khanal S, Thapa L. Five-star doctors for the 21st Century: a BPKIHS endeavour for Nepal. J Health Manag 2005 Dec 21;7(2):237-247. [doi: 10.1177/097206340500700205]
6. Chintagunta PK, Gopinath S, Venkataraman S. The effects of online user reviews on movie box office performance: accounting for sequential rollout and aggregation across local markets. Mark Sci 2010 Sep;29(5):944-957. [doi: 10.1287/mksc.1100.0572]
7. Forman C, Ghose A, Wiesenfeld B. Examining the relationship between reviews and sales: the role of reviewer identity disclosure in electronic markets. Info Sys Res 2008 Sep;19(3):291-313. [doi: 10.1287/isre.1080.0193]
8. Lu N, Wu H. Exploring the impact of word-of-mouth about physicians' service quality on patient choice based on online health communities. BMC Med Inform Decis Mak 2016 Dec 26;16(1):151 [FREE Full text] [doi: 10.1186/s12911-016-0386-0] [Medline: 27888834]
9. Geyskens I, Gielens K, Dekimpe MG. The market valuation of internet channel additions. J Mark 2018 Oct 10;66(2):102-119. [doi: 10.1509/jmkg.66.2.102.18478]
10. Maity M, Dass M. Consumer decision-making across modern and traditional channels: e-commerce, m-commerce, in-store. Decis Supp Sys 2014 May;61:34-46. [doi: 10.1016/j.dss.2014.01.008]
11. Wu H, Lu N. Online written consultation, telephone consultation and offline appointment: an examination of the channel effect in online health communities. Int J Med Inform 2017 Nov;107:107-119. [doi: 10.1016/j.ijmedinf.2017.08.009] [Medline: 29029686]
12. Bakos JY. Reducing buyer search costs: implications for electronic marketplaces. Manag Sci 1997 Dec;43(12):1676-1692. [doi: 10.1287/mnsc.43.12.1676]
13. Malone TW, Yates J, Benjamin RI. Electronic markets and electronic hierarchies. Commun ACM 1987;30(6):484-497. [doi: 10.1145/214762.214766]
14. Naik PA, Peters K. A hierarchical marketing communications model of online and offline media synergies. J Interact Mark 2009 Nov;23(4):288-299. [doi: 10.1016/j.intmar.2009.07.005]
15. Saeed KA, Grover V, Hwang Y. Creating synergy with a clicks and mortar approach. Commun ACM 2003 Dec 01;46(12):206. [doi: 10.1145/953460.953501]
16. Ward MR. Will online shopping compete more with traditional retailing or catalog shopping? Netnomics 2001;3:103. [doi: 10.1023/A:1011451228921]
17. Haodf.com. URL: https://www.haodf.com/ [accessed 2019-10-21]
18. Guahao.com. URL: https://www.guahao.com/ [accessed 2019-10-21]
19. Wu H, Lu N. How your colleagues' reputation impact your patients' odds of posting experiences: evidence from an online health community. Electron Commerce Res Appl 2016 Mar;16:7-17. [doi: 10.1016/j.elerap.2016.01.002]
20. Chevalier JA, Mayzlin D. The effect of word of mouth on sales: online book reviews. J Mark Res 2018 Oct 10;43(3):345-354. [doi: 10.1509/jmkr.43.3.345]
21. Hughes B, Joshi I, Wareham J. Health 2.0 and Medicine 2.0: tensions and controversies in the field. J Med Internet Res 2008;10(3):e23 [FREE Full text] [doi: 10.2196/jmir.1056] [Medline: 18682374]
22. Ba S, Wang L. Digital health communities: the effect of their motivation mechanisms. Decis Supp Sys 2013 Nov;55(4):941-947. [doi: 10.1016/j.dss.2013.01.003]
23. Wu H, Lu N. Service provision, pricing, and patient satisfaction in online health communities. Int J Med Inform 2018 Feb;110:77-89. [doi: 10.1016/j.ijmedinf.2017.11.009] [Medline: 29331257]
24. PatientsLikeMe. URL: https://www.patientslikeme.com/ [accessed 2019-10-23]
25. Ni J, Sun B. A dynamic game of doctors? participation in online health information platform. 2010 Presented at: Workshop on Health IT and Economics; 2010; College Park.
26. Wu H, Deng Z. Knowledge collaboration among physicians in online health communities: a transactive memory perspective. Int J Info Manag 2019 Dec;49:13-33. [doi: 10.1016/j.ijinfomgt.2019.01.003]
27. Yan L, Tan Y. Feeling blue? Go online: an empirical study of social support among patients. Info Sys Res 2014 Dec;25(4):690-709. [doi: 10.1287/isre.2014.0538]
28. Xiao N, Sharman R, Rao H, Upadhyaya S. Factors influencing online health information search: an empirical analysis of a national cancer-related survey. Decis Supp Sys 2014 Jan;57:417-427. [doi: 10.1016/j.dss.2012.10.047]
29. Bansal G, Zahedi F, Gefen D. The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. Decis Supp Sys 2010 May;49(2):138-150. [doi: 10.1016/j.dss.2010.01.010]
30. Cheung CM, Lee MK. What drives consumers to spread electronic word of mouth in online consumer-opinion platforms. Decis Supp Sys 2012 Apr;53(1):218-225. [doi: 10.1016/j.dss.2012.01.015]
31. Jeong E, Jang S. Restaurant experiences triggering positive electronic word-of-mouth (eWOM) motivations. Int J Hospitality Manag 2011 Jun;30(2):356-366. [doi: 10.1016/j.ijhm.2010.08.005]
32. Gao G, Greenwood BN, Agarwal R, McCullough JS. Vocal minority and silent majority: how do online ratings reflect population perceptions of quality? MISQ 2015;39:565. [doi: 10.2139/ssrn.2629837]
33. Lee Y, Hosanagar K, Tan Y. Do I follow my friends or the crowd? Information cascades in online movie ratings. Manag Sci 2015 Sep;61(9):2241-2258. [doi: 10.1287/mnsc.2014.2082]
34. Dichter E. How word-of-mouth advertising works. Harvard Bus Rev 1966;44:147. [doi: 10.4135/9781452229669.n3968]

XSL•FO
RenderX

35. Zhu F, Zhang X. Impact of online consumer reviews on sales: the moderating role of product and consumer characteristics. J Mark 2010 Mar;74(2):133-148. [doi: 10.1509/jm.74.2.133]
36. Ba S, Pavlou PA. Evidence of the effect of trust building technology in electronic markets: price premiums and buyer behavior. MISQ 2002 Sep;26(3):243. [doi: 10.2307/4132332]
37. Chen C. Understanding the effects of EWOM on cosmetic consumer behavioral intention. Int J Electronic Commerce Stud 2014 Jun;5(1):97-102. [doi: 10.7903/ijecs.1030]
38. Chen Y, Xie J. Online consumer review: word-of-mouth as a new element of marketing communication mix. Manag Sci 2008 Mar;54(3):477-491. [doi: 10.1287/mnsc.1070.0810]
39. Pavlou PA, Gefen D. Building effective online marketplaces with institution-based trust. Info Sys Res 2004 Mar;15(1):37-59. [doi: 10.1287/isre.1040.0015]
40. Chatterjee P. Online reviews: do consumers use them? Adv Consum Res 2001;28:129-133.
41. Yang J, Mai E. Experiential goods with network externalities effects: an empirical study of online rating system. J Bus Res 2010 Sep;63(9-10):1050-1057. [doi: 10.1016/j.jbusres.2009.04.029]
42. Sillence E, Briggs P, Harris PR, Fishwick L. How do patients evaluate and make use of online health information? Soc Sci Med 2007 May;64(9):1853-1862. [doi: 10.1016/j.socscimed.2007.01.012] [Medline: 17328998]
43. Sillence C, Hardy C, Broggs P. Why don't we trust health websites that help us help each other? An analysis of online peer-to-peer healthcare. 2013 Presented at: Proceedings of the 5th Annual ACM Web Science Conference,; 2013; Paris p. 396-404. [doi: 10.1145/2464464.2464488]
44. Clemons EK, Gao GG, Hitt LM. When online reviews meet hyperdifferentiation: a study of the craft beer industry. J Manag Info Sys 2014 Dec 08;23(2):149-171. [doi: 10.2753/mis0742-1222230207]
45. Ghose A, Ipeirotis P. Towards an understanding of the impact of customer sentiment on product sales and review quality. Info Technol Sys 2006:1-6.
46. Dellarocas C, Zhang X, Awad NF. Exploring the value of online product reviews in forecasting sales: the case of motion pictures. J Interact Mark 2007 Jan;21(4):23-45. [doi: 10.1002/dir.20087]
47. Duan W, Gu B, Whinston AB. Do online reviews matter? An empirical investigation of panel data. Decis Supp Sys 2008 Nov;45(4):1007-1016. [doi: 10.1016/j.dss.2008.04.001]
48. Lebo H. UCLA Center for Communication Policy. 2003. UCLA Internet Report: Surveying the Digital Future, Year 3 URL: https://www.digitalcenter.org/wp-content/uploads/2013/02/2003_digital_future_report-year3.pdf [accessed 2019-10-19]
49. Friedman G. In: Furey TR, editor. The Channel Advantage: Going to Market with Multiple Sales Channels to Reach More Customers, Sell More Products, Make More Profit. Oxford: Butterworth-Heinemann; 1999.
50. Frank MW. Media substitution in advertising: a spirited case study. Int J Indust Organ 2008 Jan;26(1):308-326. [doi: 10.1016/j.ijindorg.2007.01.002]
51. Athey S, Gans JS. The impact of targeting technology on advertising markets and media competition. Am Econom Rev 2010 May;100(2):608-613. [doi: 10.1257/aer.100.2.608]
52. Srivastava RK, Alpert MI, Shocker AD. A customer-oriented approach for determining market structures. J Mark 2018 Nov 02;48(2):32-45. [doi: 10.1177/002224298404800203]
53. Srivastava RK, Leone RP, Shocker AD. Market structure analysis: hierarchical clustering of products based on substitution-in-use. J Mark 2018 Nov 28;45(3):38-48. [doi: 10.1177/002224298104500303]
54. Ye Q, Deng Z, Chen Y, Liao J, Li G, Lu Y. How resource scarcity and accessibility affect patients' usage of mobile health in China: resource competition perspective. JMIR Mhealth Uhealth 2019 Aug 09;7(8):e13491 [FREE Full text] [doi: 10.2196/13491] [Medline: 31400104]
55. Lee KC, Kang I, McKnight DH. Transfer From offline trust to key online perceptions: an empirical study. IEEE Trans Eng Manag 2007 Nov;54(4):729-741. [doi: 10.1109/tem.2007.906851]
56. Daily Y. More than 80% physicians have suffered physical or language violence. 2015. URL: http://news.hexun.com/2015-04-24/175268483.html [accessed 2019-10-19]
57. lu SF, Rui H. Can we trust online physician ratings? Evidence from cardiac surgeons in Florida. Manag Sci 2015:2876-2885. [doi: 10.1109/hicss.2015.348]

Edited by Z Huang; submitted 09.09.19; peer-reviewed by X Lu, Q Ye; comments to author 08.10.19; revised version received 09.10.19; accepted 14.10.19; published 02.12.19.

Please cite as:
Lu W, Wu H
How Online Reviews and Services Affect Physician Outpatient Visits: Content Analysis of Evidence From Two Online Health Care Communities
JMIR Med Inform 2019;7(4):e16185
URL: http://medinform.jmir.org/2019/4/e16185/
doi:10.2196/16185
PMID:31789597

XSL•FO
RenderX

©Wei Lu, Hong Wu. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 02.12.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Measuring Regional Quality of Health Care Using Unsolicited Online Data: Text Analysis Study

Roy Johannus Petrus Hendrikx[1], MSc, PhD; Hanneke Wil-Trees Drewes[2], MSc, PhD; Marieke Spreeuwenberg[3,4], MSc, PhD; Dirk Ruwaard[4], MSc, PhD; Caroline Baan[1,2], MSc, PhD

[1]Tranzo Scientific Center for Care and Welfare, Tilburg University, Tilburg, Netherlands

[2]Center for Nutrition, Prevention and Health Services, National Institute for Public Health and the Environment, Bilthoven, Netherlands

[3]Zuyd University of Applied Sciences, Heerlen, Netherlands

[4]Department of Health Services Research, Care and Public Health Research Institute, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, Netherlands

Corresponding Author:
Roy Johannus Petrus Hendrikx, MSc, PhD
Tranzo Scientific Center for Care and Welfare
Tilburg University
Warandelaan 2
Tilburg, 5000 LE
Netherlands
Phone: 31 611647091
Email: roy.hendrikx@rivm.nl

Abstract

Background: Regional population management (PM) health initiatives require insight into experienced quality of care at the regional level. Unsolicited online provider ratings have shown potential for this use. This study explored the addition of comments accompanying unsolicited online ratings to regional analyses.

Objective: The goal was to create additional insight for each PM initiative as well as overall comparisons between these initiatives by attempting to determine the reasoning and rationale behind a rating.

Methods: The Dutch Zorgkaart database provided the unsolicited ratings from 2008 to 2017 for the analyses. All ratings included both quantitative ratings as well as qualitative text comments. Nine PM regions were used to aggregate ratings geographically. Sentiment analyses were performed by categorizing ratings into negative, neutral, and positive ratings. Per category, as well as per PM initiative, word frequencies (ie, unigrams and bigrams) were explored. Machine learning—naïve Bayes and random forest models—was applied to identify the most important predictors for rating overall sentiment and for identifying PM initiatives.

Results: A total of 449,263 unsolicited ratings were available in the Zorgkaart database: 303,930 positive ratings, 97,739 neutral ratings, and 47,592 negative ratings. Bigrams illustrated that feeling like not being “taken seriously” was the dominant bigram in negative ratings, while bigrams in positive ratings were mostly related to listening, explaining, and perceived knowledge. Comparing bigrams between PM initiatives showed a lot of overlap but several differences were identified. Machine learning was able to predict sentiments of comments but was unable to distinguish between specific PM initiatives.

Conclusions: Adding information from text comments that accompany online ratings to regional evaluations provides insight for PM initiatives into the underlying reasons for ratings. Text comments provide useful overarching information for health care policy makers but due to a lot of overlap, they add little region-specific information. Specific outliers for some PM initiatives are insightful.

(JMIR Med Inform 2019;7(4):e13053) doi:10.2196/13053

KEYWORDS

text mining; population health management; regional care; quality of care; online data; big data; patient-reported experience measures

XSL•FO
RenderX

Introduction

With respect to evaluating experienced quality of care, unsolicited online ratings given to health care providers have received more and more attention. This is a shift away from a past focus on solicited surveys. Studies have shown the potential of unsolicited data as a valuable resource to provide insight into the quality of care experienced at the provider level [1-3]. Furthermore, online data have some very interesting properties for policy makers and researchers, as they tend to be easier to collect, have a bigger reach, are generally cheaper, are consistently updated, and can consist of more responses than solicited surveys [1,4].

Insight into how experienced quality of care can be improved is a pivotal challenge for population management (PM) initiatives. The rising costs, changing care demands, and issues with the provided quality of care are pushing policy makers to take new approaches. Instead of health care being a reactive system based on individual demands, it should be a proactive system organized around a population's needs [5,6]. This requires a whole-system approach in order to improve quality and efficiency, including prevention. As a result, reforms designated as population health management are becoming more and more widespread in health policy. Even though different definitions exist [7], PM initiatives generally focus on the health needs of a specified population across the continuum of health and well-being by introducing multiple interventions that organize services related to health and social care, as well as prevention and welfare [7,8]. PM initiatives often strive to achieve the Triple Aim by shifting focus from individuals to populations and by integrating care across health and social domains [8,9]. The Triple Aim was introduced by Berwick et al in 2008 and requires the simultaneous pursuit of improving population health and experienced quality of care, while reducing costs [10]. Examples of PM initiatives include the American Accountable Care Organizations [11], the National Health Service's Vanguard sites [12], and the Dutch pioneer sites [13]. For the pursuit of the Triple Aim by such initiatives to be successful, each of the Triple Aim's three pillars needs to be evaluated at the population or often regional level. Unsolicited online data could be a valuable source for evaluating the experienced-quality-of-care pillar. However, a previous study, utilizing the same dataset used in this study, explored rating distributions and applied multilevel analyses. Results from these analyses suggested that when using only the available quantitative data, their use at the regional level is limited [14]. First, while differences in mean ratings between providers were caused by differences in provider-specific characteristics, regional differences could not be attributed to differences in regional characteristics. This means that any variation in mean rating between regions does not point to a structural difference in, for example, quality of care or population. Second, no insight was provided regarding the reasoning behind any given rating and why it was either negative or positive. Additional methods and/or data are needed to make unsolicited data more valuable for regional initiatives.

Text comments could be able to provide a solution for the lack of regional specificity and reasoning of unsolicited provider ratings. Much of the created online data comes in the form of text; examples include tweets, Facebook posts, forum comments, and others. In health care, most rating websites provide patients with the opportunity to add comments to their ratings as well. Comments are already used for, among other things, competitive analyses and consumer sentiment analyses [1,15]. Combining ratings with their comments in analyses can provide insight into the reasoning and rationale behind a positive or negative rating [16,17]. Typically, interviews would have to be conducted to determine reasoning. However, at the population scale, conducting interviews is a costly and time-consuming endeavor and unsolicited data could significantly help in this regard. Despite the potential, adding comments to the accompanying unsolicited provider ratings when evaluating differences in experienced quality of care between regional initiatives has not yet been explored.

This study explores whether adding text comments—that accompany ratings—to regional analysis can provide additional insight into evaluating experienced quality of care. The goal is to determine the comments' value for PM initiatives individually as well as when comparing initiatives. The largest health care ratings website in the Netherlands will be studied using different sentiment and machine learning analyses.

Methods

Dataset

The Zorgkaart Nederland website [18] provided the unsolicited online patient ratings. On this website, patients can both give and see reviews. To add a review, patients first select a care provider, which can be a care professional, such as a specific general practitioner or specialist, or an organization, such as a hospital department or nursing home. Quantitative data included ratings for six quality-of-care dimensions. These ranged from 1 to 10, with 1 indicating the worst experience and 10 the best. The six rated dimensions differed depending on the category of provider that is selected. For most providers, the dimensions were appointments, accommodation, employees, listening, information, and treatment. Qualitative data was gathered using a textbox where patients could elaborate on their ratings and add other relevant comments as well as the condition they were treated for. No further personal information about respondents was requested, but time stamps and email addresses were registered. Multimedia Appendix 1 shows a screenshot (Dutch) of the rating form from the Zorgkaart Nederland website (Figure A1.1). The Zorgkaart Nederland staff checked each submission for repeated entries, integrality, and anomalies, and gave an identifier to each one.

Regions

Ratings and providers were clustered at the regional level using nine PM initiatives' zip codes. These nine initiatives were selected in 2013 by the Dutch Minister of Health and are specified geographical areas in which different organizations cooperate to achieve the Triple Aim. They are spread out across the Netherlands and around 2 million people live in these regions in total. The Dutch National Institute for Public Health and the Environment was assigned their evaluation and set up the National Population Management Monitor for this purpose [13].

XSL•FO
RenderX

Preprocessing

An Excel file was provided by the Dutch Patient Federation (DPF), meaning no Web scraping or duplicate removal was necessary. The dataset is available from the DPF upon request. Mean ratings were calculated for each entry by averaging the six ratings provided. This combination was proven to provide an approximation of overall quality of care for an entry [19]. This mean rating was also used to assign a sentiment to each rating based on the Net Promotor Score (NPS). This is an instrument that determines consumer loyalty and whether a consumer is a promotor or a detractor for a company; sentiments are scored as follows: <6.5=negative, ≥6.5 and <8.5=neutral, and ≥8.5=positive [20]. Furthermore, providers in the Zorgkaart data were grouped into the following categories: hospital care, nursing home, general practitioner, insurer, birth care, pharmacy, physiotherapy, youth care, dental care, and *other* (see Multimedia Appendix 1, Table A1.1).

Text comments were transformed into a so-called "bag-of-words," which is required by the analyses described below. "Bag-of-words" means that any grammar, including punctuation, numbers, and capitalization, as well as word orders are removed from the text [21]. When the grammar is stripped away from a comment, that comment is then transformed into a matrix. In this matrix, each word (ie, unigram) or combination of two words (ie, bigram) is its own column. The rows are then filled with the number of times a word appears in that particular comment. This is done for all comments and creates a large matrix in which all comments and words in the whole dataset appear individually on the rows and columns. To tailor bigrams and prevent some word combinations from appearing positive, the previous words were evaluated and added if there were words such as "not" ("niet" in Dutch). For example "taken seriously" becomes "not taken seriously," essentially creating a trigram in these cases. To further prepare the dataset, stop words (eg, "and," "the," and "with") were removed. Words with a sparsity above 99% were also removed; this meant that these words only appeared in 1% or less of the comments, as it was expected that these words would not appear enough to be relevant for analyses. The dataset was transformed into a long or wide form, depending on the needs of the analyses. Finally, sentiment was established using two methods. First, the mean rating belonging to a comment was used to establish a positive, neutral, or negative sentiment to that comment (row). These categories were based on the NPS, as described above. Second, a Dutch lexicon was used to assign a polarity to each word (column) individually in the dataset [22]. The polarity in this lexicon ranged from -1 (the most negative connotation) to +1 (the most positive connotation).

Analyses

Frequencies of both unigrams and bigrams were determined for each of the rating sentiment categories using the complete dataset and then by PM initiative and provider category. Output was further tailored by excluding unigrams and bigrams that did not provide insight into the reasoning behind the rating, including terms such as "bad," "very good," or "not satisfied." This provided an overview of the most-used words or combination of words in each category. Next, the word polarity was averaged for all words in each PM initiative, which was compared to the average quantitative rating in the same initiative. The quantitative ratings have been studied in a previous study [14]. A linear regression analysis was added to determine if there was any correlation between the mean polarity and rating of each initiative.

In order to determine which words were the largest predictors of a positive, neutral, or negative rating, as determined by the NPS, supervised machine learning was used. Determining the most important predictors can provide insight into the reasoning of patients behind a rating: in other words, what patients value the most when providing a positive rating and what they dislike when they give a negative rating. The specific machine learning techniques used in this study are called naïve Bayes and random forest. The algorithms were run using the caret package in RStudio, version 1.1.383 (RStudio, Inc) [23]. Naïve Bayes is a fast method that performs well with a lot of dimensions and often performs similarly to other more complex methods [24]. A naïve Bayes model tries to predict, based on the words in a comment, the sentiment of a comment. It can be positive, neutral, or negative: the so-called classes. A naïve Bayesian classifier is based on the theorem of Bayes, in which predictors (ie, words) are assumed to be independent (ie, conditional independence); this theorem provides a method to calculate the posterior probability. The model uses this probability to predict the class (ie, sentiment). The dataset was randomly split between a training (50%) and a test (50%) dataset. The training set was used to train the models, while the test set was used to test the created models. Testing is done on an *unseen* set to prevent overfitting. The same test and training datasets were used in the random forest models. The algorithm creates multiple classification trees using a different bootstrap sample of the data. At each node of the tree, it chooses the best predictor out of a random subset of predictors [25]. The random forest model is known for its accuracy [26].Two models were run with each algorithm: one model aimed to predict the rating's sentiment using the words in the comment; a second model tried to predict from which PM initiative a rating originated to determine if word use was different between initiatives. The goal was to see whether it was possible to identify unique strengths and/or weaknesses of PM initiatives. In the second model, the PM initiatives could be considered the classes. Using these outcomes, it was possible to determine the most important predictors of rating sentiment and of differences between initiatives, which could indicate what patients value or miss the most. The models were evaluated using the accuracy metric and confusion matrices. A confusion matrix shows how many predictions a model got right and wrong in each of the categories [27].

Trial Registration

The Medical Research Involving Human Subjects Act (WMO) does not apply to this study, therefore, official approval was not required [28]. Participants agreed to the terms of service of Zorgkaart Nederland, which state that their submissions can be used anonymously for research purposes [29].

XSL•FO
RenderX

Results

Dataset

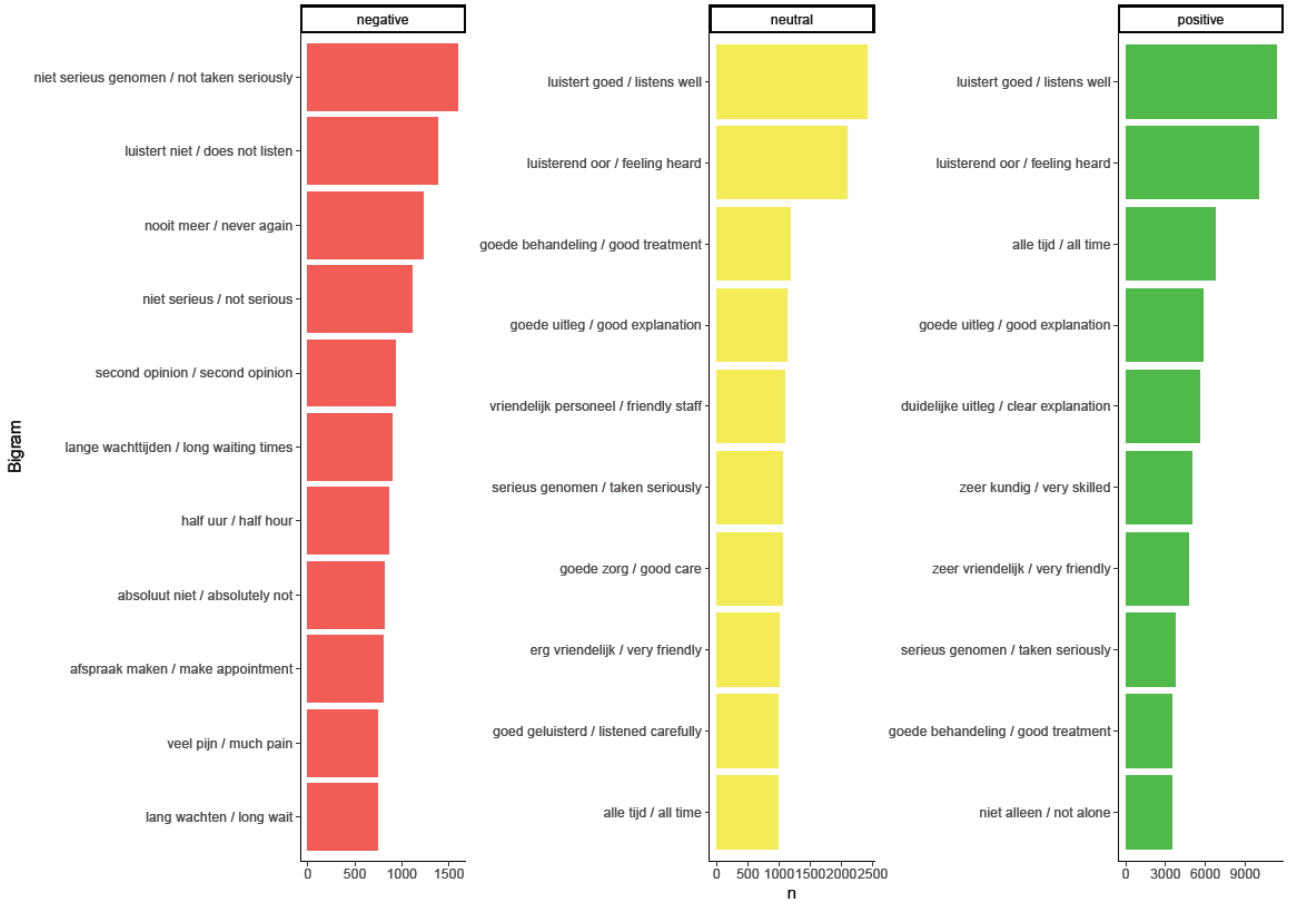
In total, 449,261 unsolicited ratings were available in the Zorgkaart database, coming from all providers in the Netherlands. These were given by 208,047 unique identifiers. Of these unsolicited ratings, 31,260 identifiers gave 70,800 ratings to providers in the PM initiatives (see Multimedia Appendix 1, Table A1.1). Of the 25,616 Dutch care providers that received at least a single rating in Zorgkaart, 4100 were located in one of the nine initiatives. The number of ratings within initiatives differed substantially, ranging from 1451 in the Vitaal Vechtdal region to 17,953 in the Slimmer met Zorg (SmZ) region (see Multimedia Appendix 1, Table A1.1). Each rating was accompanied by a comment.

Sentiment

Based on the NPS, there were 303,930 positive ratings, 97,739 neutral ratings, and 47,592 negative ratings. This illustrates that patients were generally positive about the care they received. Unigrams did not give real insight into the reasoning behind ratings; words like “very,” “good,” “treatment,” and “satisfied” were very dominant. The unigrams are, therefore, not shown in the results. Before tailoring, many bigrams did not provide insight into the reasoning. For example, in the comments of neutral and positive ratings, most bigrams were related to general satisfaction with the service, for example, “very satisfied” and “very good.” Bigrams such as the following were, therefore, excluded: “very bad,” “very good,” “very satisfied,” “bad experience,” “good experience,” “helped well,” “very nice,” “takes all,” “not again,” “not good,” “totally not,” “just only,” “totally not,” “not really,” and “a lot.”

The most-used bigrams after tailoring are shown in Figure 1. Negative bigrams were focused on listening and feeling like patients were being heard. The dominant term here was “not taken seriously.” Other bigrams within the negative ratings were mostly related to listening, waiting times, and not being satisfied with the treatment or diagnosis. Bigrams in the neutral and positive sentiment categories were similar and focused on being heard and kindness. These patterns were also seen when bigrams were split up by PM initiative (see Multimedia Appendix 2), with some exceptions. Positive ratings illustrated kindness in some and skill in other PM initiatives as the main topics, while negative ratings overall were mostly related to incorrect diagnoses and long waiting times. Standouts within the negative ratings include the region Gezonde Zorg, Gezonde Regio (ie, Healthy Care, Healthy Region), which mentioned a specific insurance company, and the mentioning of specific physicians by name in the Blauwe Zorg (ie, Blue Care) region.

Figure 1. Most-used bigrams per the Net Promotor Score (NPS) category overall. The Dutch versions of the bigrams are listed to the left of the English translations.



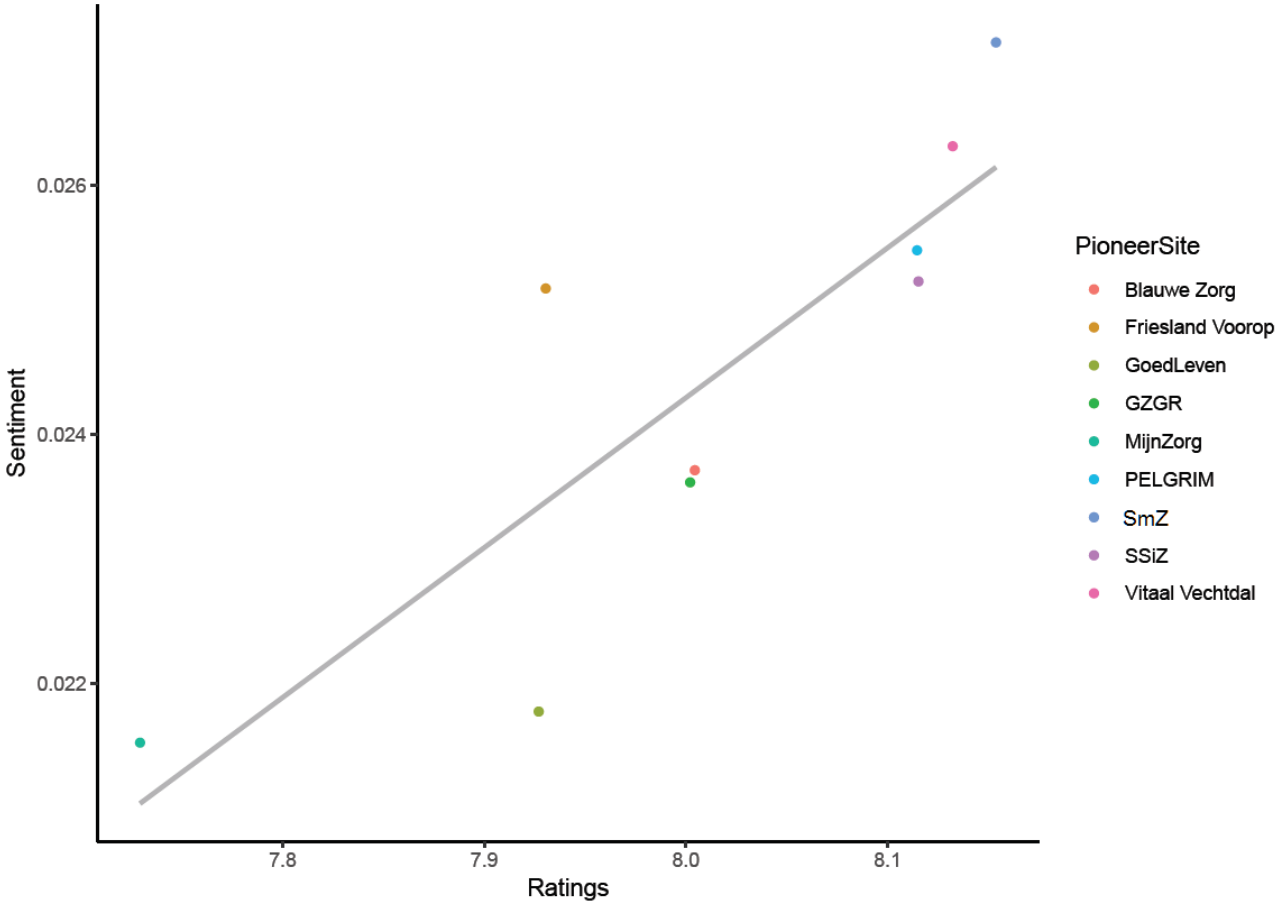
When splitting up the dataset by provider type (see Multimedia Appendix 3), it becomes clear that different aspects mattered for different providers. For example, the amount of personnel was very important in nursing homes and was often considered

XSL•FO
RenderX

in negative ratings, while the guidance by care providers was often considered a positive aspect of birth care.

Both sentiment polarity and rating did not show a large range when averaged by PM initiative. Comparing the mean sentiment with the mean ratings showed a strong positive correlation (see Figure 2). A higher mean rating in a PM initiative indicated that the sentiment was actually better within that initiative.

Figure 2. Correlation between ratings and sentiment with linear regression (r=.85): GZGR: Gezonde Zorg, Gezonde Regio; SSiZ: Samen Sterk in Zorg; SmZ: Slimmer met Zorg.



Machine Learning

Both the naïve Bayes and the random forest analyses were performed, but the results of the random forest are reported in the text, as these showed better results. The results of the naïve Bayes can be seen in Multimedia Appendix 4. Table 1 shows the confusion matrix of the sentiment model with the positive, neutral, and negative classes. The model was able to identify positive and negative ratings as such, but struggled with neutral ratings. Almost all neutral ratings were mistaken as positive ratings.

The words that had the biggest influence, including "satisfied," "good," "very," and "fine," were hard to interpret and were, therefore, not shown. Most words were simply similar to "good" or "bad," but words related to employees seemed to be additional influential factors. They did not provide a clear indication of what patients value the most in each sentiment category. The results of the naïve Bayes analysis were similar (see Multimedia Appendix 4, Table A4.1). Bigrams were not tested as predictors, as their numbers were insufficient.

Table 1. Confusion matrix of sentiment analyses using random forest machine learning.

Prediction	Actual sentiment[a], %		
	Negative	Neutral	Positive
Negative	49.2	9.7	2.7
Neutral	23.2	0.1	1.5
Positive	27.6	90.2	95.8
Total	100	100	100

[a]Accuracy=0.696.

XSL•FO
RenderX

Most ratings were positive, creating an imbalanced dataset. Additionally, the NPS scores we used could have influenced the results. Therefore, a sensitivity analysis was performed with a balanced dataset and only two sentiments: negative and positive. All ratings below 7.5 were considered to be negative, which yielded 98,974 results. An equal number of positive ratings were selected at random to create a balanced dataset. Other aspects of the analyses were kept identical to the previous analyses. This analysis showed that the accuracy improves drastically when using two sentiment categories, even when balancing them (see Table 2). The naïve Bayes variant of this analysis showed similar classifications but had worse accuracy (see Multimedia Appendix 4, Table A4.2).

Table 2. Confusion matrix of balanced sentiment analyses using random forest machine learning.

Prediction	Actual sentiment[a], %	
	Negative	Positive
Negative	83.2	20.0
Positive	16.8	80.0
Total	100	100

[a]Accuracy=0.816.

The model attempting to predict PM initiatives based on comments was not as successful; the accuracy was low (0.26). Almost all ratings were classified as either PELGRIM or SmZ, which were the PM initiatives with the most ratings. However, nine categories are a lot to predict. To fine-tune the analysis, it was repeated with only the three-largest PM initiatives. The accuracy did increase (see Table 3), also due to the reduction in categories, but ratings were still mostly classified as PELGRIM and SmZ. This indicates that the model was not able to distinguish between the different PM initiatives and that the reasoning behind ratings was similar in each. Similar results were shown by the naïve Bayes analysis (see Multimedia Appendix 4, Table A4.3).

Table 3. Confusion matrix of the largest population management (PM) initiatives analyses using random forest machine learning.

Prediction	Actual PM initiative[a], %		
	Friesland Voorop	PELGRIM	Slimmer met Zorg (SmZ)
Friesland Voorop	17.5	10.1	9.9
PELGRIM	30.6	41.4	27.2
Slimmer met Zorg (SmZ)	51.9	48.5	62.9
Total	100	100	100

[a]Accuracy=0.439.

Discussion

Principal Findings

This study explored the addition of comments accompanying unsolicited online ratings to regional analyses. The goal was to create additional insight for each PM initiative as well as for overall comparisons between initiatives by attempting to determine the reasoning and rationale behind a rating. A large online dataset provided by Zorgkaart Nederland, part of the DPF, was analyzed using sentiment analyses and machine learning techniques (naïve Bayes). Sentiment analyses illustrated that bigrams (ie, two-word combinations) proved to be more interpretable than unigrams (ie, single words). Feeling like not being “taken seriously” was the dominant bigram in negative ratings, while positive ratings mentioned mostly kindness and perceived knowledge. Comparing bigrams between PM initiatives showed a lot of overlap, but some small differences were present as well. When sentiments were quantified using a Dutch lexicon [22] and then by simply averaging the polarity of the words used, a strong correlation was found with the actual ratings. The machine learning models were able to identify sentiments of comments, especially the negative and positive comments. However, predictors did not give any meaningful insights into the underlying reasoning. When the second model tried to assign comments to PM initiatives, it could not distinguish between initiatives. This indicated that there was no clear difference in word use between initiatives.

The sentiment analyses showed that, when taken as a whole, the studied PM initiatives had mostly the same positive and negative aspects. Most ratings were positive and related to a kind and responsive staff, while negative ratings focused on being taken seriously, long waiting times, and misdiagnoses. These observations have been seen in the past in both solicited surveys [30] and interviews [31] and can be very useful for all PM initiatives, as they suggest that to get a positive rating, intangible aspects are important. Despite the amount of overlap between PM initiatives, some standout words are worth mentioning. For example, a specific care provider was mentioned often in negative ratings in a specific region. This very detailed information could prove to be very valuable for PM initiatives, as this could be used as a signal for further investigation.

XSL•FO
RenderX

Limitations

The Zorgkaart data has to be interpreted with the inherent limitations of most online datasets in mind. The data are anonymous, making it impossible to correct for potential confounders, such as age, sex, and social economic status; thus, it is impossible to correct for selection bias. It is, for example, known that a younger, more tech-savvy population tends to provide online ratings [32]. Text analysis methods also often require vast amounts of data, which were not available for each of the studied initiatives. For example, this number of comments was insufficient for the use of trigrams (ie, three-word combinations). However, as the Zorgkaart dataset shows, it is growing faster each year and this issue should resolve itself over time. Additionally, it may be possible to combine text data from different sources to increase the amount of data. The machine learning results, combined with the polarity analyses, suggest that this is possible. For example, Twitter, Facebook, and Zorgkaart data in a region could be aggregated to strengthen sentiment analyses and comparisons.

Future Research

As mentioned, online data have many benefits compared to other types of data. The biggest perk is probably the ability to monitor data close to real time. Policy makers and researchers often have to wait for survey or claims results regarding the output of an intervention; leveraging the strengths of online data could help here. In this study, the unsolicited online data have also shown that results in many regards are often similar to the results obtained from other sources. One next step for PM could, therefore, be to create a ratings dashboard for a region that keeps up with ratings and comments given in that region. It could show simple word clouds or frequencies or more advanced real-time results using machine learning; it should also be combined with more objective quality measures (eg, readmissions). This could give policy makers and researchers a more up-to-date idea of progress and might give them the opportunity to more quickly address any issues that could arise.

Conclusions

Adding information from text comments that accompany online ratings to regional evaluations provides insight for PM initiatives into the underlying reasons for the ratings. Text comments provide useful overarching information for health care policy but due to a lot of overlap, there is only limited specific information for regional policy. Specific outliers for some initiatives are insightful but comparing PM initiatives remains difficult.

Acknowledgments

This study was funded by The National Institute of Public Health and the Environment in the Netherlands (project: S/133002). The funder had no role in the design of the study; in the collection, analysis, and interpretation of the data; or in writing of the manuscript.

Authors' Contributions

All authors were responsible for the concept and design of the study. All authors were responsible for the acquisition, analysis, or interpretation of data. RJPH was responsible for drafting of the manuscript. All authors were responsible for the critical revision of the manuscript for important intellectual content. HWTD, MS, DR, and CB were responsible for study supervision. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of Zorgkaart Nederland dataset.
[DOCX File , 190 KB - medinform_v7i4e13053_app1.docx]

Multimedia Appendix 2
Most-used words (ie, unigrams and bigrams) per population management (PM) initiative.
[DOCX File , 389 KB - medinform_v7i4e13053_app2.docx]

Multimedia Appendix 3
Most-used bigrams per provider category.
[DOCX File , 430 KB - medinform_v7i4e13053_app3.docx]

Multimedia Appendix 4
Additional results of the machine learning analyses.
[DOCX File , 15 KB - medinform_v7i4e13053_app4.docx]

References

XSL•FO
RenderX

1. Ranard BL, Werner RM, Antanavicius T, Schwartz HA, Smith RJ, Meisel ZF, et al. Yelp reviews Of hospital care can supplement and inform traditional surveys of the patient experience of care. Health Aff (Millwood) 2016 Apr;35(4):697-705 [FREE Full text] [doi: 10.1377/hlthaff.2015.1030] [Medline: 27044971]
2. Greaves F, Pape UJ, King D, Darzi A, Majeed A, Wachter RM, et al. Associations between Internet-based patient ratings and conventional surveys of patient experience in the English NHS: An observational study. BMJ Qual Saf 2012 Jul;21(7):600-605. [doi: 10.1136/bmjqs-2012-000906] [Medline: 22523318]
3. Griffiths A, Leaver MP. Wisdom of patients: Predicting the quality of care using aggregated patient feedback. BMJ Qual Saf 2018 Feb;27(2):110-118 [FREE Full text] [doi: 10.1136/bmjqs-2017-006847] [Medline: 28971881]
4. Hawkins JB, Brownstein JS, Tuli G, Runels T, Broecker K, Nsoesie EO, et al. Measuring patient-perceived quality of care in US hospitals using Twitter. BMJ Qual Saf 2016 Jun;25(6):404-413 [FREE Full text] [doi: 10.1136/bmjqs-2015-004309] [Medline: 26464518]
5. Stiefel M, Nolan K. Measuring the triple aim: A call for action. Popul Health Manag 2013 Aug;16(4):219-220. [doi: 10.1089/pop.2013.0025] [Medline: 23941047]
6. Kindig D, Isham G, Siemering KQ. The business role in improving health: Beyond social responsibility. NAM Perspect 2013 Aug 08;3(8):1. [doi: 10.31478/201308b]
7. Steenkamer BM, Drewes HW, Heijink R, Baan CA, Struijs JN. Defining population health management: A scoping review of the literature. Popul Health Manag 2017 Feb;20(1):74-85. [doi: 10.1089/pop.2015.0149] [Medline: 27124406]
8. Struijs JN, Drewes HW, Heijink R, Baan CA. How to evaluate population management? Transforming the Care Continuum Alliance population health guide toward a broadly applicable analytical framework. Health Policy 2015 Apr;119(4):522-529. [doi: 10.1016/j.healthpol.2014.12.003] [Medline: 25516015]
9. Alderwick H, Ham C, Buck D. Population Health Systems: Going Beyond Integrated Care. London, UK: The King's Fund; 2015 Feb. URL: https://www.kingsfund.org.uk/sites/default/files/field/field_publication_file/population-health-systems-kingsfund-feb15.pdf [accessed 2018-09-24]
10. Berwick DM, Nolan TW, Whittington J. The triple aim: Care, health, and cost. Health Aff (Millwood) 2008;27(3):759-769. [doi: 10.1377/hlthaff.27.3.759] [Medline: 18474969]
11. Lewis VA, Colla CH, Carluzzo KL, Kler SE, Fisher ES. Accountable Care Organizations in the United States: Market and demographic factors associated with formation. Health Serv Res 2013 Dec;48(6 Pt 1):1840-1858 [FREE Full text] [doi: 10.1111/1475-6773.12102] [Medline: 24117222]
12. New Care Models: Vanguards—Developing a Blueprint for the Future of NHS and Care Services. London, UK: NHS England; 2016. URL: https://www.england.nhs.uk/wp-content/uploads/2015/11/new_care_models.pdf [accessed 2018-05-06]
13. Drewes HW, Heijink R, Struijs JN, Baan C. Landelijke Monitor Populatiemanagement. Deel 1: Beschrijving Proeftuinen. Bilthoven, the Netherlands: Rijksinstituut voor Volksgezondheid en Milieu (RIVM); 2014. URL: http://www.invoorzorg.nl/docs/ivz/bedrijfsvoering/Landelijke_monitor_populatiemanagement_RIVM.pdf [accessed 2019-11-18]
14. Hendrikx RJP, Spreeuwenberg MD, Drewes HW, Struijs JN, Ruwaard D, Baan CA. Harvesting the wisdom of the crowd: Using online ratings to explore care experiences in regions. BMC Health Serv Res 2018 Oct 20;18(1):801 [FREE Full text] [doi: 10.1186/s12913-018-3566-z] [Medline: 30342518]
15. Emmert M, Meszmer N, Schlesinger M. A cross-sectional study assessing the association between online ratings and clinical quality of care measures for US hospitals: Results from an observational study. BMC Health Serv Res 2018 Feb 05;18(1):82 [FREE Full text] [doi: 10.1186/s12913-018-2886-3] [Medline: 29402321]
16. Qu Z, Zhang H, Li H. Determinants of online merchant rating: Content analysis of consumer comments about Yahoo merchants. Decis Support Syst 2008 Dec;46(1):440-449 [FREE Full text] [doi: 10.1016/j.dss.2008.08.004]
17. Emmert M, Meier F, Heider A, Dürr C, Sander U. What do patients say about their physicians? An analysis of 3000 narrative comments posted on a German physician rating website. Health Policy 2014 Oct;118(1):66-73. [doi: 10.1016/j.healthpol.2014.04.015] [Medline: 24836021]
18. Zorgkaart Nederland. URL: https://www.zorgkaartnederland.nl/ [accessed 2018-12-06]
19. Krol MW, de Boer D, Rademakers JJ, Delnoij DM. Overall scores as an alternative to global ratings in patient experience surveys: A comparison of four methods. BMC Health Serv Res 2013 Nov 19;13:479 [FREE Full text] [doi: 10.1186/1472-6963-13-479] [Medline: 24245726]
20. Markey R, Reichheld F. Bain & Company. 2011 Dec 08. Introducing: The Net Promoter System®. URL: https://www.bain.com/insights/introducing-the-net-promoter-system-loyalty-insights/ [accessed 2019-11-18]
21. Radovanovic M, Ivanovic M. Text mining: Approaches and applications. Novi Sad J Math 2008 Jan;38(3):227-234 [FREE Full text]
22. Jijkoun V, Hofmann K. Generating a non-English subjectivity lexicon: Relations that matter. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. 2009 Presented at: 12th Conference of the European Chapter of the Association for Computational Linguistics; March 30-April 3, 2009; Athens, Greece. [doi: 10.3115/1609067.1609111]
23. Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008 Nov;28(5):1-26 [FREE Full text] [doi: 10.18637/jss.v028.i05]

XSL•FO
RenderX

24. Zhang H. The optimality of naive Bayes. In: Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference. 2004 Presented at: 17th International Florida Artificial Intelligence Research Society Conference; May 12-14, 2004; Miami Beach, FL.
25. Liaw A, Wiener M. Classification and regression by randomForest. R News 2002 Dec;2(3):18-22 [FREE Full text]
26. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 2014;15:3133-3181 [FREE Full text]
27. Ting KM. Confusion matrix. In: Sammut C, Webb G, editors. Encyclopedia of Machine Learning. Boston, MA: Springer; 2011.
28. Overheid Wettenbank. Wet medisch-wetenschappelijk onderzoek met mensen. URL: http://wetten.overheid.nl/BWBR0009408/2018-08-01 [accessed 2018-09-24]
29. Zorgkaart Nederland. Utrecht, the Netherlands: Patiëntenfederatie Nederland; 2019 Feb 07. Privacyverklaring. URL: https://www.zorgkaartnederland.nl/content/privacyverklaring [accessed 2018-09-24]
30. Rathert C, Williams ES, McCaughey D, Ishqaidef G. Patient perceptions of patient-centred care: Empirical test of a theoretical model. Health Expect 2015 Apr;18(2):199-209 [FREE Full text] [doi: 10.1111/hex.12020] [Medline: 23176054]
31. Nelson JE, Puntillo KA, Pronovost PJ, Walker AS, McAdam JL, Ilaoa D, et al. In their own words: Patients and families define high-quality palliative care in the intensive care unit. Crit Care Med 2010 Mar;38(3):808-818 [FREE Full text] [doi: 10.1097/ccm.0b013e3181c5887c] [Medline: 20198726]
32. Couper MP, Kapteyn A, Schonlau M, Winter J. Noncoverage and nonresponse in an Internet survey. Soc Sci Res 2007 Mar;36(1):131-148. [doi: 10.1016/j.ssresearch.2005.10.002]

Abbreviations

DPF: Dutch Patient Federation
NPS: Net Promotor Score
PM: population management
SmZ: Slimmer met Zorg
WMO: Medical Research Involving Human Subjects Act

Edited by G Eysenbach; submitted 07.12.18; peer-reviewed by C Fincham, R Sadeghi, B Ranard; comments to author 09.04.19; revised version received 22.08.19; accepted 26.09.19; published 16.12.19.

Please cite as:
Hendrikx RJP, Drewes HWT, Spreeuwenberg M, Ruwaard D, Baan C
Measuring Regional Quality of Health Care Using Unsolicited Online Data: Text Analysis Study
JMIR Med Inform 2019;7(4):e13053
URL: http://medinform.jmir.org/2019/4/e13053/
doi:10.2196/13053
PMID:31841116

©Roy Johannus Petrus Hendrikx, Hanneke Wil-Trees Drewes, Marieke Spreeuwenberg, Dirk Ruwaard, Caroline Baan. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 16.12.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Interpretability and Class Imbalance in Prediction Models for Pain Volatility in Manage My Pain App Users: Analysis Using Feature Selection and Majority Voting Methods

Quazi Abidur Rahman[1,2], PhD; Tahir Janmohamed[3], PEng, MBA; Hance Clarke[4], MD, PhD, FRCPC; Paul Ritvo[5], PhD; Jane Heffernan[1], PhD; Joel Katz[4,5,6], PhD

[1]Department of Computer Science, Lakehead University, Thunder Bay, ON, Canada

[2]Centre for Disease Modelling, Department of Mathematics and Statistics, York University, Toronto, ON, Canada

[3]ManagingLife, Toronto, ON, Canada

[4]Department of Anesthesia and Pain Management, Toronto General Hospital, Toronto, ON, Canada

[5]Department of Psychology, York University, Toronto, ON, Canada

[6]School of Kinesiology & Health Science, York University, Toronto, ON, Canada

Corresponding Author:
Quazi Abidur Rahman, PhD
Department of Computer Science
Lakehead University
955 Oliver Road
Thunder Bay, ON
Canada
Phone: 1 (807) 346 7789
Email: quazi.rahman@lakeheadu.ca

Abstract

Background: Pain volatility is an important factor in chronic pain experience and adaptation. Previously, we employed machine-learning methods to define and predict pain volatility levels from users of the Manage My Pain app. Reducing the number of features is important to help increase interpretability of such prediction models. Prediction results also need to be consolidated from multiple random subsamples to address the class imbalance issue.

Objective: This study aimed to: (1) increase the interpretability of previously developed pain volatility models by identifying the most important features that distinguish high from low volatility users; and (2) consolidate prediction results from models derived from multiple random subsamples while addressing the class imbalance issue.

Methods: A total of 132 features were extracted from the first month of app use to develop machine learning–based models for predicting pain volatility at the sixth month of app use. Three feature selection methods were applied to identify features that were significantly better predictors than other members of the large features set used for developing the prediction models: (1) Gini impurity criterion; (2) information gain criterion; and (3) Boruta. We then combined the three groups of important features determined by these algorithms to produce the final list of important features. Three machine learning methods were then employed to conduct prediction experiments using the selected important features: (1) logistic regression with ridge estimators; (2) logistic regression with least absolute shrinkage and selection operator; and (3) random forests. Multiple random under-sampling of the majority class was conducted to address class imbalance in the dataset. Subsequently, a majority voting approach was employed to consolidate prediction results from these multiple subsamples. The total number of users included in this study was 879, with a total number of 391,255 pain records.

Results: A threshold of 1.6 was established using clustering methods to differentiate between 2 classes: low volatility (n=694) and high volatility (n=185). The overall prediction accuracy is approximately 70% for both random forests and logistic regression models when using 132 features. Overall, 9 important features were identified using 3 feature selection methods. Of these 9 features, 2 are from the app use category and the other 7 are related to pain statistics. After consolidating models that were developed using random subsamples by majority voting, logistic regression models performed equally well using 132 or 9 features. Random forests performed better than logistic regression methods in predicting the high volatility class. The consolidated accuracy of random forests does not drop significantly (601/879; 68.4% vs 618/879; 70.3%) when only 9 important features are included in the prediction model.

XSL•FO
RenderX

Conclusions: We employed feature selection methods to identify important features in predicting future pain volatility. To address class imbalance, we consolidated models that were developed using multiple random subsamples by majority voting. Reducing the number of features did not result in a significant decrease in the consolidated prediction accuracy.

(JMIR Med Inform 2019;7(4):e15601) doi:10.2196/15601

KEYWORDS

chronic pain; pain volatility; data mining; cluster analysis; machine learning; prediction model; Manage My Pain; pain app

Introduction

Background

Pain is one of the most prevalent health-related concerns and is among the top three most frequent reasons for seeking medical help [1]. Mobile pain apps are transforming how people monitor, manage, and communicate pain-related information [2], and scientific publications on methods and results can help both consumers and health care professionals select the right app to support their treatment plans. Moreover, appropriate analyses can provide valuable insights into pain experiences over long-term periods. We previously conducted two studies [3,4] using data from a pain management app called Manage My Pain [5], where data mining and machine learning methods were employed to understand app usage patterns and define and predict pain volatility. In the first study [3], we divided users into five clusters based on their level of engagement with the app and then applied statistical methods to characterize each user cluster using six different attributes (eg, gender, age, number of pain conditions, number of medications, pain severity, and opioid use).

In the more recent study [4], we developed prediction models to identify and predict groups of users who reported improvements or decrements in their pain levels. To facilitate the development of these models, we addressed the important issue of identifying the most appropriate statistic to use when measuring pain severity over time. We proposed a measure of volatile change that captures fluctuation or variability in pain scores over time. Pain volatility is an important contributor to pain experience for people with chronic pain, particularly because of its linkage with the initiation of opioid addiction [6,7]. Moreover, pain perception and consequent disability are heightened under conditions of greater uncertainty and unpredictability [8], and greater pain volatility is one contributor to uncertainty and unpredictability. Being able to predict future pain volatility can assist patient awareness and application of self-management and appropriate medication use. We defined pain volatility as the mean of absolute changes between 2 consecutive self-reported pain severity ratings (0-10 numeric rating scale). We applied clustering methods to divide users into two classes based on their pain volatility levels: low volatility and high volatility. We then employed four machine learning methods to predict users' pain volatility level at the sixth month of app use. We developed prediction models where information related to user demographics, pain, medications, and app engagement from the first month's app use were extracted as features. The total number of features in the prediction models was 130. Prediction models, trained using random forests, performed the best, with the accuracy for the low and the high volatility class reasonably high.

One major drawback of using random forests and similar black-box methods is the lack of interpretability of the trained models. This is especially true when the application domain is medicine and the set of features is large. An interpretable prediction model should incorporate a way to identify a subset of important features that are significantly better predictors than other members of the large features set. This provides health care providers with a practical model to apply and may result in increased confidence in the model. Moreover, the important features may help health care professionals and patients develop appropriate interventions and pain management plans for the future.

While developing volatility prediction models, we addressed the issue of class imbalance because the number of low volatility users was much higher than that of high volatility users. We employed random under-sampling methods to make two equal class sizes. We repeated this under-sampling procedure three times to ensure the stability of the results. Because we did not consolidate the result of these multiple models trained on random subsamples into a single unified model in previous work, we were intent on consolidation in this research.

Objectives

Accordingly, the present study has two objectives. The first was to identify important features in the pain volatility prediction models that have significantly higher predictive capability than other features. We used two criteria to rank features based on their importance: Gini impurity and information gain. We also applied the Boruta feature selection algorithm to identify a subset of important features. The second objective was to consolidate prediction results from models trained on multiple random subsamples of data where the number of users in the low and high volatility classes was equalized. We employed the majority voting approach to achieve this. Training and testing were conducted using standard 5-fold cross validation. Accuracy for the low and high volatility classes and overall accuracy were calculated to measure and compare the performance of individual and consolidated prediction models developed.

Methods

Manage My Pain

Manage My Pain [5], developed by ManagingLife, helps people living with pain to track their pain and functioning daily using an Android or Apple smartphone app. The central feature of Manage My Pain is the pain record that enables users to enter

XSL•FO
RenderX

details about their pain experience. Each record contains only one mandatory item, a rating of pain severity using a slider on a visual analogue scale. Users have the option of completing seven more items to describe their pain experience more comprehensively. The app issues daily reminders and prompts users to reflect on their daily accomplishments through a daily reflection. Users can also add pain conditions, gender, age, and medications to their profile in the app. As of March 1, 2019, Manage My Pain had 31,700 users and 949,919 pain records.

Procedure

The present study was reviewed and approved by the Research Ethics Board at York University (Human Participants Review Committee, Certificate number: e2015-160). The user database was accessed and downloaded in two separate files (using plain text format): (1) deidentified user information; and (2) pain records. The user information file contains the following fields: user ID, age at date of app registration, gender, self-reported pain conditions and self-reported medications. The pain record file contains the following fields: user ID, date, pain severity rating (0-10), body location(s) of pain, pain type, pain duration, other associated symptoms, characteristics, relieving factors, ineffective factors, aggravating factors, and environments of pain occurrence. All fields in the text files are delimited using special characters. The data used in this study were downloaded on March 1, 2019. This study covers pain records entered by users between January 1, 2013 and March 1, 2019.

Data

The primary dataset includes 949,919 pain records from 31,700 users. The outcome period for predicting pain volatility is the sixth month of app usage. The sixth month was chosen because pain lasting at least 6 months meets most generally accepted definitions of chronic pain [9]. In the present study, as in our previous work, we used the first month as the predictor period and we thus collected features from the first month of engagement with Manage My Pain to predict pain volatility during the sixth month of Manage My Pain engagement. The mathematical minimum for calculating pain volatility is 2 pain severity records. However, to increase the reliability of prediction results, users with at least 5 pain records in both the predictor and outcome periods were required for prediction experiments in this study. The number of users in the primary dataset meeting this criterion was 879 and there were 391,255 pain records in the dataset. This is an increase of 97 users and 62,185 records over the number of users and pain records used in our previous study. These 879 users had a mean of 370.09 pain records and a median of 213 pain records.

Pain Volatility Definition and Prediction

We first briefly summarized the methods used in our previous work [4] to develop volatility prediction models. We defined pain volatility as the mean of absolute changes between two consecutive pain severity ratings within each of the two observation periods. We also applied the k-means clustering method [10] to divide users into two distinct classes (high volatility and low volatility) using a threshold on the pain volatility measure. We extracted 130 features from each user to develop prediction models. Four machine learning methods were employed to develop prediction models: (1) logistic regression with ridge estimators [11]; (2) logistic regression with least absolute shrinkage and selection operator (LASSO) [12]; (3) random forests [13]; and (4) support vector machines (SVM) [14].

The stratified 5-fold cross-validation procedure was used for training and testing. Initial experiments employing 10-fold cross-validation produced similar prediction performance. Data preprocessing was conducted in R (version 3.5.0) (R Core Team, Vienna, Austria). R package glmnet (version 2.0-16) [15] was used for training and testing logistic regression models. We applied the standard Random Forests classification package in WEKA (version 3.8) (University of Waikato, Hamilton, New Zealand) [16], using 100 trees in the Random Forests implementation. The number of features selected at random at each tree-node was set to , where n is the total number of features. For SVM implementation, we used the WEKA libsvm, employing the Gaussian radial basis function kernel.

The following three measures were used to measure prediction performance:

If we consider users in the low volatility class to be the control group in our experiments, the accuracy of the low volatility class and that of the high volatility class are Specificity and Sensitivity, respectively.

In the present study, the same methods were employed for defining and predicting pain volatility. We added 2 new features to the previous list of 130 features: (1) the standard deviation of the mean of the absolute values of changes between consecutive pain severity ratings; and (2) the absolute value of the difference in pain severity ratings between the end point and the starting point of the trend line of the ratings from the predictor and outcome periods. These two are added to complement 2 existing features: (1) the mean of the absolute values of changes between consecutive pain severity ratings; and (2) the difference in pain severity ratings between the end point and the starting point of the trend line of ratings from the predictor and outcome periods.

Thus, we extracted 132 features from each of the 879 users for developing prediction models. We divided these 132 features into 8 broad categories to facilitate discussions on results from feature selection experiments. The 8 categories are listed below:

1. Demographic (2 features)
   - Gender
   - Age
2. App usage (2 features)
   - Number of pain records (1 features)
   - Number of days with at least one pain record (1 features)
3. Pain statistics (8 features)
   - Mean and standard deviation of pain severity ratings (2 features)

XSL•FO
RenderX

- Mean and standard deviation of absolute values of changes between consecutive severity ratings (2 features)
- The difference and the absolute value of the difference between the end point and the starting point of a trend line fitted through the severity ratings (2 features)
- Pain severity level (1 feature)
- Pain volatility level (1 feature)

4. Pain descriptors (64 features)
   - Pain locations (24 features)
   - Associated symptoms (20 features)
   - Characteristics (13 features)
   - Environments (7 features)
5. Factors impacting pain (43 features)
   - Aggravating (15 features)
   - Alleviating (14 factors) and Ineffective (14 features)
6. Pain conditions (6 features)
7. Medications (5 features)
8. Mental health conditions (2 features)

Feature Selection

Summary

We applied three different methods to identify features important in predicting pain volatility: the Gini Impurity criterion, the Information Gain criterion, and Boruta.

Gini Impurity Criterion

The Gini impurity measure [13] is defined as the probability of an incorrect prediction of a random instance in a dataset, assuming it is randomly predicted according to the distribution of the outcomes in the dataset. This criterion is used while training a Random Forests prediction model [17] to help choose the best feature for splitting a node of a tree. Once a feature has been selected to split a node in a tree in the model, the Gini impurity for the descendants is less than the parent node. The importance is calculated as the difference between the parent node's impurity and the weighted sum of the children's nodes' impurity. This is averaged over the whole forest and a higher mean value indicates higher importance.

Information Gain Criterion

For each feature, the information gain [18] measures how much information is gained about the outcome when the value of the feature is obtained. It is calculated as the difference between the unconditional entropy associated with the outcome and the conditional entropy of the outcome given the value of a feature. A higher value of information gain indicates higher importance in prediction.

Boruta

Boruta [19] is a wrapper feature selection algorithm built around Random Forests. This method adds randomly permutated copies of all features to the dataset and trains a Random Forests model on this extended dataset. For each feature and its copy, an importance score (mean decreased accuracy) is calculated by the Random Forests training algorithm. A feature is identified to be important by the Boruta algorithm if its importance score is determined to be higher than the best importance score of the permutated copies through a statistical significance test. Similarly, a feature is labeled as not important if the importance is lower than the best importance score of the permutated copies by a statistically significant margin. This process is repeated iteratively until all features have been assigned an important or not important label.

In our prediction experiments, we had five different training sets as we conducted 5-fold cross validation. For each of these five training sets, we applied random under-sampling 5 times, resulting in a total of 25 different training sets. We applied each of the three described algorithms on all 25 sets and identified the features that were common across these sets. We then combined the three groups of important features determined by the three algorithms to produce the final list of important features.

Class Imbalance

After defining the low and high volatility classes using the clustering approach, the number of low volatility users is much higher than that of high volatility users (approximately 3 times), as will be detailed in the Results section. In our previous study, we addressed this class imbalance issue by repeated, random under-sampling from the majority class to create a balanced dataset for training prediction models. Under the under-sampling method, instances are chosen at random from the majority class to make the size of the two classes equal. We repeated the under-sampling procedure 3 times to ensure stability of the results.

In the present study, we repeated the under-sampling 5 times and employed a majority voting approach to consolidate the prediction results of these multiple subsamples. In this majority voting method, we assigned a user to the high volatility class when the output was predicted to be high volatility by at least 3 models trained on different subsamples.

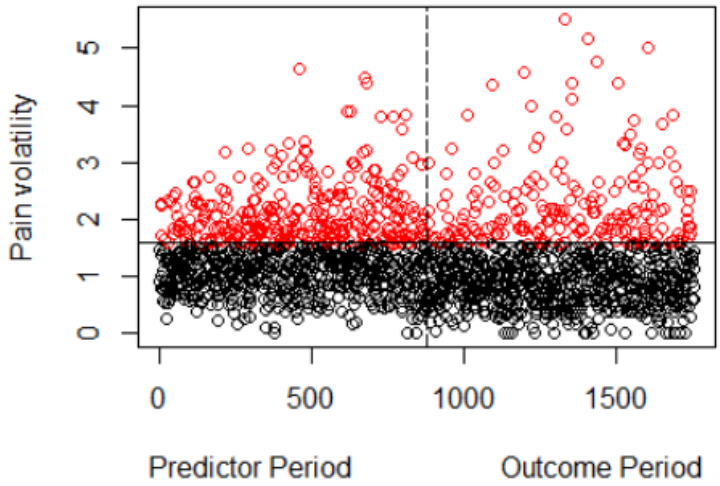
Results

Prediction Results

We first employed the k-means clustering method on the pain volatility measures to distinguish between low and high volatility classes. Figure 1 shows the clustering output. Each of 879 users has two values in the figure: one from the predictor and one from the outcome period. Low and high volatility classes are indicated by black and red colors, respectively. The numerical threshold distinguishing these two classes of users is approximately 1.6, which is the same as our previous study.

XSL•FO
RenderX

Figure 1. Clustering pain volatility measures. The total number of data points is 1758. Each user has two data points, one each from the predictor and outcome periods. Data points with index (x-axis) 1 to 879 are volatility values from the predictor period and 880 to 1758 are from the outcome period. Black and red colors indicate low and high volatility levels, respectively. The horizontal solid line shows the volatility threshold of 1.6 and the vertical dotted line indicates the cut-off between the predictor and the outcome period.



We further validated this threshold by reapplying the clustering algorithm on randomly chosen subsamples of the pain volatility values.

Using the pain volatility threshold of 1.6 resulted in the following division of users in the outcome period: 694 had low volatility and 185 had high volatility. We addressed the class imbalance issue by random under-sampling (repeating 5 times) of the majority class (ie, the low volatility class). We then developed prediction models using logistic regression with ridge estimators and LASSO, random forests, and SVM. The prediction results are shown in Table 1 and Figure 2.

Table 1. Prediction performance using all 132 features. Random under-sampling of the majority class (low volatility) was applied and repeated 5 times to make class sizes equal in the training dataset.

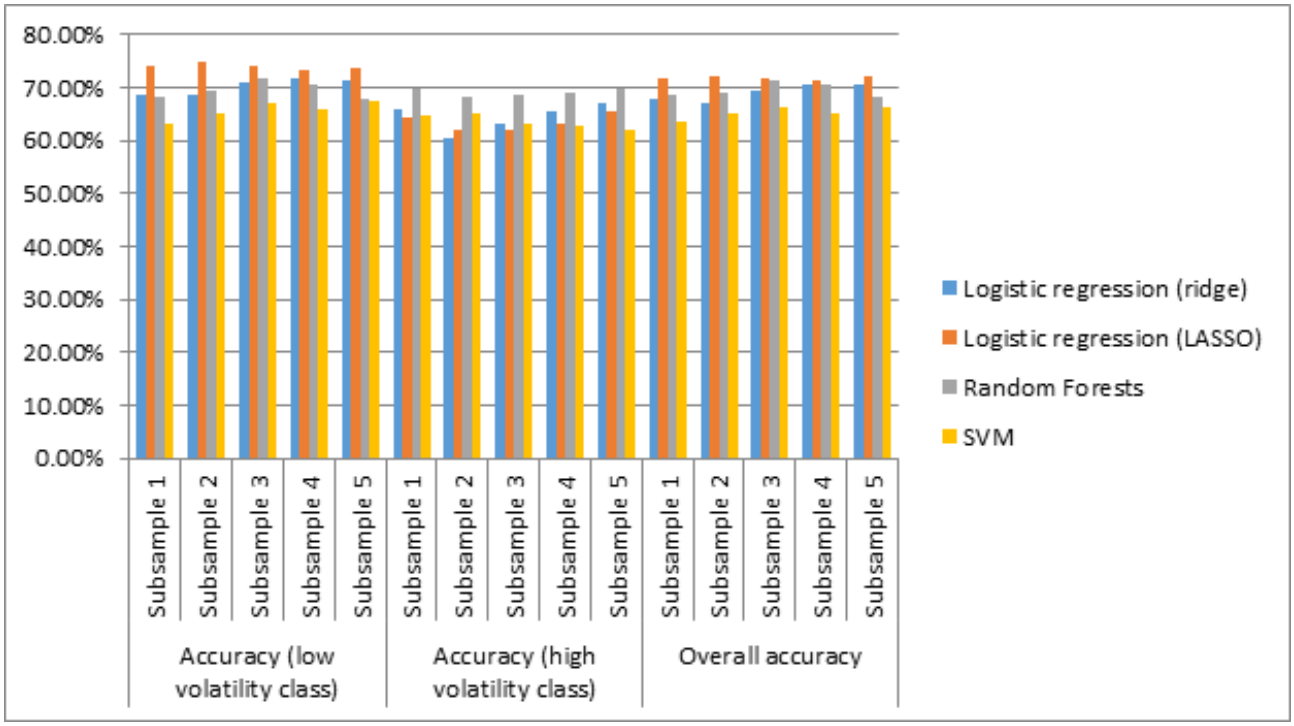
Performance measure, subsamples	Logistic regression (ridge), n (%)	Logistic regression (LASSO[a]), n (%)	Random forests, n (%)	SVM[b], n (%)
Accuracy (low volatility class; n=694)				
Subsample 1	476 (68.6)	513 (73.9)	473 (68.2)	437 (63.0)
Subsample 2	476 (68.6)	520 (74.9)	482 (69.5)	453 (65.3)
Subsample 3	492 (70.9)	514 (74.1)	499 (71.9)	465 (67.0)
Subsample 4	499 (71.9)	509 (73.3)	491 (70.7)	458 (66.0)
Subsample 5	495 (71.3)	512 (73.8)	471 (67.9)	469 (67.6)
Accuracy (high volatility class; n=185)				
Subsample 1	122 (65.9)	119 (64.3)	129 (69.7)	120 (64.9)
Subsample 2	112 (60.5)	115 (62.2)	126 (68.1)	118 (65.3)
Subsample 3	117 (63.2)	115 (62.2)	127 (68.6)	117 (63.2)
Subsample 4	121 (65.4)	117 (63.2)	128 (69.2)	116 (62.7)
Subsample 5	124 (67.0)	121 (65.4)	129 (69.7)	115 (62.2)
Overall accuracy (n=879)				
Subsample 1	598 (68.0)	632 (71.9)	602 (68.5)	557 (63.4)
Subsample 2	588 (66.9)	635 (72.2)	608 (69.2)	571 (65.0)
Subsample 3	609 (69.3)	629 (71.6)	626 (71.2)	582 (66.21)
Subsample 4	620 (70.5)	626 (71.2)	619 (70.4)	574 (65.3)
Subsample 5	619 (70.4)	633 (72.0)	600 (68.3)	584 (66.4)

[a]LASSO: least absolute shrinkage and selection operator.

[b]SVM: support vector machines.

XSL•FO
RenderX

Figure 2. Prediction performance using all 132 features (graphical representation of <xref ref-type="table" rid="31table1">Table 1</xref>). LASSO: least absolute shrinkage and selection operator; SVM: support vector machines.

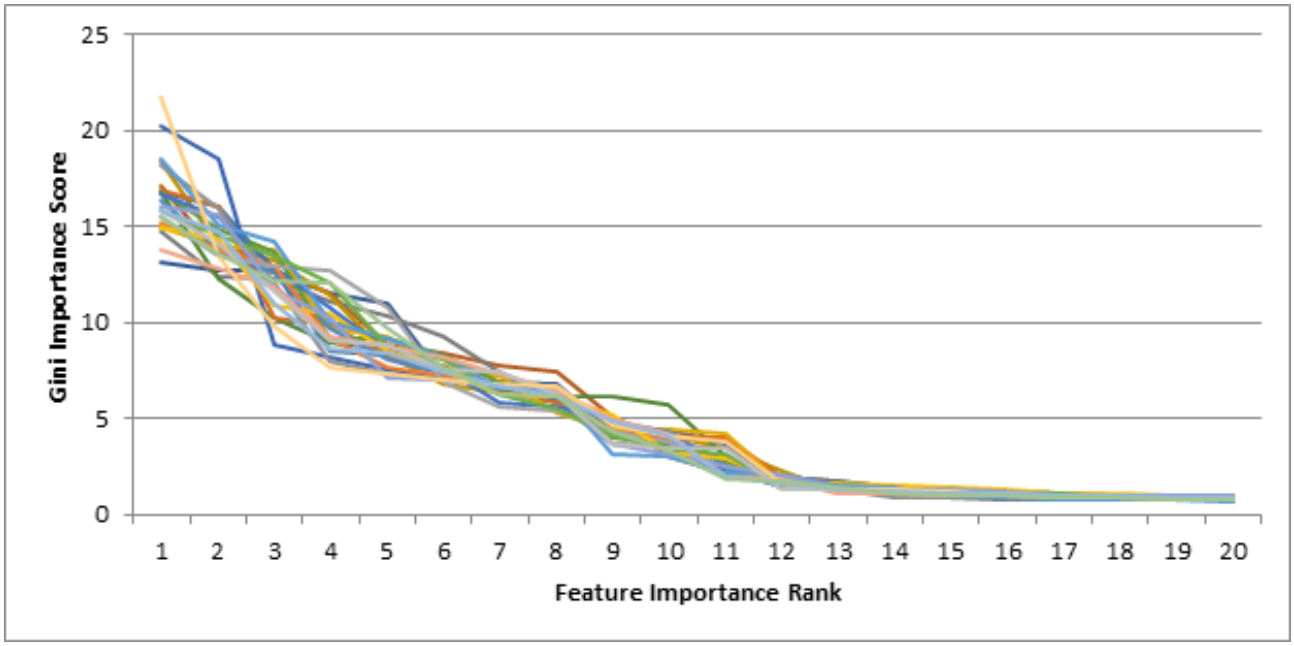


The overall accuracy is approximately 70% for both random forests and logistic regression models. Random forests consistently achieved the same accuracy in predicting both low and high volatility classes across all subsamples. However, SVM did not perform well in predicting future pain volatility levels. Therefore, SVM was not used for features selection experiments in this study.

Feature Selection Results

We first used the Gini importance criterion to identify important features in distinguishing high from low volatility users. Mean decreased Gini was calculated for each training set. In 5-fold cross validation experiments, we had 5 different training sets. As we conducted under-sampling 5 times for each training set, we eventually had 25 training sets for this study. For each of these 25 training sets, we trained models using random forests and all 132 features. We then calculated the Gini importance score for each feature to create a ranking based on importance. In Figure 3, we show this importance score for the top 20 features of all 25 training sets.

Figure 3. Ranking of importance of features according to Gini importance criterion for all 25 different training sets.



XSL•FO
RenderX

The importance of features does not decrease significantly beyond the top 11 features in all the training sets. In the list of these top 11 features, we identified 8 that were common across all training sets. They are: (1) the number of days with at least one pain record; (2) the number of pain records; (3) the mean pain severity rating; (4) the standard deviation of the pain severity ratings; (5) the mean of the absolute changes between consecutive pain ratings; (6) the standard deviation of the absolute changes between consecutive pain ratings; (7) the change between the start and end of the trend line fitted through the severity ratings; and (8) the absolute value of the change between the start and end of the trend line fitted through the severity ratings.

The second criterion that we used to calculate the importance of features was information gain. Figure 4 shows the top 20 features in all 25 training sets and the corresponding information gain values.

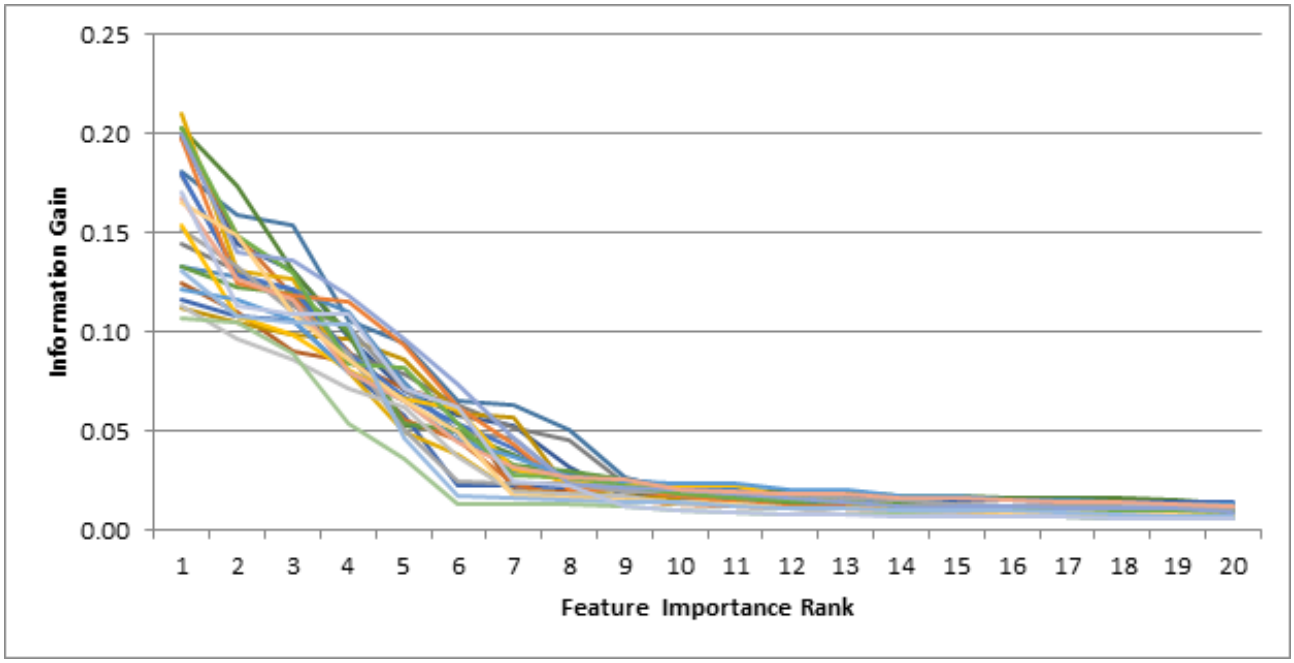
The information gain drops significantly between the sixth and the ninth feature across different training sets. The following features are the common ones among the top features as ranked by the information gain criterion: (1) the number of days with at least one pain record; (2) the standard deviation of the pain severity ratings; (3) the mean of the absolute changes between consecutive pain ratings (pain volatility scores); (4) the standard deviation of the absolute changes between consecutive pain ratings; and (5) the pain volatility levels in the predictor period.

Lastly, we applied the Boruta method to identify important features in each of the 25 training sets. The number of features considered important by this method varied between 4 and 7 across training sets, with the following 4 common in all sets: (1) the number of days with at least one pain record; (2) the standard deviation of the pain severity ratings; (3) the mean of the absolute changes between consecutive pain ratings; and (4) the standard deviation of the absolute changes between consecutive pain ratings.

Combining the features identified to be important by the three methods leads to the following 9 features: (1) the number of days with at least one pain record; (2) the number of pain records; (3) the mean pain severity rating; (4) the standard deviation of the pain severity ratings; (5) the mean of the absolute changes between consecutive pain ratings (pain volatility scores); (6) the standard deviation of the absolute changes between consecutive pain ratings; (7) the change between the start and end of the trend line fitted through the severity ratings; (8) the absolute value of the change between the start and end of the trend line fitted through the severity ratings; and (9) the pain volatility levels in the predictor period.

The first 2 features are related to app usage and the other 7 features are from the pain statistics category. We used these 9 features to develop volatility prediction models using random forests and logistic regression methods. We then applied majority voting to consolidate the five models developed using subsamples. The prediction results of the individual and consolidated models are presented in Table 2. In Table 2, random under-sampling of the majority class (low volatility) was applied and repeated 5 times to make class sizes equal in the training dataset. Although low volatility and overall accuracy is higher for logistics regression models, random forests perform better when predicting high volatility class across different random subsamples.

Figure 4. Ranking of features based on the importance calculated using information gain for all 25 different training sets.



XSL•FO
RenderX

Table 2. Prediction performance using the 9 selected important features.

Performance measure	Logistic regression (ridge), n (%)	Logistic regression (LASSO[a]), n (%)	Random forests, n (%)
Accuracy (low volatility class; n=694)			
Subsample 1	510 (73.5)	513 (73.9)	461 (67.4)
Subsample 2	518 (74.6)	516 (74.4)	475 (68.4)
Subsample 3	511 (73.6)	518 (74.6)	474 (68.3)
Subsample 4	511 (73.6)	506 (72.9)	454 (65.4)
Subsample 5	504 (72.6)	506 (72.9)	455 (65.6)
Consolidated	510 (73.5)	515 (74.2)	476 (68.6)
Accuracy (high volatility class; n=185)			
Subsample 1	114 (61.6)	122 (65.9)	119 (64.3)
Subsample 2	116 (62.7)	117 (63.2)	129 (69.7)
Subsample 3	114 (61.6)	115 (62.2)	121 (65.4)
Subsample 4	118 (63.8)	121 (65.4)	124 (67.0)
Subsample 5	120 (64.9)	119 (64.3)	123 (66.5)
Consolidated	115 (62.2)	121 (65.4)	125 (67.6)
Overall accuracy (n=879)			
Subsample 1	624 (71.0)	635 (72.2)	587 (66.8)
Subsample 2	634 (72.1)	633 (72.0)	604 (68.7)
Subsample 3	625 (71.1)	633 (72.0)	595 (65.8)
Subsample 4	629 (71.6)	627 (71.3)	578 (65.8)
Subsample 5	624 (71.0)	625 (71.1)	578 (65.8)
Consolidated	625 (71.1)	636 (72.4)	601 (68.4)

[a]LASSO: least absolute shrinkage and selection operator.

We applied majority voting to the models developed using all 132 features to compare to the consolidated performances of the models developed using the 9 selected features. Figure 5 shows the comparative performance of the consolidated models developed using 132 or 9 features.

Logistic regression–based models perform equally well when developed using 132 or 9 features. This was expected because both methods used regularization to reduce the magnitude of some features' coefficients. As such, the effect of a redundant feature was significantly diminished even when all features were included in the model.

Random forests performed better than logistic regression methods in predicting the high volatility class. The consolidated overall accuracy measure did not drop significantly (601/879; 68.4% vs 618/879; 70.3%) when only 9 important features were included in the prediction model. The consolidated model's accuracy was very close to the best performing model (Subsample 2) and was better than four other models (Subsample 1, Subsample 3, Subsample 4, and Subsample 5). Thus, consolidating prediction output from models trained on multiple random subsamples using majority voting performs well when random forests are employed with all features as well as selected important features.

XSL•FO
RenderX

Figure 5. Comparison of consolidated prediction accuracy achieved by all 132 features versus the 9 selected important features. LASSO: least absolute shrinkage and selection operator.

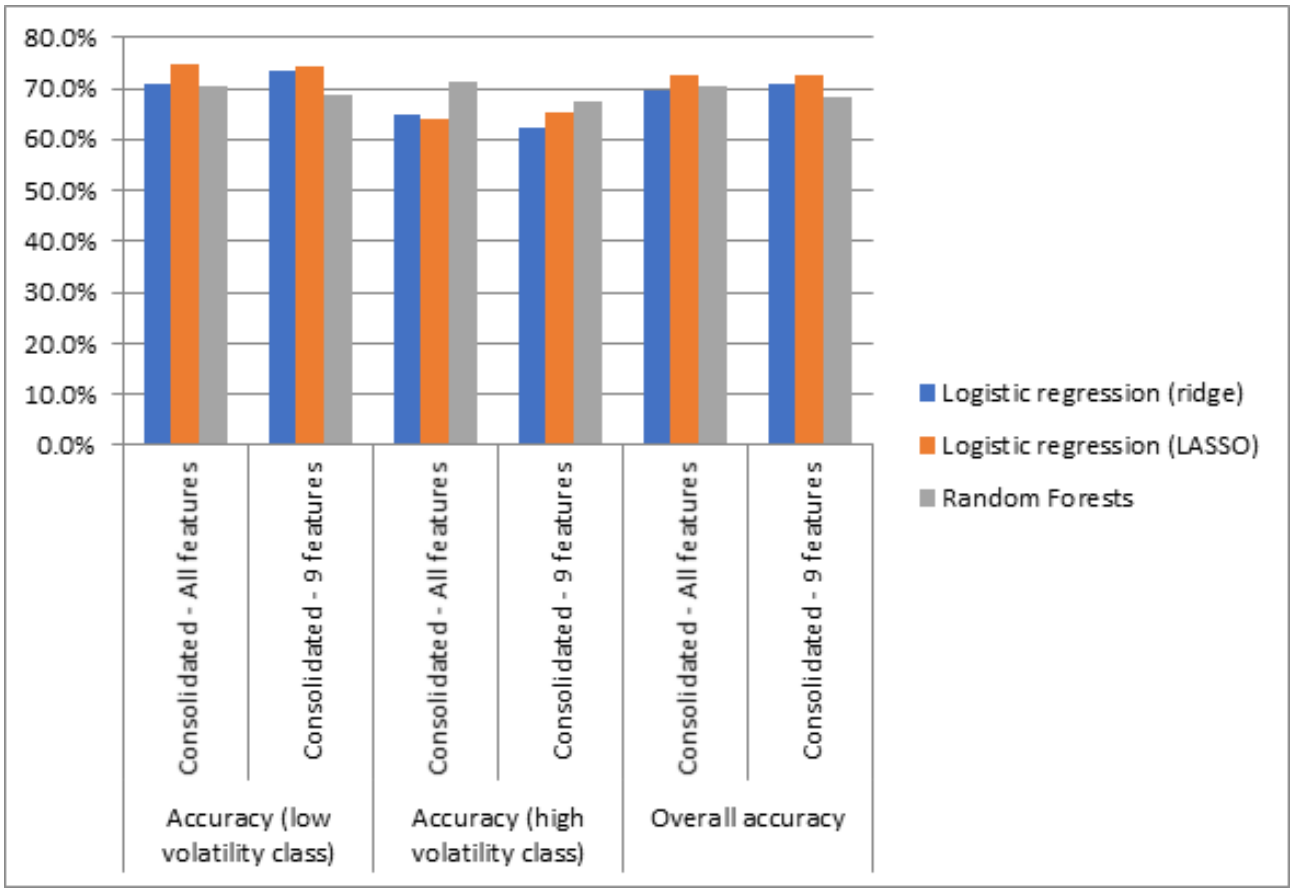


Figure 7. Standalone Equation 2.

$$\textit{Accuracy of the low volatility class} = \frac{\textit{Number of correctly predicted low volatility users}}{\textit{Total number of low volatility users}} \times 100\%,$$

$$\textit{Accuracy of the high volatility class} = \frac{\textit{Number of correctly predicted high volatility users}}{\textit{Total number of high volatility users}} \times 100\%,$$

$$\textit{Overall Accuracy} = \frac{\textit{Number of correctly predicted low and high volatility users}}{\textit{Total number of users}} \times 100\%.$$

Discussion

Principal Findings

In this study, we identified features with high predictive capability that could play an important role in predicting pain volatility in users of Manage My Pain, a digital health app for recording pain experiences. Initially, 132 features were extracted, and four methods were used to develop prediction models (logistic regression with ridge estimators, logistic regression with LASSO, random forests, and SVM). We used Gini impurity and information gain criteria to rank features based on their importance. We also employed the Boruta feature selection method to identify a subset of important features. We conducted 5-fold cross validations for training and testing, and repeated random under-sampling 5 times to address the class imbalance issue. Thus, there were 25 different training sets, and for each feature selection method the common important features across all these 25 sets were identified. Finally, we combined the feature sets selected by the three methods to create a list of 9 important features. Two of these 9 features are from the app usage category and the other 7 are from users' self-reported pain statistics.

XSL•FO
RenderX

Majority voting was utilized to consolidate prediction models trained on the multiple random subsamples to address class imbalance in the dataset. This method worked effectively in achieving the prediction performance closest to the best performing trained model. After feature reduction, the prediction accuracy achieved by two logistic regression methods did not decrease. This shows that the regularization technique embedded in these two methods effectively minimized the impact of redundant features. The consolidated random forests–based models using 9 selected features achieved approximately 68% accuracy for both low and high volatility classes. This is close to the 70% accuracy achieved when models were developed using all 132 features. Logistic regression methods performed better than random forests in predicting the low volatility class while random forests achieved better accuracy for the high volatility class.

Major Contributions and Future Work

This study continues two prior studies [3,4] where data mining and machine learning methods were used to analyze mobile app users' pain data. We effectively reduced the set of 132 features drawn from 8 different categories to 9 important features from 2 categories, extracted over the first month of app use, to predict pain volatility at the sixth month. We achieved this without significant reduction in prediction accuracy. Thus, the prediction models developed can be more effectively interpreted and applied as reducing the number of features from 130 to 9 with little loss in accuracy aids interpretability. This is in part because accuracy and facility of medical decision-making depends on the quantity and complexity of information [20]. Health care providers and patients have a limited capacity to absorb and synthesize a predictor set that contains 130 features, making interpretability a challenge. Moreover, increasing interpretability by reducing the important features to a set of 9 may help health care professionals and patients develop appropriate interventions and pain management plans for the future.

Moreover, the approach of majority voting performed well in consolidating models trained on multiple random subsamples while also addressing the class imbalance issue. Notably, random forests using the selected 9 important features performed better in predicting the high volatility class than logistic regression methods. Correctly predicting future high volatility patients is desirable in many ways even with a minor reduction in the accuracy of low volatility prediction. Accordingly, identifying the important features through the 3 different methods used in this study and then developing nonlinear prediction models using random forests is preferable to developing linear models using logistic regression methods.

In our previous study [4] we noted that mean pain intensity among those affected by chronic pain tends not to change significantly over time, given that the pain is, by definition, chronic. As such, mean changes are not always informative, whereas volatility, the degree of change in pain the person must cope with daily, weekly, or hourly, can be helpful in adaptation and management plans. In this current study, 6/9 of the features that were deemed important were useful in measuring change in pain during the predictor period. This variability in pain during the predictor period strongly predicts future volatility at six months. However, two other interesting features related to app use were also significant predictors of pain volatility in our experiments: the number of pain records and the number of days when users created pain records. While the number of pain records may be interpreted as the number of data points in the predictor period, the significance of the number of days and the correlation between these 2 features requires additional analysis. In the future, we shall conduct additional analyses to investigate the possible multicollinearity among the 9 important features and shall also analyze the effects of interactions between features on future pain volatility.

Limitations

It is usually recommended that datasets used in analytics research are made publicly available for reproducibility and independent verification. However, ManagingLife, the developers of Manage My Pain, is a private organization that serves as the custodian of the data collected by users of Manage My Pain. To ensure it complies with privacy legislation and its own internal privacy policy, ManagingLife cannot make its users' data public.

Acknowledgments

QAR is supported by Mitacs. JK is supported by a Tier 1 Canadian Institutes of Health Research Canada Research Chair in Health Psychology at York University. HC is supported by a Merit award from the University of Toronto, Department of Anesthesia. JMH is a York University Research Chair.

Conflicts of Interest

TJ is the founder and CEO of ManagingLife, Inc. JK, HC, and QR are unpaid members of the ManagingLife Advisory Board, providing guidance on the product and the company's research initiatives.

References

1. St Sauver JL, Warner DO, Yawn BP, Jacobson DJ, McGree ME, Pankratz JJ, et al. Why patients visit their doctors: assessing the most prevalent conditions in a defined American population. Mayo Clin Proc 2013 Jan;88(1):56-67 [FREE Full text] [doi: 10.1016/j.mayocp.2012.08.020] [Medline: 23274019]

XSL•FO
RenderX

2. McKay FH, Cheng C, Wright A, Shill J, Stephens H, Uccellini M. Evaluating mobile phone applications for health behaviour change: A systematic review. J Telemed Telecare 2018 Jan;24(1):22-30. [doi: 10.1177/1357633X16673538] [Medline: 27760883]
3. Rahman QA, Janmohamed T, Pirbaglou M, Ritvo P, Heffernan JM, Clarke H, et al. Patterns of User Engagement With the Mobile App, Manage My Pain: Results of a Data Mining Investigation. JMIR Mhealth Uhealth 2017 Jul 12;5(7):e96 [FREE Full text] [doi: 10.2196/mhealth.7871] [Medline: 28701291]
4. Rahman QA, Janmohamed T, Pirbaglou M, Clarke H, Ritvo P, Heffernan JM, et al. Defining and Predicting Pain Volatility in Users of the Manage My Pain App: Analysis Using Data Mining and Machine Learning Methods. J Med Internet Res 2018 Nov 15;20(11):e12001 [FREE Full text] [doi: 10.2196/12001] [Medline: 30442636]
5. ManagingLife Inc. Manage My Pain URL: https://managinglife.com/ [accessed 2019-07-01]
6. Worley MJ, Heinzerling KG, Shoptaw S, Ling W. Pain volatility and prescription opioid addiction treatment outcomes in patients with chronic pain. Exp Clin Psychopharmacol 2015 Dec;23(6):428-435 [FREE Full text] [doi: 10.1037/pha0000039] [Medline: 26302337]
7. Worley MJ, Heinzerling KG, Shoptaw S, Ling W. Volatility and change in chronic pain severity predict outcomes of treatment for prescription opioid addiction. Addiction 2017 Jul;112(7):1202-1209 [FREE Full text] [doi: 10.1111/add.13782] [Medline: 28164407]
8. Bélanger C, Blais Morin B, Brousseau A, Gagné N, Tremblay A, Daigle K, et al. Unpredictable pain timings lead to greater pain when people are highly intolerant of uncertainty. Scand J Pain 2017 Oct;17:367-372. [doi: 10.1016/j.sjpain.2017.09.013] [Medline: 29033299]
9. Merskey H, Bogduk N. Classification of chronic pain: Descriptions of chronic pain syndromes and definitions of pain terms, 2nd Edition. In: Task Force on Taxonomy of the International Association for the Study of Pain. Seattle, WA: IASP Press; 1994.
10. Kaufman L, Rousseeuw P. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley-Interscience; 2005.
11. Cessie SL, Houwelingen JCV. Ridge Estimators in Logistic Regression. Applied Statistics 1992;41(1):191. [doi: 10.2307/2347628]
12. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J Royal Stat Soc: Series B (Methodological) 2018 Dec 05;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
13. Breiman L. Random forests. Machine Learning 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
14. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/BF00994018]
15. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Soft 2010;33(1) [FREE Full text] [doi: 10.18637/jss.v033.i01]
16. Frank E, Hall M, Witten I. Appendix: The WEKA Workbench. In: Data Mining: Practical Machine Learning Tools and Techniques. Burlington, MA: Morgan Kaufmann; 2016.
17. Liaw A, Wiener M. Classification and Regression by randomForest. R News 2002;2(3):18-22 [FREE Full text]
18. Mitchell T. Machine Learning. New York: McGraw-Hill; 1997.
19. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. J Stat Soft 2010;36(11). [doi: 10.18637/jss.v036.i11]
20. Pohlig C. Medical Decision-Making Factors Include Quantity of Information, Complexity. Hospitalist 2012;2012(1) [FREE Full text]

Abbreviations

LASSO: least absolute shrinkage and selection operator
MMP: Manage My Pain
SVM: support vector machines

Edited by G Eysenbach; submitted 23.07.19; peer-reviewed by G Page, M Sokolova, CH Li, Y Jafer, WD Dotson; comments to author 21.08.19; revised version received 11.09.19; accepted 28.09.19; published 20.11.19.

Please cite as:
Rahman QA, Janmohamed T, Clarke H, Ritvo P, Heffernan J, Katz J
Interpretability and Class Imbalance in Prediction Models for Pain Volatility in Manage My Pain App Users: Analysis Using Feature Selection and Majority Voting Methods
JMIR Med Inform 2019;7(4):e15601
URL: http://medinform.jmir.org/2019/4/e15601/
doi:10.2196/15601
PMID:31746764

XSL•FO
RenderX

©Quazi Abidur Rahman, Tahir Janmohamed, Hance Clarke, Paul Ritvo, Jane Heffernan, Joel Katz. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 20.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Age Assessment of Youth and Young Adults Using Magnetic Resonance Imaging of the Knee: A Deep Learning Approach

Ana Luiza Dallora[1], MSc; Johan Sanmartin Berglund[1], PhD; Martin Brogren[2], MSc; Ola Kvist[3], MD; Sandra Diaz Ruiz[3], MD, PhD; André Dübbel[2], MSc; Peter Anderberg[1], PhD

[1]Department of Health, Blekinge Institute of Technology, Karlskrona, Sweden

[2]Optriva AB, Stockholm, Sweden

[3]Department of Pediatric Radiology, Karolinska University Hospital, Stockholm, Sweden

Corresponding Author:
Peter Anderberg, PhD
Department of Health
Blekinge Institute of Technology
Valhallavägen 1
Karlskrona, 37141
Sweden
Phone: 46 0734223736
Email: pan@bth.se

Abstract

Background: Bone age assessment (BAA) is an important tool for diagnosis and in determining the time of treatment in a number of pediatric clinical scenarios, as well as in legal settings where it is used to estimate the chronological age of an individual where valid documents are lacking. Traditional methods for BAA suffer from drawbacks, such as exposing juveniles to radiation, intra- and interrater variability, and the time spent on the assessment. The employment of automated methods such as deep learning and the use of magnetic resonance imaging (MRI) can address these drawbacks and improve the assessment of age.

Objective: The aim of this paper is to propose an automated approach for age assessment of youth and young adults in the age range when the length growth ceases and growth zones are closed (14-21 years of age) by employing deep learning using MRI of the knee.

Methods: This study carried out MRI examinations of the knee of 402 volunteer subjects—221 males (55.0%) and 181 (45.0%) females—aged 14-21 years. The method comprised two convolutional neural network (CNN) models: the first one selected the most informative images of an MRI sequence, concerning age-assessment purposes; these were then used in the second module, which was responsible for the age estimation. Different CNN architectures were tested, both training from scratch and employing transfer learning.

Results: The CNN architecture that provided the best results was GoogLeNet pretrained on the ImageNet database. The proposed method was able to assess the age of male subjects in the range of 14-20.5 years, with a mean absolute error (MAE) of 0.793 years, and of female subjects in the range of 14-19.5 years, with an MAE of 0.988 years. Regarding the classification of minors—with the threshold of 18 years of age—an accuracy of 98.1% for male subjects and 95.0% for female subjects was achieved.

Conclusions: The proposed method was able to assess the age of youth and young adults from 14 to 20.5 years of age for male subjects and 14 to 19.5 years of age for female subjects in a fully automated manner, without the use of ionizing radiation, addressing the drawbacks of traditional methods.

(JMIR Med Inform 2019;7(4):e16291) doi:10.2196/16291

KEYWORDS

age assessment; bone age; skeletal maturity; deep learning; convolutional neural networks; transfer learning; machine learning; magnetic resonance imaging; medical imaging; knee

XSL•FO
RenderX

Introduction

Background

Bone age and skeletal maturity are closely related concepts that measure the stage of bone development of an individual [1,2]. When compared to the chronological age, they aid in the diagnosis and in determining the time of treatment of many pediatric disorders related to orthodontics, orthopedics, and endocrinology. Further, they are also used in estimations about the final height of an individual [3].

From a legal standpoint, bone age assessment (BAA) also plays an important role in the estimation of chronological age. In this sense, the estimation of the bone age is employed when determining if an individual is a minor in the absence of valid documents, which is the case for numerous unaccompanied minors seeking asylum [2], as well as in adoption, imputability, and pedopornography judicial and civil issues [4]. The estimation of chronological age is also used in age-related sports competitions to guarantee fair play [5,6]. In all of these cases, BAA is an important tool that is used to make important legal decisions that can enormously affect an individual's life.

The traditional methods for performing BAA are the Greulich-Pyle (GP) atlas and the Tanner-Whitehouse (TW) scoring system. The GP atlas [7] comprises hand and wrist radiograph reference images of subjects from 0 to 19 years of age for males and 0 to 18 years of age for females. The process for determining bone age is done by comparing the nearest matching reference image in the atlas to the image of the individual being assessed [3]. The TW scoring system [8] first analyzes the hand and wrist radiograph of a subject and categorizes the skeletal maturity scores of the ossification centers of the radius, ulna, and 13 short bones of the hand and carpals into stages ranging from A to I. Then, all of the stages are aggregated into a numerical score that is converted to the bone age [2].

Drawbacks of the Traditional Age-Assessment Methods

The drawbacks of the GP and TW methods derive from the fact that they are done manually by radiologists; thus, they can be prone to inter- and intrarater variability, in addition to being time-consuming tasks [9,10].

Also, there is an important ethical issue related to submitting healthy subjects to ionizing radiation without therapeutic purposes, which is especially important in the case of assessing if an individual is a minor for legal purposes [10]. This scenario suggests that new approaches for the assessment of age should be explored by research in order to address these drawbacks.

The use of radiation-free medical imaging can be achieved by the employment of magnetic resonance imaging (MRI). An additional advantage of MRI technology is that it supports the manipulation of the image's contrast, granting the possibility of highlighting different tissue types and allowing better visualization of ossification centers [11,12]. Additionally, since MRI images are volumetric, more information can be extracted and analyzed when compared to 2D radiographs [13].

The issues related to rater variability and time spent in the assessment are big motivators for the use of more automated techniques like deep learning. Deep learning is a type of machine learning technique, which refers to algorithms that are able to learn a task from a set of training examples; in view of a new set of data, this task can be reproduced with an acceptable performance [14]. The use of machine learning for health applications is not new and is broadly employed for disease prediction and prognosis [14,15], genomics, proteomics, and microarrays [16]; it has also been used to predict health care utilization through Web search logs [17]. Contrary to many machine learning techniques, deep learning methods perform feature engineering: instead of having a domain expert specify important data characteristics, it learns the informative representations in the data and performs a task of classification or regression [18,19]. When working with medical images, this is especially advantageous since image features are difficult to translate into descriptive means [20]. That is the reason why the first applications of deep learning with health data were aimed at analyzing medical images, specifically MRI images of the brain for the prediction of Alzheimer disease and MRI images of the knee to estimate the risk of osteoarthritis [21]. In the specific area of BAA, most computerized approaches extract features following established procedures (eg, TW or GP), which can be limiting in terms of the information available in the image [22]. When using deep learning, the algorithm finds the important representations in the images without any constraint, which could allow more features in the image to be considered in the classification or regression task not previously known by the current methods [22].

Goal of This Study

Taking into account the numerous settings in which the estimation of chronological age is employed and their importance and potential effect on individuals' lives, it is important to address the drawbacks in the methods currently in use. Thus, this paper proposes an automated approach for age assessment of youth and young adults (14-21 years of age) employing deep learning methods with MRI images of the knee.

The knee region aggregates four ossification centers—femur, tibia, fibula, and patella—but it has not been explored very much by the research in BAA, which is mostly focused on the hand and wrist regions; this research makes use of radiograph images, due to the impact the GP method, which is still considered by many to be the gold standard for BAA [23]. The choice of the knee region in this study was motivated by findings in the research with MRI images that reported the presence of cartilage signal intensity at the knee ossification centers in male individuals from 17.8 to 30.0 years of age and female individuals from 16.6 to 29.6 years of age, which could imply later fusion of maturation centers [24]. Additionally, recent findings in the research of BAA with MRI images of the knee also reported a uniform spatial pattern of maturation of ossification centers in the knee in both male and female individuals [12].

XSL•FO
RenderX

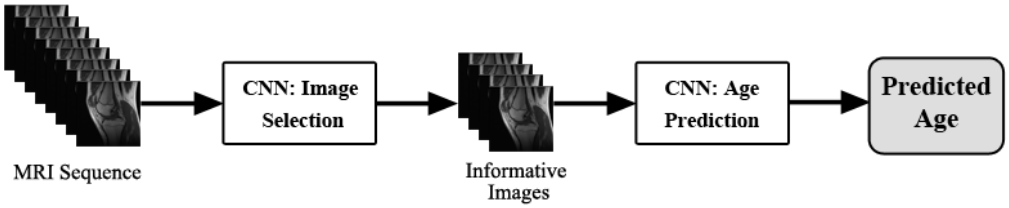
Methods

Overview

The fully automated age-assessment method proposed in this paper uses MRI images of the knee and the subjects' chronological ages to train deep learning models for continuous age estimation with convolutional neural networks (CNNs).

An overview of the method is shown in Figure 1. It comprises two CNN models: the first one is responsible for selecting the most informative images of an MRI sequence for age-assessment purposes; these are then fed to the age-prediction CNN, which outputs an estimated age. The remainder of this section further details the process of training, deploying, and evaluating the CNN models of the proposed method as well as the materials used in the experiments.

Figure 1. Overview of the proposed automated age-assessment method. CNN: convolutional neural network; MRI: magnetic resonance imaging.



Recruitment

This study prospectively acquired MRI images of the knee region of 402 volunteer subjects—221 males (55.0%) and 181 (45.0%) females—aged 14.0-21.5 years (see Table 1) between 2017 and 2018. It is important to note that throughout the text of this paper, the mention of an age group X refers to an age span from X to X.5 (eg, the age group 14 refers to an age span of 14 to 14.5 years). The criteria used for subject recruitment in the study were as follows:

1. Inclusion criteria: subjects (1) were born in Sweden and (2) have a birth certificate verified by national authorities.
2. Exclusion criteria: subjects (1) have a history of bilateral fractures or trauma near the growth plate, (2) have a history of chronic disease or long-term medication, (3) exhibit noncompliance during MRI examinations, (4) have resided outside Sweden for more than 6 consecutive months, and (5) experienced a past pregnancy or were pregnant at the time of recruitment: all female volunteer subjects were tested.

Table 1. Age distribution of the volunteer subjects[a] (N=402).

Gender	Subject age group[b], years, n (%)								Total, n (%)
	14	15	16	17	18	19	20	21	
Male (N=221)	22 (10.0)	26 (11.8)	31 (14.0)	25 (11.3)	24 (10.9)	25 (11.0)	35 (15.8)	33 (14.9)	221 (100)
Female (N=181)	22 (12.2)	21 (11.6)	30 (16.6)	27 (14.9)	20 (11.0)	12 (6.6)	25 (13.8)	24 (13.3)	181 (100)
Total (N=402)	44 (10.9)	47 (11.7)	61 (15.2)	52 (12.9)	44 (10.9)	37 (9.2)	60 (14.9)	57 (14.1)	402 (100)

[a]All data were acquired within a maximum of 6 months after the subjects' birth dates.

[b]Age group X refers to an age span from X to X.5 (eg, the age group 14 refers to an age span of 14 to 14.5 years).

Magnetic Resonance Imaging Examinations

The MRI examinations were performed on 1.5 Tesla whole-body MRI scanners with dedicated knee coils. The images were taken from the nondominant side of the knee; however, in the case of previous fracture or trauma near these regions, the dominant side was imaged.

The examinations were performed in two sites, with the same protocol, 256 x 256-pixel resolution, and 160 x 160 mm field of view. The following machinery was used:

1. Site 1: MAGNETOM Avanto Fit (Siemens Healthcare Gmbh) and Achieva (Philips Healthcare) whole-body scanners.
2. Site 2: SIGNA (GE Healthcare) whole-body scanner.

Data Privacy and Study Ethics

All acquired data were anonymized and stratified by age and gender. The study was approved by the local ethics committee and was conducted in accordance with the Declaration of Helsinki. Written informed consent was acquired from all subjects and legal guardians, in the case of minors.

Image Selection

Each MRI examination produced 17-35 images per subject, however, not all of them were equally informative in regard to the assessment of the age of an individual. To simplify the age estimation learning task, only the best images were considered for the *CNN: Age Prediction* model. To make the method fully automated without any need for human input, a CNN classifier was trained to be able to select the most informative images in an MRI sequence. An *informative* image in the context of the proposed method corresponds to the part of the bone that contains anatomical structures of interest, which include the

XSL•FO
RenderX

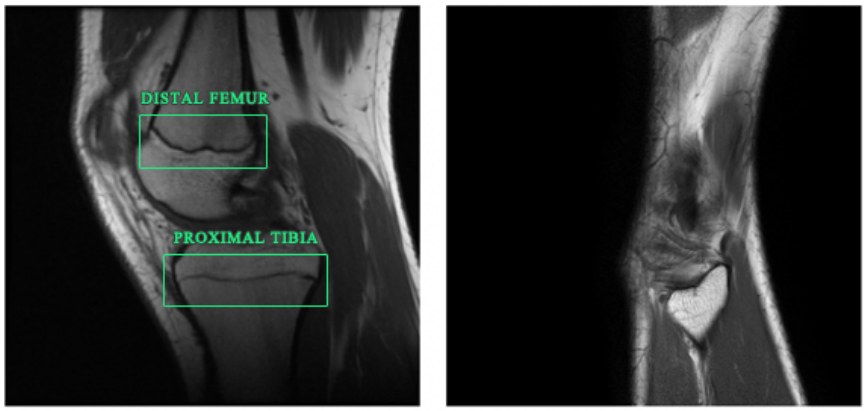
growth plate, epiphysis, and metaphysis. This classifier corresponds to the *CNN: Image Selection* block in Figure 1.

The CNN architecture used was GoogLeNet [25], a model that has been shown to generalize well to a wide variety of image classification tasks, medical and otherwise [26].

To be able to train this classifier, one image from each MRI sequence that had growth zones clearly visible was annotated as *informative*. Also, one image from each MRI sequence in which the growth zones were occluded by other tissue types was selected and labelled as *noninformative*. Examples of informative and noninformative images are shown in Figure 2.

The output of the CNN model is the confidence levels of the two classes—informative and noninformative—for the given MRI image. The confidence level is a continuous value between 0 and 1, where 1 is the highest confidence level and the confidence levels of the two classes sum up to 1. In later steps, only images with a confidence level for the informative class above a threshold C on the test set were used.

Figure 2. Examples of informative and noninformative images from the same subject.



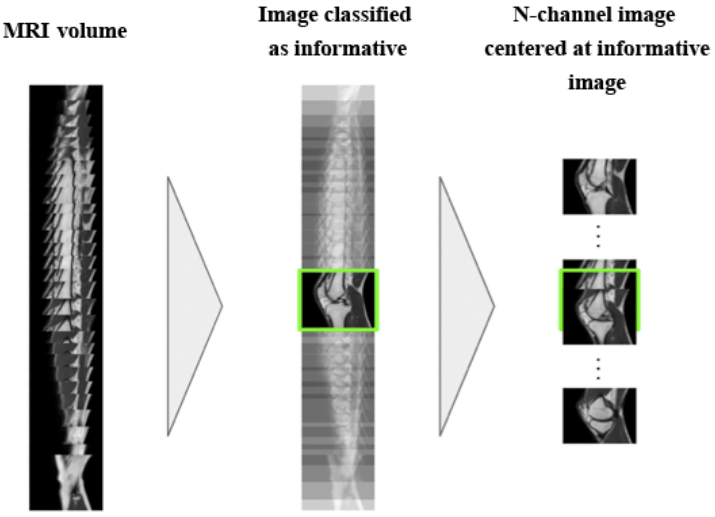
Age Prediction

For predicting the age of an individual from the MRI images, another CNN model was built. This model corresponds to the *CNN: Age Prediction* block in Figure 1. Seven different CNN architectures were considered; these were as follows: GoogLeNet [25], ResNet-50 [27], Inception-v3 [28], Visual Geometry Group (VGG) [29], AlexNet [30], DenseNet [31], and U-Net [32].

The final classification layer of these networks was replaced with a linear scalar output providing the age estimation. The only exception from this was U-Net, which is a fully connected model without classification layers in the end. Here, the linear scalar output was added after the last convolutional layer instead.

The age-prediction model takes an MRI image with N channels as input, then outputs the estimated chronological age of the subject. To create an image with N channels, a subset of the MRI volume, centered on an image classified as informative, is extracted (see Figure 3).

Figure 3. Example of how an N-channel image is created from one of the images in the magnetic resonance imaging (MRI) volume classified as informative.



Input images of 1-9 channels were tested. The idea was that the model might be able to use information from neighboring images to improve results and make the model more robust to mistakes in the image-selection process.

XSL•FO
RenderX

Training the Models

Training and Evaluation

The Convolutional Architecture for Fast Feature Embedding (Caffe) deep learning framework [33] was used to train the models. Training and evaluation were done on Amazon Web Services on an Elastic Compute Cloud (EC2) P3.2xlarge with a Tesla V100 Nvidia graphics processing unit.

Optimization

The Adam optimizer [34] was used to minimize the cross-entropy loss when training the classifier and the Euclidean loss when training the regressor. Cross-entropy loss for binary classification is calculated as follows:

$$-1/N \Sigma^{N}_{i=1} y_i \times log(p(y_i)) + (1-y_i) \times log(1-p(y_i)) \quad (1)$$

with N being the number of training samples per batch, y being a binary indicator (0 or 1) of the correctness of classification for an observation o being of class c, and p being the predicted probability of an observation o being of class c. Euclidean loss is calculated as follows:

$$1/2N \Sigma^{N}_{i=1} \left| |x_i^1 = x_i^2| \right|^2_2 \quad (2)$$

with N being the number of training samples per batch, x^1 the estimated age, and x^2 the verified chronological age.

Cross-Validation

All experiments were performed using six-fold cross-validation, including the test set. The dataset was split into six equal-sized parts, with data stratified for age and gender. This data partition followed the procedure that all of the images from a subject were assigned to a single fold. Four parts were used for training, one part was used for validation during training, and one part was used to finally evaluate and measure the model's performance. This was done to be able to evaluate the models on the full dataset.

Before performing a full cross-validation, a sparse grid search was performed for each model to find good hyperparameters. This was done using the validation set of the first cross-validation split only. The hyperparameters tuned during the grid search were as follows: learning rate, weight decay, momentum, dropout ratio, and batch size.

Transfer Learning

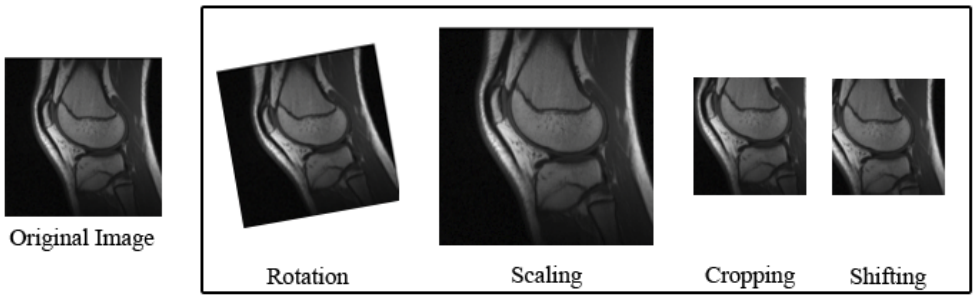
Both training from scratch and transfer learning were tested. Transfer learning is a technique that, instead of using randomly initialized weights, takes the weights from a CNN that has already been trained to perform well on a generic task as a starting point. The model is then adapted by carefully updating the weights using the task-specific training data. This makes it possible to leverage larger datasets to avoid overfitting when the task-specific dataset is small [35,36]. All pretrained models used in this paper were trained on ImageNet [37]. During the task-specific training, the weights of all layers were updated.

Data Augmentation

Data augmentation is a technique that aims to synthetically increase the size of the training set from existing data without additional labelling work, using geometric or photometric transformations, noise injections, and color jittering operations. It is used to prevent overfitting when training CNNs on small datasets [38,39].

In the proposed method, data augmentation was performed on all training samples to increase the dataset. The images were randomly cropped, shifted, rotated at a maximum of five degrees, and scaled up to 20%. Figure 4 shows examples of the applied data augmentation operations.

Figure 4. Examples of data augmentation operations applied in the proposed method.



Estimation

When estimating the age on the test set for each subject, all images with a confidence higher than threshold C of 0.95 for the informative class were used. Each of these test images were used to create a number of copies with different augmentations applied to each copy. All augmented test images were fed through the network to produce one result each. Finally, the results from the augmented versions of the images were used to estimate a final result. This technique has been shown to improve the performance of the predictions and is widely used within deep learning [25].

In this method, each image was augmented 15 times, using the same augmentations as during training, generating 15 new images. If none of the images for a subject had a confidence higher than the threshold, the image with the highest confidence was used instead. This was the case for two subjects only. The highest confidence value for these subjects were 0.91 and 0.81. If more than 10 images had a confidence level higher than the threshold, only the 10 images with the highest confidence were used in order to set a maximum limit on the processing time.

Age was estimated for all augmented images and, finally, the median of all estimated ages for each subject was computed to get the final prediction. For example, if a subject had eight

XSL•FO
RenderX

images with high-enough confidence, 120 augmented images were created and 120 ages were estimated, of which the median was used as the final estimated age.

Results

Overview

Hyperparameters and settings were tuned to optimize the models' performance. This was done through a sparse grid search on the first cross-validation split, as specified previously. The validation set was used for tuning in order to avoid tuning specifically toward the test set and thereby overestimating the models' performance on new data. The final results reported in this section were evaluated on the full dataset from the cross-validation test sets in terms of the mean absolute error (MAE), calculated as follows:

$$MAE = 1/n \sum_{i=1}^{n} |x_i - x| \quad (3)$$

with n being the number of samples, x_i being the estimated age, and x being the verified chronological age.

Conclusions From Experiments

Fine-tuning pretrained models showed significantly better results compared to training the models from scratch. The two architectures that showed best results were GoogLeNet and ResNet-50. Training on men and women subjects separately gave better results for both groups compared to single training using all data.

The best results were achieved using a confidence threshold C of 0.95 in the image selection data preprocessing stage for choosing the most informative MRI images. The results did not change much using different thresholds. MAE differed only by 0.004 years when using thresholds in the range of 0.5-0.99.

Results were very similar when using MRI images with one or three channels, but with more channels than three the performance dropped. This can be due to the increasing number of parameters in the models when using more channels, which might lead to overfitting. Using one channel gave a slightly better result, which is why we used this in our final models.

The hyperparameters that gave the best results were as follows:

1. Learning rate: 1e-4
2. Weight decay: 1e-2
3. Momentum: 0.83
4. Dropout ratio: 0.7 for GoogLeNet and 0.6 for ResNet-50
5. Batch size: 66 for GoogLeNet and 30 for ResNet-50

The best results were achieved when resizing the images to 256×256 pixels for both GoogLeNet and ResNet-50. Both these architectures use cropped images of size 224×224 pixels as input.

Results for the Best Models

The results for the experiments with the best-performing models, GoogLeNet and ResNet-50, in terms of the MAE and SD per age group is shown in Figure 5 and detailed in Table 2 below. The acquisition of the MRI images happened in a window within 6 months from the subjects' birthdays. The best overall results for male subjects were achieved by the GoogLeNet model using knee MRI images. When training the age-prediction model for women, only the architecture performing best on men was considered.

There is a clear trend on all of the experiments among male subjects in which the MAE increases substantially from the age of 21. The same phenomenon occurs for the model among women subjects but from the age of 20. These results lead us to believe that after the ages of 20.5 for men and 19.5 for women, no information regarding older ages can be extracted from the MRI image data, regarding the knee region. This is also supported by Figure 6 and Table 3, which show that the mean estimated age planes out around these ages for the respective genders. The models underestimated the age more and more the older the subjects got after these ages. In conclusion, the presented method is not able to estimate ages above 20.5 for men and above 19.5 for women. Therefore, these ages were removed in the results below, which focus on the applicable age ranges for the models: 14 to 20.5 years for men and 14 to 19.5 years for women.

XSL•FO
RenderX

Figure 5. Comparison of the best-performing models: GoogLeNet and ResNet-50. MAE: mean absolute error.

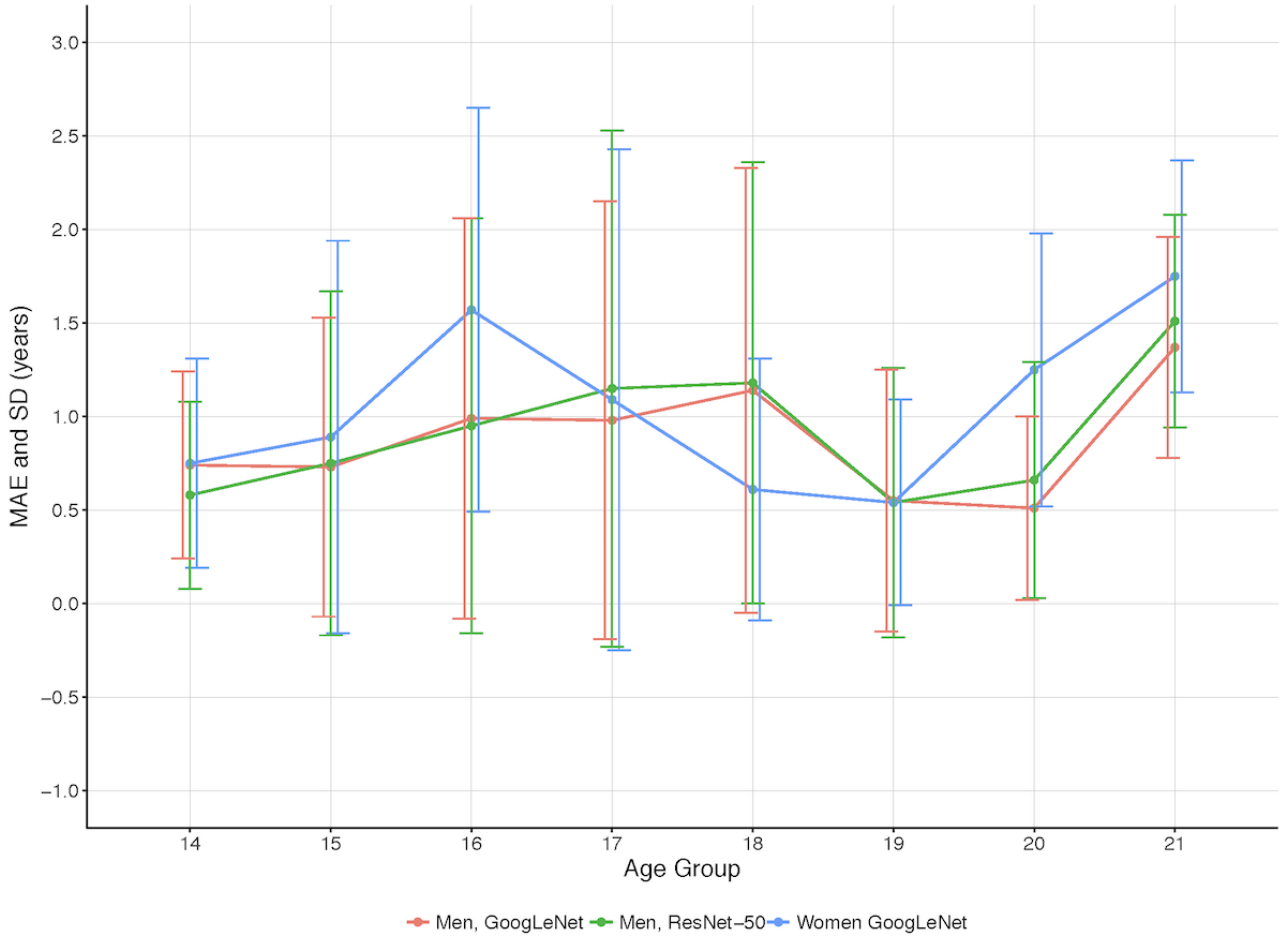


Table 2. Results from the experiments with the best-performing models: GoogLeNet and ResNet-50.

Gender, model	Subject age group[a] in years, MAE[b] (SD)							
	14	15	16	17	18	19	20	21
Men, GoogLeNet	0.74 (0.50)	0.73 (0.80)	0.99 (1.07)	0.98 (1.17)	1.14 (1.19)	0.55 (0.70)	0.51 (0.49)	1.37 (0.59)
Men, ResNet-50	0.58 (0.50)	0.75 (0.92)	0.95 (1.11)	1.15 (1.38)	1.18 (1.18)	0.54 (0.72)	0.66 (0.63)	1.51 (0.57)
Women, GoogLeNet	0.75 (0.56)	0.89 (1.05)	1.57 (1.08)	1.09 (1.34)	0.61 (0.70)	0.54 (0.55)	1.25 (0.73)	1.75 (0.62)

[a]Age group X refers to an age span from X to X.5 (eg, the age group 14 refers to an age span of 14 to 14.5 years).

[b]MAE: mean absolute error.

XSL•FO
RenderX

Figure 6. Mean age and mean estimated age per age group with the best-performing model, GoogLeNet, on male and female subjects.

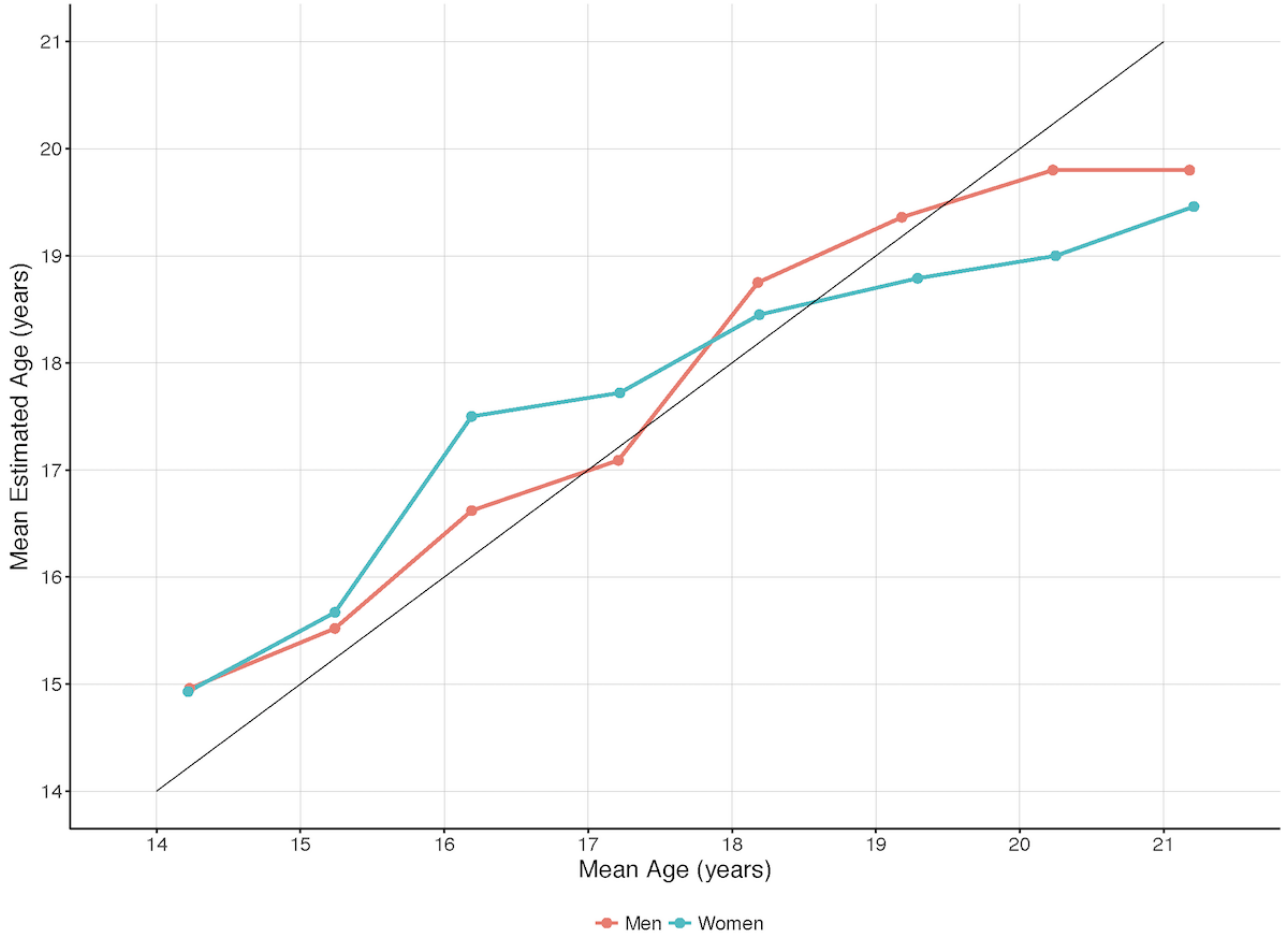


Table 3. Mean age and mean estimated age per age group by the best-performing model, GoogLeNet, on male and female subjects.

Gender	Subject age group[a], years							
	14	15	16	17	18	19	20	21
Men, mean age	14.23	15.24	16.19	17.21	18.18	19.18	20.23	21.18
Men, mean estimated age	14.96	15.52	16.62	17.09	18.75	19.36	19.80	19.80
Women, mean age	14.22	15.24	16.19	17.22	18.19	19.29	20.25	21.21
Women, mean estimated age	14.93	15.67	17.50	17.72	18.45	18.79	19.00	19.00

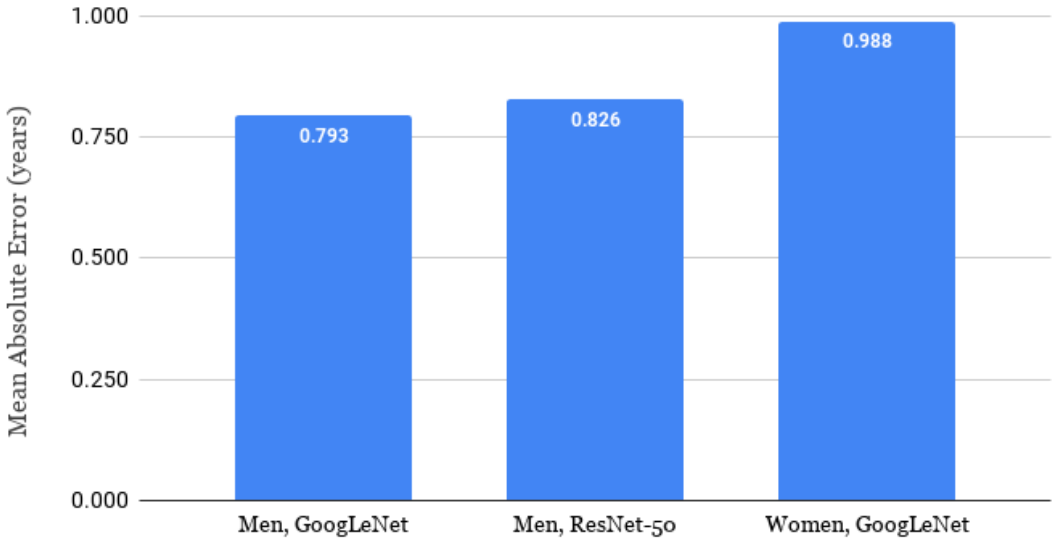
[a]Age group X refers to an age span from X to X.5 (eg, the age group 14 refers to an age span of 14 to 14.5 years).

Results for the Best Models in the Applicable Age Ranges

Figure 7 shows the MAE in years for the best models in their applicable ranges: 14-20.5 years for men and 14-19.5 years for women. The best achieved result for the age prediction of youth and young adult individuals in this study corresponds to an MAE of 0.793 years for men and 0.988 years for women, using the GoogleNet architecture.

XSL•FO
RenderX

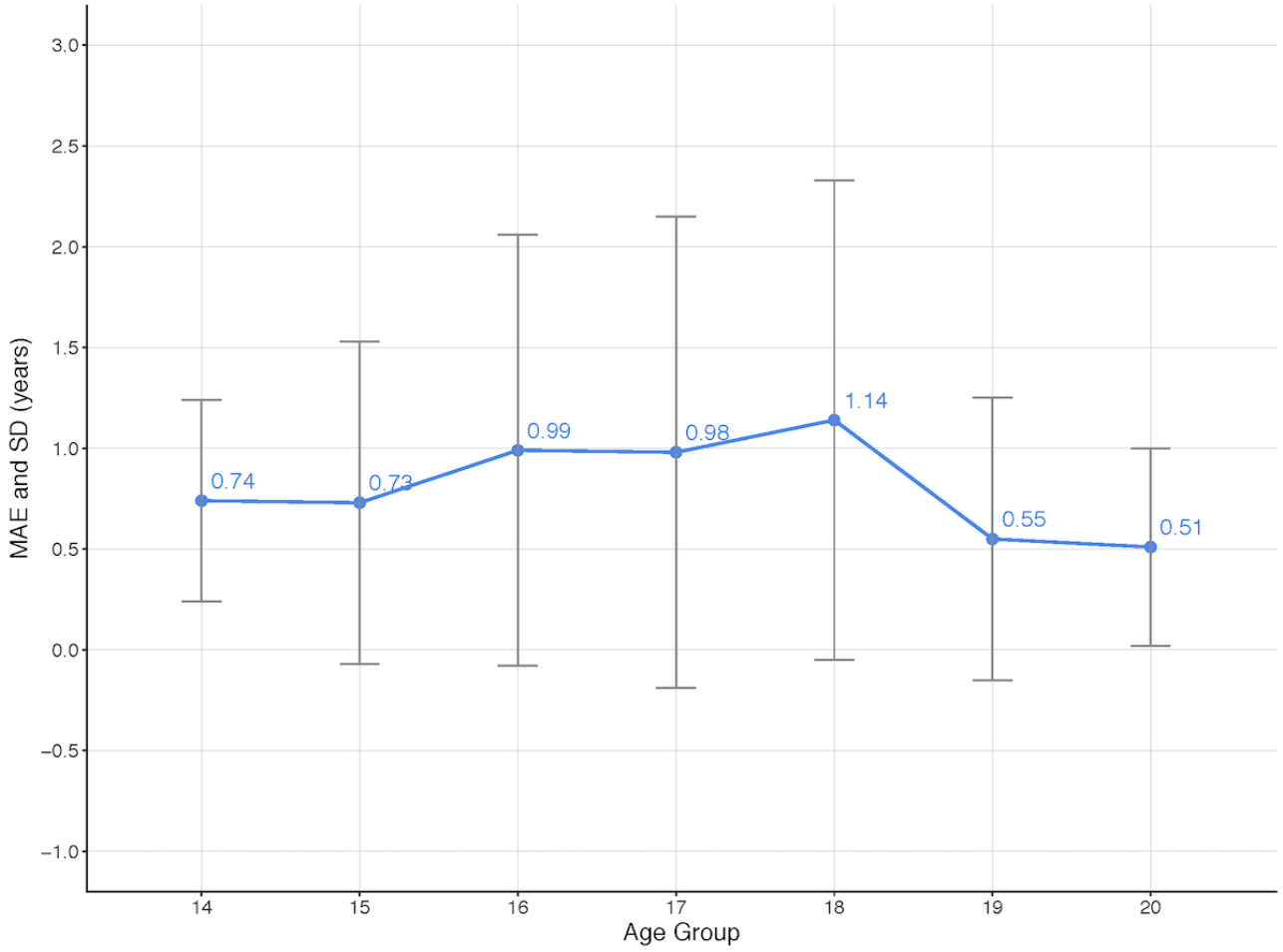
Figure 7. Mean absolute error (MAE) of the best-performing models in the applicable age ranges.



Results for the GoogLeNet Model in the Applicable Age Ranges for Male and Female Subjects

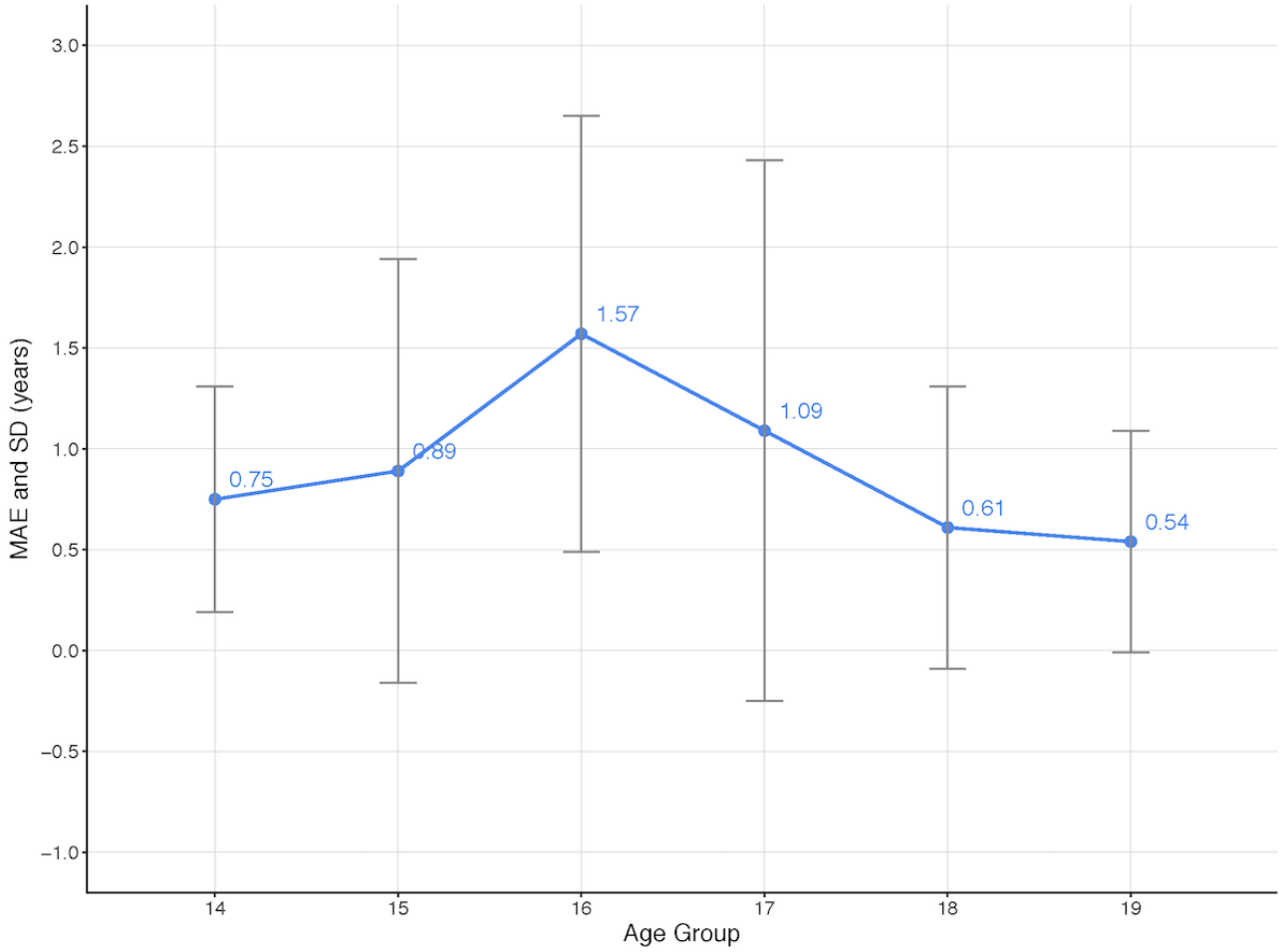
Figures 8 and 9 show the MAE for the GoogLeNet model applied to male and female subjects, respectively, in the applicable age ranges. It is interesting to notice that the age range with the highest error occurs earlier for females (age group of 16) compared to men (age group of 18). This goes in line with previous knee studies where findings showed that women mature earlier than men [40].

Figure 8. Mean absolute error (MAE) for the GoogLeNet model for male subjects in the applicable age ranges.



XSL•FO
RenderX

Figure 9. Mean absolute error (MAE) for the GoogLeNet model for female subjects in the applicable age ranges.



Classification Performance of Minors Versus Adults

Experiments were also performed for classification of subjects as being adults or minors, considering the age of 18 years old as the adulthood threshold. This classification is especially important in cases regarding the age assessment of minors from a legal standpoint.

No new training of models was performed. Instead, the classification of adults and minors was performed by applying a threshold to the estimated age from the best-performing models trained in the age-assessment experiments.

Three different strategies for setting the threshold were evaluated:

1. Setting the threshold to increase the accuracy for minors and sacrificing accuracy for adults.
2. Setting the threshold to get as equal accuracy as possible for adults and minors.
3. Using the threshold of 18 years of age without any modification.

The results for male subjects are shown in Figure 10 and Table 4. The same procedures and reasoning were also applied to the women's case and the results are shown in Figure 11 and Table 5.

XSL•FO
RenderX

Figure 10. Accuracies for minor versus adult classification of male subjects, using threshold to increase accuracy for minors.

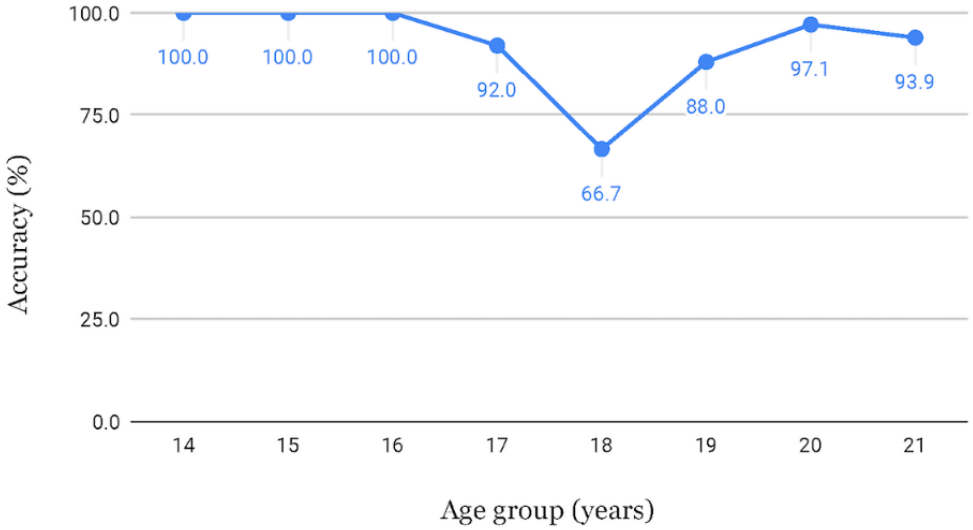


Table 4. Accuracies for minor versus adult classification of male subjects.

Strategy for setting the threshold	Threshold in years	Accuracy for minors, %	Accuracy for adults, %
Using the threshold to get lower errors for minors	18.73	98.1	88.0
Using the threshold to get as equal accuracy for adults and minors as possible	18.38	93.3	93.2
Using estimated age without modifying the threshold	18.00	90.4	95.7

Figure 11. Accuracies for minor versus adult classification of female subjects, using threshold to increase accuracy for minors.

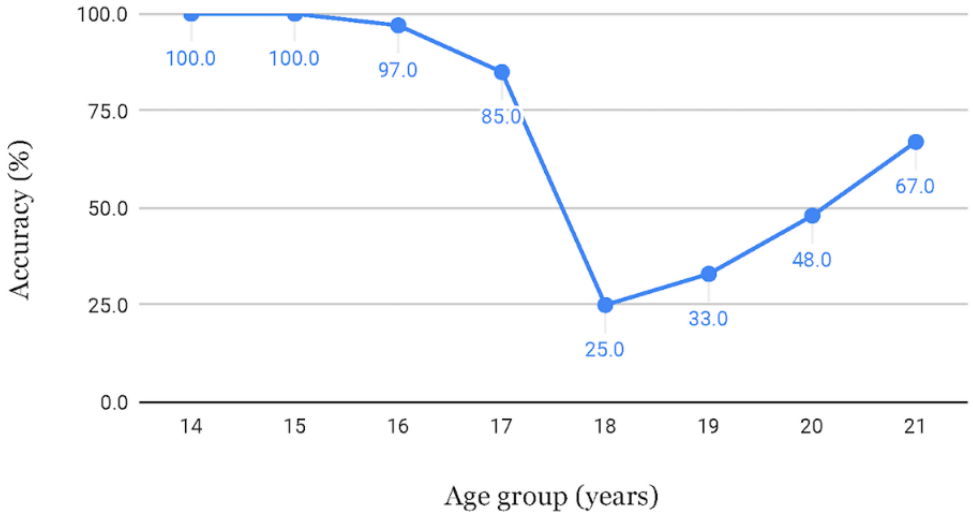


Table 5. Accuracies for minor versus adult classification of female subjects.

Strategy for setting the threshold	Threshold in years	Accuracy for minors, %	Accuracy for adults, %
Using threshold to get lower errors for minors	19.11	95.0	45.7
Using threshold to get as equal accuracy for adults and minors as possible	18.20	85.0	85.2
Using estimated age without modifying the threshold	18.00	77.0	88.9

XSL•FO
RenderX

Discussion

Principal Findings

This paper proposed a fully automated method, free from ionizing radiation, for age assessment based on MRI images of the knee using CNNs. The method was able to assess the age of male subjects in the range of 14-20.5 years of age, with an MAE of 0.793 years, and of female subjects in the range of 14-19.5 years of age, with an MAE of 0.988 years.

The method developed in this paper addresses and proposes solutions to the drawbacks in age-assessment research, which currently deals with the following:

1. Ethical issues of submitting healthy individuals to ionizing radiation for nontherapeutic purposes [10], since most of the established methods (ie, GP and TW) and recently published methods make use, mostly, of radiographs as the analysis input [23]. This paper showed that it is possible to achieve a good estimation of age by employing MRI images instead.
2. Lowering the risk of intra- and interrater variability, which can be very high when general radiologists are employed in the assessment of age instead of high-expertise pediatric radiologists [41,42]. Also, there is limited evidence that contrasts with the findings of manual raters and automatic systems regarding chronological age assessment, since most of the published material is directed to predict bone age [23]. However, a novel study reports a higher rate of false positives in classifying adults—with a threshold of 18 years—from hand images for manual raters compared to a deep learning system [43].
3. Time spent on assessment [9] addressed by the automation of the proposed method, which is able to perform evaluations in real time.

It is also important to mention that the proposed method in this paper provides the estimation of chronological age based on MRI images of the knee, contrary to most previous research, which aimed at estimating bone age and evaluating the methods using bone age and not chronological age. While the concept of bone age is certainly useful and important in many clinical settings, it was not conceived as a method to determine the chronological age of an individual. It was used to examine the developmental status of children and adolescents in comparison to their known chronological age, which can be advanced or delayed due to a multitude of factors that include chronic illnesses, hormonal disorders, etc [7,10]. The widespread use of BAA as an estimation of chronological age sometimes confuses these concepts and they are erroneously used interchangeably, as in many studies to justify the execution of BAA to judicial and civil issues. Also, it can be argued that the bone age attributed to an individual may be subjective and there is no objective way to obtain a confirmation of the exact number. In a clinical setting this may not be a problem since doctors can work with secure thresholds, but if the estimation is done for legal purposes it can become problematic, since decisions based on this estimation, especially regarding the ages of adulthood, can greatly affect the life of the individual in question.

Regarding our experiments, it is shown that for the male subjects, after the age of 20.5 the model could not identify any more information in the MRI images to discriminate the age of individuals. The same phenomenon occurred at the age of 19.5 for female subjects, which could indicate that the transformations that occur in the knee area related to the maturation process occur earlier in women than in men. This is in line with prior research on the knee region [12,24,44].

We also had satisfactory results for the problem regarding the classification of minors versus adults, considering the threshold of 18 years of age, which can be especially important in civil and judicial scenarios. Misclassification of minors as adults can often be viewed as much more problematic than the inverse, since the imputability for the application of laws, as well as guaranteed rights, may be different for these groups of individuals and usually harsher for adults. Our method can reduce that problem by distributing the errors depending on the application, using a modifiable threshold applied to the estimated age. Our method achieved an accuracy of 98.1% for male subjects and 95.0% for female subjects when it came to correctly classifying minors from the MRI images, when using a threshold that increased the accuracy for minors and sacrificed accuracy for adults.

From an operational point of view, the CNN technology employed with transfer learning can be seen as an enabler in performing research with medical images. The high cost for medical imaging can result in smaller datasets for many studies, but this caveat can be partially addressed when using the transfer learning technology on pretrained CNNs that have learned features from generic images. In this study, even if the features changed during training they were not changed much in our case. Generic features seem to work in a satisfactory way for MRI images; it is just detecting edges, corners, and blobs, which are relevant in MRI images as well as in generic images. Therefore, there is a possibility of applying automated methods even for smaller datasets. The study by Spampinato et al reported similar conclusions, but for radiographs of the hand [36].

Comparison With Prior Work

We propose a fully automated and radiation-free method for chronological age assessment based on MRI images of the knee region, employing deep learning techniques. We could not find prior published work with the same attributes in the literature, as not much work has been done in estimating chronological age per se.

A recent study by Stern et al [43] employed MRI volumes of the hand with CNNs in order to predict chronological age of male subjects from 13 up to 19 years of age. They reported an MAE of 0.82 years for subjects under 18 years of age. They also reported results on majority age classification for male subjects between the ages of 13 and 25 years. An error of 5% for minors gave an error of 27.5% for adults, and an error of 1% for minors gave an error of 67.2% for adults. This can be compared to our results where an error of 1.9% for minors gave an error of 12% for adults on male subjects between the ages of 14 and 22 years. In an earlier study by Stern et al [45], they proposed a multi-factorial age estimation method using MRI

XSL•FO
RenderX

volumes of the hand, clavicle, and teeth with CNNs. With this approach, they managed to predict chronological age of male subjects from 13 up to 25 years of age with an MAE of 1.01 years. They also reported results on majority age classification, where an error of 0.5% for minors gave an error of 25.0% for adults, and an error of 3% for minors gave an error of 18.1% for adults. This can be compared to our results, where an error of 1.9% for minors gave an error of 12% for adults on male subjects between 14 and 22 years of age. The results on majority age classification in these two papers by Stern et al [43,45] are the best published results so far, using one or multiple body parts. However, our results are significantly better even compared to their method using MRI data from three different body parts.

The study by Tang et al [46] proposed an artificial neural network model for estimating the chronological age of subjects (12-17 years old) using MRI images of the hand and wrist and other skeletal maturity factors of 79 subjects. In this study, the authors chose as the performance metric the comparison between the mean chronological age for all subjects and the mean estimated age for all subjects (ie, mean disparity), not calculating the error per subject, which could be misleading. The mean disparity measures whether there is a constant offset in the estimations, not the performance of the model on a per-subject level, like MAE does. A model can, therefore, have large errors in age estimation for all subjects and high MAE but can still have a small mean disparity; the MAE was not reported in this paper. Additionally, the reported results were on the validation set, probably due to the small sample size. In this fashion, the authors reported a mean disparity of 0.1 years between the estimated skeletal age and the chronological age.

Prior published methods for BAA that employed automated methods still focused mostly on the hand and wrist regions for the age assessment and made heavy use of radiographs as the input for their systems, as reported by a recent systematic literature review (SLR) and meta-analysis on BAA systems [23].

In this SLR, only two studies were reported to have made assessments based on the knee. The study by O'Connor et al [44] proposed a scoring system based on the assessment of knee radiographs as to the stage of epiphyseal fusion of the femur, tibia, and fibula on subjects from 9 to 19 years of age, employing regression model-building techniques. This study reported residuals of more than 2 bone-age years for both male and female individuals. The study by Fan et al [24] aimed to compare the age assessment based on the knee region from radiographs and MRI images on subjects from 11 to 25 years of age. They built regression models for bone age based on the scoring system by Krämer et al [47] for both image modalities, yielding better results for the MRI images, achieving R^2 values (eg, the variance in the dependent variable that is predicted from the independent variables in regression models) of 0.634 and 0.654 for female and male subjects, respectively.

On the choice of medical imaging, the referred SLR reported only three studies that built systems for BAA based on MRI images; one of these was the study by Tang et al [46], mentioned previously. The study by Urchsler et al [13] designed a system with the deep learning technology to automatically locate the ossification centers on MRI images of the hand and wrist to assess the bone age of individuals, 13-20 years of age, with random forests. This study obtained an MAE of 0.850 bone-age years. The study by Hillewig et al [48] obtained MRI images from the clavicle and radiograph images from the hand and wrist of 220 subjects, 16-26 years of age, and evaluated these regions according to the Schmeling et al [49] and Kreitner et al [50] scoring systems for the clavicle and the hand and wrist, respectively. The study concluded that the assessment of the clavicle alone was not sufficient to discriminate individuals as younger or older than 18 years of age, thus requiring the information from the hand and wrist for the assessment.

Another noninvasive and radiation-free medical imaging method for the estimation of age that is reported in the literature is the assessment of retinal images, which is an approach that provides diagnostic evidence about important diseases, such as cardiovascular disease and diabetes. Retinal images were assessed with deep learning in the study by Poplin et al [51] in predicting a variety of cardiovascular risk factors, including age, which achieved an MAE of 3.26 years. Retinal images were also assessed by Ting et al [52] in estimating the prevalence and systematic risk factors for diabetic retinopathy, which included young age.

In regard to approaches that make use of deep learning methods in the field of BAA, the biggest initiative posed in recent years was done so by the Radiological Society of North America (RSNA) for the prediction of bone age: the RSNA 2018 Pediatric Bone Age Challenge [53]. This challenge aimed to encourage participants to develop algorithms that could most-accurately determine the bone age of subjects from 0 to 19 years of age, providing a database of around 12,000 radiograph images of the hand and wrist, labeled as to their bone age [53]. The participants proposed CNN models, like the ones by Iglovikov et al [54], Zhao et al [55], and Ren et al [22], which achieved MAEs of 7.52, 7.66, and 5.2 months. However good the obtained results were, they were not comparable to our results, since our aim was to predict the chronological age of a subject, and the RSNA project's goal was to predict the bone age. It is also important to note that although these studies made use of large-enough sample sizes, the data were not uniformly distributed, as only 0.1% of the dataset was composed of individuals of 18 and 19 years of age. Additionally, Dallora et al [23] provided a meta-analysis on the performances based on seven studies, which contained all three deep learning studies mentioned previously, where the age ranges were mostly within 0-19 years of age and the performance metrics were given in MAE (bone-age months). The weighted average by the dataset size resulted in 9.96 MAE (bone-age months), which is higher than the results presented in this paper.

Limitations

Regarding the limitations of this study, it could be argued that the sample size would not be big enough to be generalizable; therefore, we employed methods to ensure that the models did not overfit by using test sets separated from the training and validation sets. The results showed that the model was able to generalize to new data in the test sets. Additionally, further

XSL•FO
RenderX

work will be directed to the collection of more data, which may improve the precision and MAE of our models.

Also, we aimed at having a uniform number of subjects for each age group, which was achieved by the data acquisition process; an exception was for the 19-year-old female subjects, who accounted for only 12 subjects, which could be seen as a caveat to the female model.

Additionally, the acquisition of ages for the first half year from each age group may interfere with the estimation accuracy of the minor versus adult classification. The largest impact occurs for the ages closest to 18 years. The missing data for those 17.5-17.99 years of age is important and we plan to collect new data to complement those ages in future work. Concerning the MAE numbers, these missing ages do not have as much impact as for the accuracy numbers.

Finally, the method was built upon data from healthy youth and young adult subjects and the effect of disorders that can affect growth was not explored.

Conclusions

This paper proposed a model for the estimation of chronological age in youth and young adults using MRI images of the knee. Our method demonstrated good results and addressed the biggest drawbacks in the traditional age-estimation procedures that are still currently in use. Our results on majority age classification were significantly better than the best results previously published.

Acknowledgments

We would like to express our greatest appreciation to the participants and staff who took part in our study. This work was supported by the National Board of Health and Welfare of Sweden (Socialstyrelsen). The funding source had no involvement regarding study design, data collection, analysis, interpretation, or reporting of this work.

Conflicts of Interest

None declared.

References

1. Gilsanz V, Ratib O. Hand Bone Age: A Digital Atlas of Skeletal Maturity. Berlin, Germany: Springer-Verlag; 2005.
2. Manzoor Mughal A, Hassan N, Ahmed A. Bone age assessment methods: A critical review. Pak J Med Sci 2014 Jan;30(1):211-215 [FREE Full text] [doi: 10.12669/pjms.301.4295] [Medline: 24639863]
3. Satoh M. Bone age: Assessment methods and clinical applications. Clin Pediatr Endocrinol 2015 Oct;24(4):143-152 [FREE Full text] [doi: 10.1297/cpe.24.143] [Medline: 26568655]
4. Cunha E, Baccino E, Martrille L, Ramsthaler F, Prieto J, Schuliar Y, et al. The problem of aging human remains and living individuals: A review. Forensic Sci Int 2009 Dec 15;193(1-3):1-13. [doi: 10.1016/j.forsciint.2009.09.008] [Medline: 19879075]
5. Fatehi M, Nateghi R, Pourakpour F. Automatic bone age determination using wrist MRI based on FIFA grading system for athletes: Deep learning approach. In: Proceedings of the 26th Annual Scientific Meeting of the European Society of Musculoskeletal Radiology (ESSR). 2019 Presented at: 26th Annual Scientific Meeting of the European Society of Musculoskeletal Radiology (ESSR); June 26-29, 2019; Lisbon, Portugal. [doi: 10.1055/s-0039-1692580]
6. Dvorak J, George J, Junge A, Hodler J. Application of MRI of the wrist for age determination in international U-17 soccer competitions. Br J Sports Med 2007 Aug;41(8):497-500 [FREE Full text] [doi: 10.1136/bjsm.2006.033431] [Medline: 17347314]
7. Greulich WW, Pyle SI. Radiographic atlas of skeletal development of the hand and wrist. Am J Med Sci 1959;238(3):393. [doi: 10.1097/00000441-195909000-00030]
8. Tanner JM, Whitehouse RH, Cameron N, Marshall WA, Healy MJR, Goldstein H. Assessment of Skeletal Maturity and Prediction of Adult Height (TW2 Method). 2nd edition. London, UK: Academic Press; 1975.
9. Mansourvar M, Ismail M, Herawan T, Raj R, Kareem S, Nasaruddin F. Automated bone age assessment: Motivation, taxonomies, and challenges. Comput Math Methods Med 2013;2013:391626 [FREE Full text] [doi: 10.1155/2013/391626] [Medline: 24454534]
10. Hjern A, Brendler-Lindqvist M, Norredam M. Age assessment of young asylum seekers. Acta Paediatr 2012 Jan;101(1):4-7. [doi: 10.1111/j.1651-2227.2011.02476.x] [Medline: 21950617]
11. Crema MD, Roemer FW, Marra MD, Burstein D, Gold GE, Eckstein F, et al. Articular cartilage in the knee: Current MR imaging techniques and applications in clinical practice and research. Radiographics 2011;31(1):37-61. [doi: 10.1148/rg.311105084] [Medline: 21257932]
12. Margalit A, Cottrill E, Nhan D, Yu L, Tang X, Fritz J, et al. The spatial order of physeal maturation in the normal human knee using magnetic resonance imaging. J Pediatr Orthop 2019;39(4):e318-e322. [doi: 10.1097/bpo.0000000000001298]
13. Urschler M, Grassegger S, Štern D. What automated age estimation of hand and wrist MRI data tells us about skeletal maturation in male adolescents. Ann Hum Biol 2015;42(4):358-367. [doi: 10.3109/03014460.2015.1043945] [Medline: 26313328]

XSL•FO
RenderX

14. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8-17 [FREE Full text] [doi: 10.1016/j.csbj.2014.11.005] [Medline: 25750696]
15. Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. IEEE Access 2017;5:8869-8879. [doi: 10.1109/access.2017.2694446]
16. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine learning in bioinformatics. Brief Bioinform 2006 Mar;7(1):86-112. [doi: 10.1093/bib/bbk007] [Medline: 16761367]
17. Agarwal V, Zhang L, Zhu J, Fang S, Cheng T, Hong C, et al. Impact of predicting health care utilization via Web search behavior: A data-driven analysis. J Med Internet Res 2016 Sep 21;18(9):e251 [FREE Full text] [doi: 10.2196/jmir.6240] [Medline: 27655225]
18. van Hartskamp M, Consoli S, Verhaegh W, Petkovic M, van de Stolpe A. Artificial intelligence in clinical health care applications: Viewpoint. Interact J Med Res 2019 Apr 05;8(2):e12100 [FREE Full text] [doi: 10.2196/12100] [Medline: 30950806]
19. Shen D, Wu G, Suk H. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017 Jun 21;19(1):221-248 [FREE Full text] [doi: 10.1146/annurev-bioeng-071516-044442] [Medline: 28301734]
20. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. IEEE J Biomed Health Inform 2017 Jan;21(1):4-21. [doi: 10.1109/jbhi.2016.2636665]
21. Miotto R, Wang F, Wang S, Jiang X, Dudley J. Deep learning for healthcare: Review, opportunities and challenges. Brief Bioinform 2018 Nov 27;19(6):1236-1246 [FREE Full text] [doi: 10.1093/bib/bbx044] [Medline: 28481991]
22. Ren X, Li T, Yang X, Wang S, Ahmad S, Xiang L, et al. Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph. IEEE J Biomed Health Inform 2019 Sep;23(5):2030-2038. [doi: 10.1109/JBHI.2018.2876916] [Medline: 30346295]
23. Dallora AL, Anderberg P, Kvist O, Mendes E, Diaz Ruiz S, Sanmartin Berglund J. Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis. PLoS One 2019 Jul 25;14(7):e0220242 [FREE Full text] [doi: 10.1371/journal.pone.0220242] [Medline: 31344143]
24. Fan F, Zhang K, Peng Z, Cui J, Hu N, Deng Z. Forensic age estimation of living persons from the knee: Comparison of MRI with radiographs. Forensic Sci Int 2016 Nov;268:145-150. [doi: 10.1016/j.forsciint.2016.10.002] [Medline: 27770721]
25. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 7-12, 2015; Boston, MA. [doi: 10.1109/cvpr.2015.7298594]
26. Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. IEEE Access 2018;6:9375-9389. [doi: 10.1109/access.2017.2788044]
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 27-30 , 2016; Las Vegas, NV URL: http://toc.proceedings.com/32592webtoc.pdf [doi: 10.1109/cvpr.2016.90]
28. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 27-30, 2016; Las Vegas, NV. [doi: 10.1109/cvpr.2016.308]
29. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations. 2015 Presented at: 3rd International Conference on Learning Representations; May 7-9, 2015; San Diego, CA URL: http://arxiv.org/abs/1409.1556
30. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May 24;60(6):84-90. [doi: 10.1145/3065386]
31. Huang G, Liu Z, Van DML, Weinberger K. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; July 21-26 , 2017; Honolulu, HI. [doi: 10.1109/cvpr.2017.243]
32. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. 2015 Presented at: 18th International Conference on Medical Image Computing and Computer-Assisted Intervention; October 5-9, 2015; Munich, Germany. [doi: 10.1007/978-3-319-24574-4_28]
33. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R. Caffe: Convolutional Architecture for Fast Feature Embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia. 2014 Presented at: 22nd ACM International Conference on Multimedia; November 3-7, 2014; Orlando, FL. [doi: 10.1145/2647868.2654889]
34. Kingma D, Ba J. Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations. 2015 Presented at: International Conference on Learning Representations; May 7-9, 2015; San Diego, CA URL: http://arxiv.org/abs/1412.6980
35. Kumar A, Kim J, Lyndon D, Fulham M, Feng D. An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE J Biomed Health Inform 2017 Jan;21(1):31-40. [doi: 10.1109/jbhi.2016.2635663]

XSL•FO
RenderX

36. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in X-ray images. Med Image Anal 2017 Feb;36:41-51. [doi: 10.1016/j.media.2016.10.010] [Medline: 27816861]
37. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2009 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 20-25, 2009; Miami, FL. [doi: 10.1109/cvprw.2009.5206848]
38. Zhong Z, Zheng L, Kang G, Li S, Yang Y. arXiv. 2017. Random erasing data augmentation URL: https://arxiv.org/pdf/1708.04896.pdf [accessed 2019-11-20]
39. Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas BA, Alkasab TK, et al. Fully automated deep learning system for bone age assessment. J Digit Imaging 2017 Aug 8;30(4):427-441 [FREE Full text] [doi: 10.1007/s10278-017-9955-8] [Medline: 28275919]
40. Dedouit F, Auriol J, Rousseau H, Rougé D, Crubézy E, Telmon N. Age assessment by magnetic resonance imaging of the knee: A preliminary study. Forensic Sci Int 2012 Apr 10;217(1-3):232.e1-232.e7. [doi: 10.1016/j.forsciint.2011.11.013] [Medline: 22153621]
41. Kaplowitz P, Srinivasan S, He J, McCarter R, Hayeri MR, Sze R. Comparison of bone age readings by pediatric endocrinologists and pediatric radiologists using two bone age atlases. Pediatr Radiol 2011 Jun 16;41(6):690-693. [doi: 10.1007/s00247-010-1915-0] [Medline: 21161206]
42. Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, et al. Artificial intelligence versus clinicians in disease diagnosis: Systematic review. JMIR Med Inform 2019 Aug 16;7(3):e10010 [FREE Full text] [doi: 10.2196/10010] [Medline: 31420959]
43. Štern D, Payer C, Urschler M. Automated age estimation from MRI volumes of the hand. Med Image Anal 2019 Dec;58:101538 [FREE Full text] [doi: 10.1016/j.media.2019.101538] [Medline: 31400620]
44. O'Connor JE, Coyle J, Bogue C, Spence LD, Last J. Age prediction formulae from radiographic assessment of skeletal maturation at the knee in an Irish population. Forensic Sci Int 2014 Jan;234:188.e1-188.e8. [doi: 10.1016/j.forsciint.2013.10.032] [Medline: 24262807]
45. Stern D, Payer C, Giuliani N, Urschler M. Automatic age estimation and majority age classification from multi-factorial MRI data. IEEE J Biomed Health Inform 2019 Jul;23(4):1392-1403. [doi: 10.1109/jbhi.2018.2869606]
46. Tang FH, Chan JL, Chan BK. Accurate age determination for adolescents using magnetic resonance imaging of the hand and wrist with an artificial neural network-based approach. J Digit Imaging 2019 Apr 15;32(2):283-289. [doi: 10.1007/s10278-018-0135-2] [Medline: 30324428]
47. Krämer JA, Schmidt S, Jürgens KU, Lentschig M, Schmeling A, Vieth V. Forensic age estimation in living individuals using 3.0 T MRI of the distal femur. Int J Legal Med 2014 May 7;128(3):509-514. [doi: 10.1007/s00414-014-0967-3] [Medline: 24504560]
48. Hillewig E, Degroote J, Van der Paelt T, Visscher A, Vandemaele P, Lutin B, et al. Magnetic resonance imaging of the sternal extremity of the clavicle in forensic age estimation: Towards more sound age estimates. Int J Legal Med 2013 May 9;127(3):677-689. [doi: 10.1007/s00414-012-0798-z] [Medline: 23224029]
49. Schmeling A, Schulz R, Reisinger W, Mühler M, Wernecke K, Geserick G. Studies on the time frame for ossification of the medial clavicular epiphyseal cartilage in conventional radiography. Int J Legal Med 2004 Feb 1;118(1):5-8. [doi: 10.1007/s00414-003-0404-5] [Medline: 14534796]
50. Kreitner K, Schweden FJ, Riepert T, Nafe B, Thelen M. Bone age determination based on the study of the medial extremity of the clavicle. Eur Radiol 1998 Sep 2;8(7):1116-1122. [doi: 10.1007/s003300050518] [Medline: 9724422]
51. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018 Mar 19;2(3):158-164. [doi: 10.1038/s41551-018-0195-0] [Medline: 31015713]
52. Ting DSW, Cheung CY, Nguyen Q, Sabanayagam C, Lim G, Lim ZW, et al. Deep learning in estimating prevalence and systemic risk factors for diabetic retinopathy: A multi-ethnic study. NPJ Digit Med 2019 Apr 10;2(1):24 [FREE Full text] [doi: 10.1038/s41746-019-0097-x] [Medline: 31304371]
53. Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mamonov AB, Bilbily A, Cicero M, et al. The RSNA Pediatric Bone Age Machine Learning Challenge. Radiology 2019 Feb;290(2):498-503. [doi: 10.1148/radiol.2018180736] [Medline: 30480490]
54. Iglovikov V, Rakhlin A, Kalinin A, Shvets A. Paediatric bone age assessment using deep convolutional neural networks. In: Proceedings of the 4th International Workshop on Deep Learning in Medical Image Analysis. 2018 Presented at: 4th International Workshop on Deep Learning in Medical Image Analysis; September 20, 2018; Granada, Spain. [doi: 10.1101/234120]
55. Zhao C, Han J, Jia Y, Fan L, Gou F. Versatile framework for medical image processing and analysis with application to automatic bone age assessment. J Electr Comput Eng 2018 Dec 31;2018:1-13. [doi: 10.1155/2018/2187247]

Abbreviations

BAA: bone age assessment
Caffe: Convolutional Architecture for Fast Feature Embedding
CNN: convolutional neural network

XSL•FO
RenderX

EC2: Elastic Compute Cloud
GP: Greulich-Pyle
MAE: mean absolute error
MRI: magnetic resonance imaging
RSNA: Radiological Society of North America
SLR: systematic literature review
TW: Tanner-Whitehouse
VGG: Visual Geometry Group

Edited by G Eysenbach; submitted 18.09.19; peer-reviewed by A Korchi, L Zhang, G Lim; comments to author 08.10.19; revised version received 31.10.19; accepted 13.11.19; published 05.12.19.

Please cite as:
Dallora AL, Berglund JS, Brogren M, Kvist O, Diaz Ruiz S, Dübbel A, Anderberg P
Age Assessment of Youth and Young Adults Using Magnetic Resonance Imaging of the Knee: A Deep Learning Approach
JMIR Med Inform 2019;7(4):e16291
URL: http://medinform.jmir.org/2019/4/e16291/
doi:10.2196/16291
PMID:31804183

©Ana Luiza Dallora, Johan Sanmartin Berglund, Martin Brogren, Ola Kvist, Sandra Diaz Ruiz, André Dübbel, Peter Anderberg. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 05.12.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Impact on Readmission Reduction Among Heart Failure Patients Using Digital Health Monitoring: Feasibility and Adoptability Study

Christopher Park[1], BA; Emamuzo Otobo[1], MPH, MD; Jennifer Ullman[1], NP; Jason Rogers[1], BA; Farah Fasihuddin[1], MPH; Shashank Garg[1], MS; Sarthak Kakkar[1], MS; Marni Goldstein[1], MPH; Sai Vishudhi Chandrasekhar[1]; Sean Pinney[1], MD; Ashish Atreja[1], MPH, MD

Icahn School of Medicine at Mount Sinai, New York, NY, United States

Corresponding Author:
Ashish Atreja, MPH, MD
Icahn School of Medicine at Mount Sinai
1 Gustave L Levy Pl
New York, NY, 10029
United States
Phone: 1 2122418100
Email: ashish.atreja@mssm.edu

Abstract

Background: Heart failure (HF) is a condition that affects approximately 6.2 million people in the United States and has a 5-year mortality rate of approximately 42%. With the prevalence expected to exceed 8 million cases by 2030, projections estimate that total annual HF costs will increase to nearly US $70 billion. Recently, the advent of remote monitoring technology has significantly broadened the scope of the physician's reach in chronic disease management.

Objective: The goal of our program, named the Heart Health Program, was to examine the feasibility of using digital health monitoring in real-world home settings, ascertain patient adoption, and evaluate impact on 30-day readmission rate.

Methods: A digital medicine software platform developed at Mount Sinai Health System, called RxUniverse, was used to prescribe a digital care pathway including the HealthPROMISE digital therapeutic and iHealth mobile apps to patients' personal smartphones. Vital sign data, including blood pressure (BP) and weight, were collected through an ambulatory remote monitoring system that comprised a mobile app and complementary consumer-grade Bluetooth-connected smart devices (BP cuff and digital scale) that send data to the provider care teams. Care teams were alerted via a Web-based dashboard of abnormal patient BP and weight change readings, and further action was taken at the clinicians' discretion. We used statistical analyses to determine risk factors associated with 30-day all-cause readmission.

Results: Overall, the Heart Health Program included 58 patients admitted to the Mount Sinai Hospital for HF. The 30-day hospital readmission rate was 10% (6/58), compared with the national readmission rates of approximately 25% and the Mount Sinai Hospital's average of approximately 23%. Single marital status (P=.06) and history of percutaneous coronary intervention (P=.08) were associated with readmission. Readmitted patients were also less likely to have been previously prescribed angiotensin-converting enzyme inhibitors or angiotensin II receptor blockers (P=.02). Notably, readmitted patients utilized the BP and weight monitors less than nonreadmitted patients, and patients aged younger than 70 years used the monitors more frequently on average than those aged over 70 years, though these trends did not reach statistical significance. The percentage of the 58 patients using the monitors at least once dropped from 83% (42/58) in the first week after discharge to 46% (23/58) in the fourth week.

Conclusions: Given the increasing burden of HF, there is a need for an effective and sustainable remote monitoring system for HF patients following hospital discharge. We identified clinical and social factors as well as remote monitoring usage trends that identify targetable patient populations that could benefit most from integration of daily remote monitoring. In addition, we demonstrated that interventions driven by real-time vital sign data may greatly aid in reducing hospital readmissions and costs while improving patient outcomes.

(JMIR Med Inform 2019;7(4):e13353) doi:10.2196/13353

XSL•FO
RenderX

KEYWORDS

heart failure; blood pressure; body weight; mHealth; remote consultation; patient care management; patient readmission; cell phone; mobile phone; blood pressure monitors; mobile apps

Introduction

Background

Heart failure (HF) currently affects about 6.2 million people in the United States and has an approximate 5-year mortality rate of 42% [1,2]. With the incidence rate projected to rise by 46% to exceed 8 million cases by 2030, HF is an increasing public health concern [2]. Furthermore, although admission rates for HF have declined over the past two decades [3], readmission rates have not [1,4]; Centers for Medicare and Medicaid Services (CMS) data suggest that about 25% of HF patients are readmitted within 30 days after initial hospitalization, and 35% of these readmissions are because of HF [5]. About half of discharged HF patients are readmitted within 6 months post discharge. In 2012, the HF cost burden was estimated to be US $30 billion, and projections show that by 2030, total HF costs will increase 125% from 2012 to nearly US $70 billion, which averages to US $244 for each US adult [2].

Numerous studies have been aimed at understanding the greatest risk factors for HF readmission, and results are widely varied [6-13]. The American Heart Association reports that old age, African-American race, hypertension, diabetes mellitus (DM), and low socioeconomic status were associated with higher incidence of HF [1,2]. Other reported factors include nonclinical factors such as single marital status and Medicare/Medicaid status, and clinical factors such as renal failure, chronic obstructive pulmonary disease (COPD), and history of drug use [8,9,11,14-22]. Many studies have sought to build models that use these factors to predict readmission for HF patients [14,16,17,22-26]. However, these prediction schemes have struggled to universally reduce readmission rates for HF patients because of lack of evidence and inconsistencies in determining relevant factors [1,22,26].

Owing to the advent of remote monitoring technology, recent attempts to reduce readmission rates have begun to integrate modes by which physicians directly track their patients' health statuses [27,28]. In addition to promoting a steadier communication between physician and patient, such protocols actively encourage patient participation in their own health, which has been shown to facilitate informed decision making and increase health literacy [29-31]. Studies utilizing remote monitoring for HF patients have shown varying but promising results; however, many of the reported solutions were done in research settings and required significant staff resources and costs associated with integration of telemonitoring [28,29,32,33].

Objectives

In our quality improvement (QI) initiative, named the Heart Health Program, we developed a time-efficient and relatively cost-effective remote monitoring mobile device platform, prescribed from electronic health record (EHR)–connected platform, which helped to reduce hospital readmissions among HF patients. In addition to assessing readmission rates, we aimed to identify and analyze monitor usage and adherence trends in real-world home settings to inform future interventions to be conducted on larger scales.

Methods

Participants

The program included 60 patients who were admitted to the Mount Sinai Hospital for a diagnosis of acute HF. Under an approved QI protocol, eligible patients were approached around 2 days before discharge, and interested patients were enrolled in the Heart Health Program, a digital health monitoring initiative.

Intervention Protocol

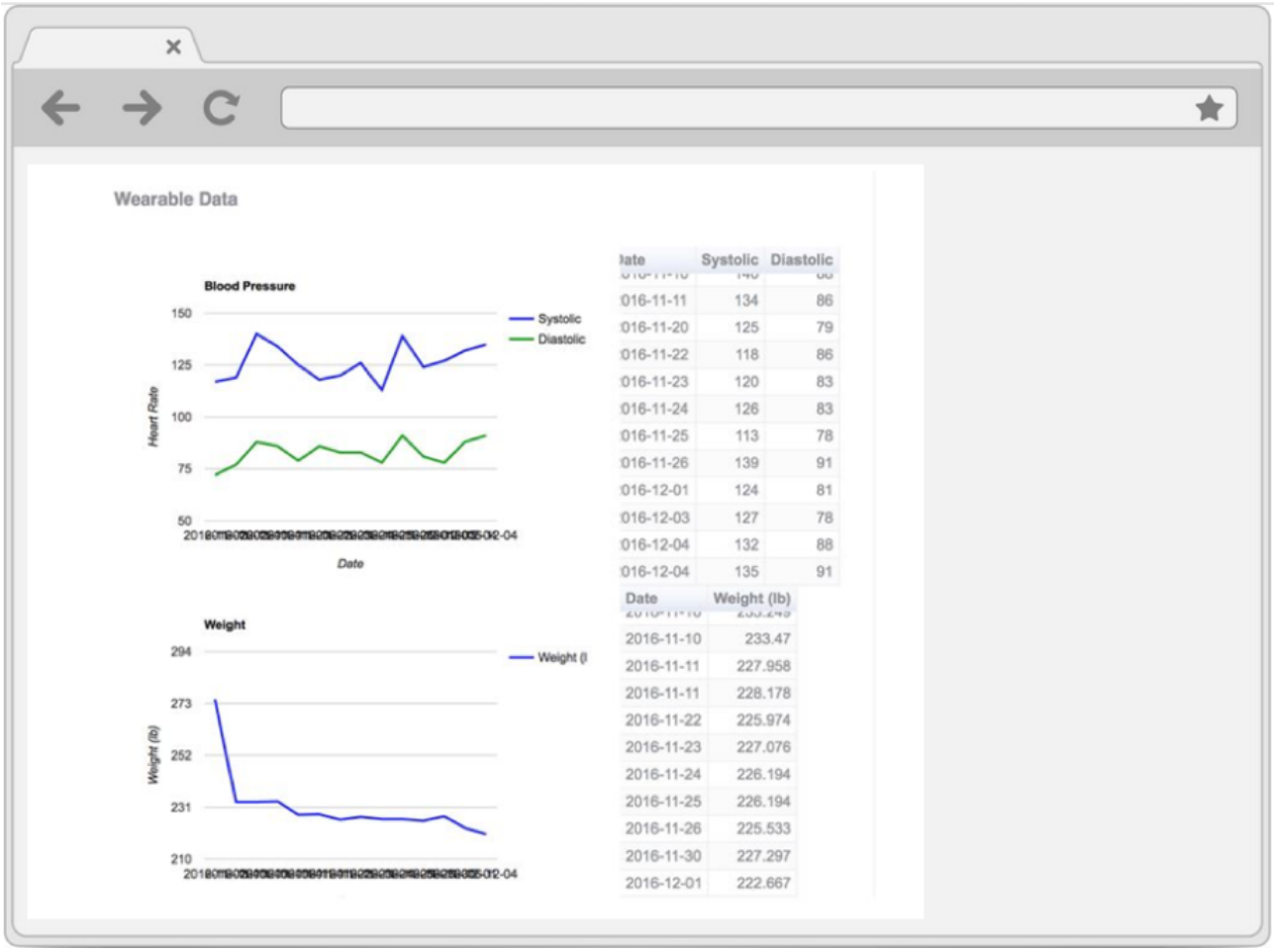
Our digital care pathway "solution toolkit" consisted of apps (HealthPROMISE digital therapeutic and iHealth mobile apps) that were prescribed using RxUniverse Digital Medicine platform (Rx.Health, Inc). RxUniverse is an EHR-integrated platform that allows digital prescription of automated care pathways including apps, digital therapeutics, education, and reminders directly to patients. This platform has been previously shown to have high usability and patient activation rate [34]. Patients were provided with HF education content, SMS or text reminders, and the ability to track their patient-reported outcomes (PROs) via preselected symptom checkboxes on HealthPROMISE digital therapeutic. In addition, each patient was provided with a complementary, consumer-grade Bluetooth-connected weight scale and blood pressure (BP) monitor that connected to the iHealth app installed on their own smartphone. Patients were instructed to measure their weight and BP each day. The total cost of equipment for each patient was US $110, paid for by the institution (Figure 1).

Electronic PRO and Bluetooth scale and BP monitor data were sent to a Web-based dashboard (Figure 2), which was monitored daily by a patient health coordinator (coauthor, EO) who worked with the clinical team to assess symptoms and respond as needed. The clinical team comprised the nurse manager for the patient floor and a cardiologist. Any critical red-flag values, for example, a greater than 2-pound weight gain within 24 hours or greater than 5-pound weight gain within a week automatically alerted the physician and prompted the patient to seek medical attention. The alerted physician could then elect to have the patient contacted to set up an appointment with the cardiologist. Physicians were also able to contact patients based on individual judgment of changes in BP or weight. This pipeline is illustrated in Figure 3. This program was exempted by the institutional review board and classified as a QI study by the Department of Medicine's QI Board.

XSL•FO
RenderX

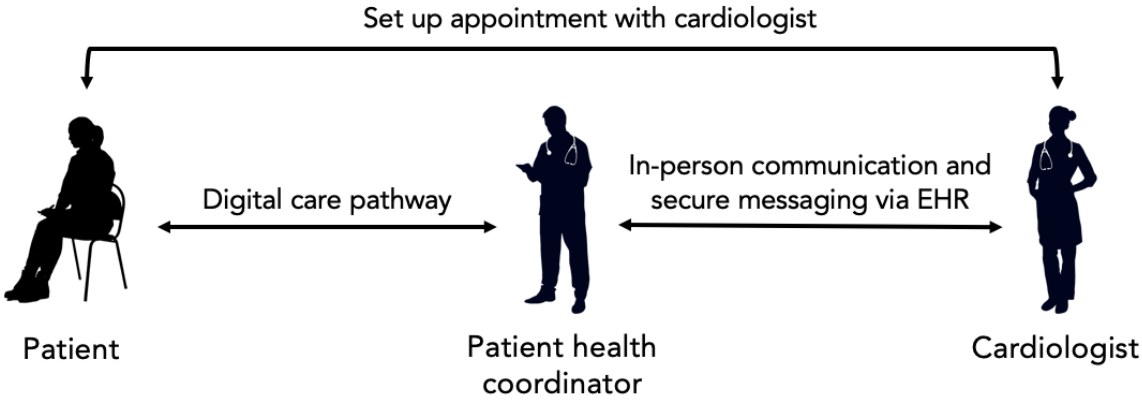
Figure 1. Digital care pathway "solution toolkit" for heart failure patients' remote monitoring.

Figure 2. Screenshot of the electronic patient-reported outcomes data dashboard.



XSL•FO
RenderX

Figure 3. Intervention protocol pipeline and roles. EHR: electronic heath record.



Dependent Variables

The primary outcome measured was 30-day all-cause readmission. In addition, usage of the monitors was defined as the number of times patients would use the weight and BP monitors per week following discharge. Usage for each monitor was calculated separately.

Covariates

All patient data were pulled from EHRs, HealthPROMISE digital therapeutic, and Bluetooth-connected devices. Demographic variables included age, sex, race, and marital status. Clinical and social factors included insurance status; previous percutaneous coronary intervention (PCI); drug abuse; prescription of angiotensin-converting enzyme inhibitor (ACEI) or angiotensin II receptor blocker (ARB); prior HF; prior hospitalization within the last 12 or 6 months; and diagnoses of systolic HF, atrial fibrillation, COPD, depression, anxiety, cancer, coronary artery disease, HIV, DM, and anemia. Factors were selected based on a comprehensive literature search of factors correlated with HF readmission.

Analysis

Readmission rates were compared with the reported Mount Sinai Hospital's standard HF all-cause 30-day readmission rate as well as the national readmission rate. Patient characteristics were compared between the readmission and nonreadmission group using *t* tests and Fischer exact tests as appropriate. Finally, BP and weight monitor usage trends were qualitatively and quantitatively analyzed based on weekly usage percentage. All statistical tests are reported with the absolute *P* values.

Results

Patient Characteristics

A total of 60 patients were enrolled in the Heart Health demonstration program. Two patients dropped out because of personal reasons. Of the remaining 58 patients in the program, 33% (n=19) were female and 67% (n=39) were male, with a median age of 62 years (Table 1). Of the 58 patients, 57 were on Medicare or Medicaid. A majority of patients had been hospitalized in the prior 12 months before discharge (60%, n=36) and were single (55%, n=32).

Readmission Statistics

Overall, there were 6 hospital readmissions (10%, 6/58) after 30 days, compared with the national 30-day readmission rate of 25%, which denotes a significant decrease (*P*=.06), and Mount Sinai Hospital's readmission rate of 23%. As shown in Table 1, readmission was associated with single status (*P*=.06), a previous PCI (*P*=.08), and no prescription of ACEI or ARB (*P*=.02).

XSL•FO
RenderX

Table 1. Patient characteristics for 30-day readmitted and nonreadmitted patients.

Characteristics	Readmitted (n=6)	Not readmitted (n=52)	*P* value
Age (years), mean (SD)	63 (9.1)	58.7 (14.1)	.37
Sex, n (%)			.38
Male	3 (50)	36 (69)	
Female	3 (50)	16 (31)	
Race, n (%)			.46
African-American	3 (50)	11 (21)	
White	1 (17)	17 (33)	
Other	2 (33)	23 (44)	
Unknown	0 (0)	1 (2)	
Marital status, n (%)			.06
Single	5 (83)	27 (52)	
Not single	0 (0)	24 (46)	
Unknown	1 (17)	1 (2)	
Systolic HF[a], n (%)	5 (83)	37 (71)	>.99
Depression, n (%)	1 (17)	4 (8)	.43
Anxiety, n (%)	1 (17)	2 (4)	.28
Cancer, n (%)	2 (33)	8 (15)	.27
Coronary artery disease, n (%)	3 (50)	17 (33)	.41
History of drug abuse, n (%)	0 (0)	1 (2)	>.99
HIV, n (%)	1 (17)	0 (0)	.10
Atrial fibrillation, n (%)	1 (17)	17 (33)	.65
Chronic obstructive pulmonary disease, n (%)	0 (0)	4 (8)	>.99
Diabetes mellitus, n (%)	3 (50)	18 (35)	.66
Anemia, n (%)	4 (67)	16 (31)	.17
Percutaneous coronary intervention, n (%)	3 (50)	8 (15)	.08
Prior HF, n (%)	4 (67)	21 (40)	.39
Prior 12 months hospitalization, n (%)	5 (83)	31 (60)	.39
Prior 6 months hospitalization, n (%)	3 (50)	14 (27)	.34
Taking angiotensin-converting enzyme inhibitor/angiotensin II receptor blocker, n (%)	3 (50)	48 (92)	.02
Medicare/aid, n (%)	6 (100)	51 (98)	>.99

[a]HF: heart failure.

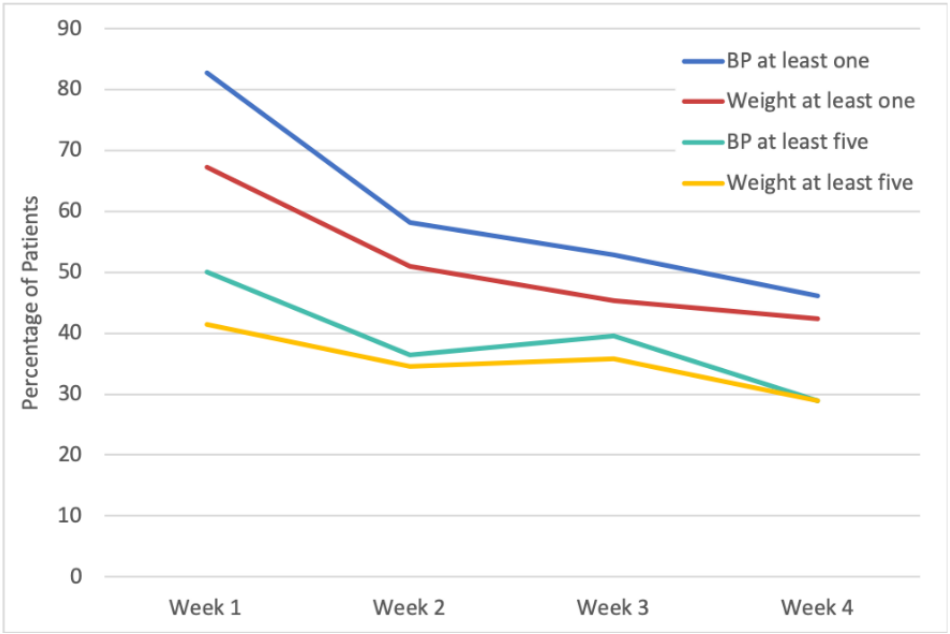
Device Usage

The percentage of the total 58 patients using the monitors at least once dropped steadily from 83% (n=48) in the first week after discharge to 46% (n=27) in the fourth week (Figure 4; ie, in the fourth week following discharge, 46% of the original 58 patients, or 27 patients, were still using the BP monitor at least once). The percentage of patients using the monitors at least five times per week dropped from around 50% (n=29) to 30% (n=17) by the fourth week. Patients, on average, used the BP monitor more often than the weight scale. Device usage was not correlated with any patient characteristic or demographic factor.

Readmitted patients utilized the BP and weight monitors less frequently than nonreadmitted patients, and patients aged less than 70 years used the monitors more frequently on average than those aged more than 70 years. Furthermore, patients were more likely to use the monitors on weekdays compared with weekends. These trends, however, did not reach statistical significance.

XSL•FO
RenderX

Figure 4. Absolute percentage of the total 58 patients with at least 1 or at least 5 blood pressure (BP) and weight checks each week following discharge.



Discussion

Overview

Readmission rates for HF patients have not decreased in the last few decades, and in fact, have risen, likely because of the decreased mortality and subsequent increased prevalence of HF patients [1]. Thus, strategies aimed at reducing readmissions are essential. The goal of this QI initiative was to assess feasibility (in part measured by readmission reduction and adherence rates) of using affordable digital health monitoring in real-world home settings, ascertain patient adoption trends, and evaluate impact on 30-day readmission rate as the primary outcome. We demonstrated that remote monitoring of BP and weight data via an app-based platform can be affordable, be adopted by patients irrespective of age, and reduce 30-day readmissions in real-world setting. We further identified usage trends that will be informative for future population health-wide implementations.

Readmission Statistics

Our program's readmission rate of 10% represented a significant decrease in 30-day readmission rate compared with both the Mount Sinai Hospital's rate (23%) as well as the national rate (25%). Readmitted patients utilized the BP and weight monitors less frequently than nonreadmitted patients, and device usage was not correlated with any patient characteristic or demographic factor. These results support the published literature on the potential role of technology-aided remote care in HF patient management [27,30-32]. However, many of the published solutions were conducted in research settings and required invasive monitor placement, significant staff resources, or expensive equipment, which may not be necessary for all HF patients [31,32,34,35].

We were able to integrate an intervention protocol that reduced barriers to adherence on both the physician side and patient side by developing mobile apps that automated many parts of the workflow: PROs and monitor data were sent directly to a dashboard via mobile apps, and the patient health coordinator would only need to notify the physician of suspicious readings or symptoms. In addition, the RxUniverse platform would automatically notify patients to enter their symptoms on the app and measure their weight and BP. Thus, we were able to remove the need for telemonitoring, frequent check-ins for every patient, and the hassle that many patients may face in recording daily symptom and weight readings manually.

Besides physician monitoring of patient health status, interventions such as ours may also play a greater role in facilitating patient engagement with their own health [36]. For example, in receiving a weight scale, patients can be educated that quick increases in weight may reflect fluid retention because of high salt intake, which can exacerbate the HF symptoms. In this way, patients may be educated on their own condition, which has been shown to decrease mortality, morbidity, and costs associated with the disease [36-38].

Device Usage

Decreased patient adherence to discharge protocols is one of the most prominent obstacles in HF patient management [39,40]. In our program, the vast majority of patients used the monitors in the week following discharge, and about half of the patients continued to measure their weight and BP at least once in the fourth week following discharge. The first week after discharge is the most vulnerable period for patients, and high engagement during that time was one of the important factors in reducing readmission rate. This drop in adherence from the first week onward may be attributable to numerous context-dependent factors, such as low motivation once patients know they are out of the "danger zone," lack of long-term nursing or social support and/or living alone, and low technological proficiency [41]. Although some of these factors were beyond the scope of our program, these data provide insight into how intervention timing may be improved. As adherence falls after the first week, perhaps a front-loaded approach to increase patient engagement (via automated SMS text messages, direct interactive voice

XSL•FO
RenderX

response calls, telemedicine and incentives when needed, etc) rather than an evenly spaced approach would increase and have a more long-lasting effect.

It is important to note that the number of patients taking BP and weight readings at least five times per week did not fall as drastically as the number of patients taking readings at least once per week. Furthermore, the majority of patients using the monitor at least five times in week 1 after discharge were the same as those in week 4. In other words, frequent monitor users are more likely to stay frequent users in the long run. Thus, patients that exhibit drastic decreases in monitor usage after the first week may represent a more specific targetable population.

In the past decade, remote monitoring, namely telemonitoring, has become a large field of interest in chronic disease management, and studies such as the Weight and Activity with Blood Pressure Monitoring System have described detailed protocols for such modalities [34]. In addition, studies that have implemented these monitoring solutions have showed great promise in reducing mortality and readmission [27]. However, many of the proposed workflows, in addition to the telemonitoring, require personnel-heavy intervention strategies, such as frequent phone calls and videoconferencing [27,29].

Many mobile apps have also been developed to help patients better manage HF, such as PatientConnect and HeartMapp [42]. Our app toolkit is similar to these apps in that it provides educational materials, care plans, reminders, in-app communication, and input options for PROs. In addition, our toolkit (1) integrates Bluetooth connectivity and automated updates of patient BP and weight data on the physician's dashboard and (2) implements newly developed digital care pathways that can be delivered to patients without the need for mobile app downloads (eg, via SMS text messages). In our Heart Health Program, we reduced readmission by exclusively using remotely collected BP and weight readings, significantly reducing the cost and staff burden compared with other more resource-heavy interventions [29,32]. Coupled with the development of cheaper Bluetooth monitors and more effective risk stratification of patients who need monitors, our approach may be scaled as larger health systems look to expand digital health monitoring enterprise-wide to all HF patients. The Heart Health Program is now being expanded across multiple health systems through a nationwide digital transformation network in partnership with the American College of Cardiology.

Limitations

Owing to our program being a demonstration program, our sample size was limited to 60 participants. Future research should use similar interventions on larger populations as well as different population demographics. Our sample size was also influenced by barriers to enrollment, including onboarding time, competition with other hospital initiatives and research trials, missed recruiting opportunities because of limited notice of discharge, language barriers, and smartphone ownership. Our program addressed these issues by enrolling HF patients around 2 days before expected discharge and adding a patient health coordinator to hospital rounds. Future research studies using device-dependent remote interventions on a larger scale could establish personnel with specific roles to address these issues and to maximize enrollment efficiency. Finally, consistent with real-world QI initiatives, specific physician-patient communications were not protocolized, and it was left to individual physicians to follow their best practices when patients' alerts were seen. Additional work could be done to integrate a mobile app–enabled protocol where physicians may report their communications with patients to gain more granular data.

Conclusions

Despite rigorous ongoing research, HF continues to contribute to increased health care costs and hospitalizations [1]. As technology advances, there is an increasing opportunity for a comprehensive, affordable, and personalized solution that can leverage existing technology to provide care to patients post discharge in real-world settings [31,43]. Recently, CMS approved reimbursement codes for remote monitoring and nonface-to-face chronic care disease management [44]. Diseases with increasing cost and health burdens such as HF could particularly benefit from such solutions. In our program, we demonstrated that interventions driven by real-time vital sign data may greatly aid in reducing hospital readmissions and costs, improving patient outcomes. Remote monitoring protocols such as ours both facilitate caregiver-patient interactions as well as engage the patient in his or her own health. Furthermore, remote monitoring programs have the potential to be scalable, especially when combined with digital medicine platforms integrated with EHRs. Future studies should seek to measure population health-wide impact on outcomes and revenue as digital health remote monitoring is expanded enterprise-wide.

Acknowledgments

This initiative is supported in part by National Institutes of Health grants (K23, NIDDK, K23DK097451-01A1: AA) and CTSA (UL1TR000067: H Sampson).

Conflicts of Interest

The RxUniverse software platform and HealthPROMISE digital therapeutic are licensed technologies from Icahn School of Medicine at Mount Sinai to Rx.Health, Inc (New York, NY). AA, SK, and JR own stock in Rx.Health and have recused themselves from data analysis. SP receives consulting fees from Abbott, CareDx, Medtronic, and Procyrion and has recused himself from data analysis.

References

XSL•FO
RenderX

1. Roger VL. Epidemiology of heart failure. Circ Res 2013 Aug 30;113(6):646-659 [FREE Full text] [doi: 10.1161/CIRCRESAHA.113.300268] [Medline: 23989710]
2. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics-2019 update: a report from the American heart association. Circulation 2019 Mar 5;139(10):e56-528. [doi: 10.1161/CIR.0000000000000659] [Medline: 30700139]
3. Chen J, Normand ST, Wang Y, Krumholz HM. National and regional trends in heart failure hospitalization and mortality rates for Medicare beneficiaries, 1998-2008. J Am Med Assoc 2011 Oct 19;306(15):1669-1678 [FREE Full text] [doi: 10.1001/jama.2011.1474] [Medline: 22009099]
4. Bueno H, Ross JS, Wang Y, Chen J, Vidán MT, Normand ST, et al. Trends in length of stay and short-term outcomes among Medicare patients hospitalized for heart failure, 1993-2006. J Am Med Assoc 2010 Jun 2;303(21):2141-2147 [FREE Full text] [doi: 10.1001/jama.2010.748] [Medline: 20516414]
5. Dharmarajan K, Hsieh AF, Lin Z, Bueno H, Ross JS, Horwitz LI, et al. Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. J Am Med Assoc 2013 Jan 23;309(4):355-363 [FREE Full text] [doi: 10.1001/jama.2012.216476] [Medline: 23340637]
6. Aizawa H, Imai S, Fushimi K. Factors associated with 30-day readmission of patients with heart failure from a Japanese administrative database. BMC Cardiovasc Disord 2015 Oct 24;15:134 [FREE Full text] [doi: 10.1186/s12872-015-0127-9] [Medline: 26497394]
7. Anderson MA, Levsen J, Dusio ME, Bryant PJ, Brown SM, Burr CM, et al. Evidenced-based factors in readmission of patients with heart failure. J Nurs Care Qual 2006;21(2):160-167. [Medline: 16540785]
8. Arora S, Lahewala S, Hassan Virk HU, Setareh-Shenas S, Patel P, Kumar V, et al. Etiologies, trends, and predictors of 30-day readmissions in patients with diastolic heart failure. Am J Cardiol 2017 Aug 15;120(4):616-624. [doi: 10.1016/j.amjcard.2017.05.028] [Medline: 28648393]
9. Chamberlain RS, Sond J, Mahendraraj K, Lau CS, Siracuse BL. Determining 30-day readmission risk for heart failure patients: the Readmission After Heart Failure scale. Int J Gen Med 2018;11:127-141 [FREE Full text] [doi: 10.2147/IJGM.S150676] [Medline: 29670391]
10. Kaneko H, Suzuki S, Goto M, Arita T, Yuzawa Y, Yagi N, et al. Incidence and predictors of rehospitalization of acute heart failure patients. Int Heart J 2015;56(2):219-225 [FREE Full text] [doi: 10.1536/ihj.14-290] [Medline: 25740584]
11. Mirkin KA, Enomoto LM, Caputo GM, Hollenbeak CS. Risk factors for 30-day readmission in patients with congestive heart failure. Heart Lung 2017;46(5):357-362. [doi: 10.1016/j.hrtlng.2017.06.005] [Medline: 28801110]
12. Ponce SG, Norris J, Dodendorf D, Martinez M, Cox B, Laskey W. Impact of ethnicity, sex, and socio-economic status on the risk for heart failure readmission: the importance of context. Ethn Dis 2018;28(2):99-104 [FREE Full text] [doi: 10.18865/ed.28.2.99] [Medline: 29725194]
13. Sherer AP, Crane PB, Abel WM, Efird J. Predicting heart failure readmissions. J Cardiovasc Nurs 2016;31(2):114-120. [doi: 10.1097/JCN.0000000000000225] [Medline: 25513988]
14. Felker GM, Leimberger JD, Califf RM, Cuffe MS, Massie BM, Adams KF, et al. Risk stratification after hospitalization for decompensated heart failure. J Card Fail 2004 Dec;10(6):460-466. [doi: 10.1016/j.cardfail.2004.02.011] [Medline: 15599835]
15. Chin MH, Goldman L. Correlates of early hospital readmission or death in patients with congestive heart failure. Am J Cardiol 1997 Jun 15;79(12):1640-1644. [doi: 10.1016/s0002-9149(97)00214-2] [Medline: 9202355]
16. Philbin EF, DiSalvo TG. Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. J Am Coll Cardiol 1999 May;33(6):1560-1566 [FREE Full text] [doi: 10.1016/s0735-1097(99)00059-5] [Medline: 10334424]
17. Krumholz HM, Chen YT, Wang Y, Vaccarino V, Radford MJ, Horwitz RI. Predictors of readmission among elderly survivors of admission with heart failure. Am Heart J 2000 Jan;139(1 Pt 1):72-77. [doi: 10.1016/s0002-8703(00)90311-9] [Medline: 10618565]
18. Amarasingham R, Moore BJ, Tabak YP, Drazner MH, Clark CA, Zhang S, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Med Care 2010 Nov;48(11):981-988. [doi: 10.1097/MLR.0b013e3181ef60d9] [Medline: 20940649]
19. Bradford C, Shah BM, Shane P, Wachi N, Sahota K. Patient and clinical characteristics that heighten risk for heart failure readmission. Res Social Adm Pharm 2017 Nov;13(6):1070-1081. [doi: 10.1016/j.sapharm.2016.11.002] [Medline: 27888091]
20. Korda RJ, Du W, Day C, Page K, Macdonald PS, Banks E. Variation in readmission and mortality following hospitalisation with a diagnosis of heart failure: prospective cohort study using linked data. BMC Health Serv Res 2017 Mar 21;17(1):220 [FREE Full text] [doi: 10.1186/s12913-017-2152-0] [Medline: 28320381]
21. Ketterer MW, Draus C, McCord J, Mossallam U, Hudson M. Behavioral factors and hospital admissions/readmissions in patients with CHF. Psychosomatics 2014;55(1):45-50. [doi: 10.1016/j.psym.2013.06.019] [Medline: 24016384]
22. Betihavas V, Davidson PM, Newton PJ, Frost SA, Macdonald PS, Stewart S. What are the factors in risk prediction models for rehospitalisation for adults with chronic heart failure? Aust Crit Care 2012 Feb;25(1):31-40. [doi: 10.1016/j.aucc.2011.07.004] [Medline: 21889893]

XSL•FO
RenderX

23. Eapen ZJ, Liang L, Fonarow GC, Heidenreich PA, Curtis LH, Peterson ED, et al. Validated, electronic health record deployable prediction models for assessing patient risk of 30-day rehospitalization and mortality in older heart failure patients. JACC Heart Fail 2013 Jun;1(3):245-251 [FREE Full text] [doi: 10.1016/j.jchf.2013.01.008] [Medline: 24621877]
24. Formiga F, Masip J, Chivite D, Corbella X. Applicability of the heart failure Readmission Risk score: a first European study. Int J Cardiol 2017 Jun 1;236:304-309. [doi: 10.1016/j.ijcard.2017.02.024] [Medline: 28407978]
25. Hernandez MB, Schwartz RS, Asher CR, Navas EV, Totfalusi V, Buitrago I, et al. Predictors of 30-day readmission in patients hospitalized with decompensated heart failure. Clin Cardiol 2013 Sep;36(9):542-547 [FREE Full text] [doi: 10.1002/clc.22180] [Medline: 23929763]
26. Ross JS, Mulvey GK, Stauffer B, Patlolla V, Bernheim SM, Keenan PS, et al. Statistical models and patient predictors of readmission for heart failure: a systematic review. Arch Intern Med 2008 Jul 14;168(13):1371-1386. [doi: 10.1001/archinte.168.13.1371] [Medline: 18625917]
27. Conway A, Inglis SC, Clark RA. Effective technologies for noninvasive remote monitoring in heart failure. Telemed J E Health 2014 Jun;20(6):531-538 [FREE Full text] [doi: 10.1089/tmj.2013.0267] [Medline: 24731212]
28. Zai AH, Ronquillo JG, Nieves R, Chueh HC, Kvedar JC, Jethwani K. Assessing hospital readmission risk factors in heart failure patients enrolled in a telemonitoring program. Int J Telemed Appl 2013;2013:305819 [FREE Full text] [doi: 10.1155/2013/305819] [Medline: 23710170]
29. Rosen D, McCall JD, Primack BA. Telehealth protocol to prevent readmission among high-risk patients with congestive heart failure. Am J Med 2017 Nov;130(11):1326-1330. [doi: 10.1016/j.amjmed.2017.07.007] [Medline: 28756266]
30. Currie K, Strachan PH, Spaling M, Harkness K, Barber D, Clark AM. The importance of interactions between patients and healthcare professionals for heart failure self-care: a systematic review of qualitative research into patient perspectives. Eur J Cardiovasc Nurs 2015 Dec;14(6):525-535. [doi: 10.1177/1474515114547648] [Medline: 25139468]
31. Kvedar J, Coye MJ, Everett W. Connected health: a review of technologies and strategies to improve patient care with telemedicine and telehealth. Health Aff (Millwood) 2014 Feb;33(2):194-199. [doi: 10.1377/hlthaff.2013.0992] [Medline: 24493760]
32. Chaudhry SI, Phillips CO, Stewart SS, Riegel B, Mattera JA, Jerant AF, et al. Telemonitoring for patients with chronic heart failure: a systematic review. J Card Fail 2007 Feb;13(1):56-62 [FREE Full text] [doi: 10.1016/j.cardfail.2006.09.001] [Medline: 17339004]
33. Bashi N, Karunanithi M, Fatehi F, Ding H, Walters D. Remote monitoring of patients with heart failure: an overview of systematic reviews. J Med Internet Res 2017 Jan 20;19(1):e18 [FREE Full text] [doi: 10.2196/jmir.6571] [Medline: 28108430]
34. Suh MK, Chen CA, Woodbridge J, Tu MK, Kim JI, Nahapetian A, et al. A remote patient monitoring system for congestive heart failure. J Med Syst 2011 Oct;35(5):1165-1179 [FREE Full text] [doi: 10.1007/s10916-011-9733-y] [Medline: 21611788]
35. Guidi G, Pollonini L, Dacso CC, Iadanza E. A multi-layer monitoring system for clinical management of congestive heart failure. BMC Med Inform Decis Mak 2015;15 Suppl 3:S5 [FREE Full text] [doi: 10.1186/1472-6947-15-S3-S5] [Medline: 26391638]
36. Cajita MI, Cajita TR, Han HR. Health literacy and heart failure: a systematic review. J Cardiovasc Nurs 2016;31(2):121-130 [FREE Full text] [doi: 10.1097/JCN.0000000000000229] [Medline: 25569150]
37. Haun JN, Patel NR, French DD, Campbell RR, Bradham DD, Lapcevic WA. Association between health literacy and medical care costs in an integrated healthcare system: a regional population based study. BMC Health Serv Res 2015 Jun 27;15:249 [FREE Full text] [doi: 10.1186/s12913-015-0887-z] [Medline: 26113118]
38. McNaughton CD, Cawthon C, Kripalani S, Liu D, Storrow AB, Roumie CL. Health literacy and mortality: a cohort study of patients hospitalized for acute heart failure. J Am Heart Assoc 2015 Apr 29;4(5) [FREE Full text] [doi: 10.1161/JAHA.115.001799] [Medline: 25926328]
39. Ding H, Jayasena R, Maiorana A, Dowling A, Chen SH, Karunanithi M, et al. Innovative Telemonitoring Enhanced Care Programme for Chronic Heart Failure (ITEC-CHF) to improve guideline compliance and collaborative care: protocol of a multicentre randomised controlled trial. BMJ Open 2017 Oct 8;7(10):e017550 [FREE Full text] [doi: 10.1136/bmjopen-2017-017550] [Medline: 28993389]
40. Stut W, Deighan C, Cleland JG, Jaarsma T. Adherence to self-care in patients with heart failure in the HeartCycle study. Patient Prefer Adherence 2015;9:1195-1206 [FREE Full text] [doi: 10.2147/PPA.S88482] [Medline: 26316725]
41. Strachan PH, Currie K, Harkness K, Spaling M, Clark AM. Context matters in heart failure self-care: a qualitative systematic review. J Card Fail 2014 Jun;20(6):448-455. [doi: 10.1016/j.cardfail.2014.03.010] [Medline: 24735549]
42. Athilingam P, Labrador MA, Remo EF, Mack L, San Juan AB, Elliott AF. Features and usability assessment of a patient-centered mobile application (HeartMapp) for self-management of heart failure. Appl Nurs Res 2016 Nov;32:156-163. [doi: 10.1016/j.apnr.2016.07.001] [Medline: 27969021]
43. Davarzani N, van Wijk SS, Maeder MT, Rickenbacher P, Smirnov E, Karel J, TIME-CHF Investigators. Novel concept to guide systolic heart failure medication by repeated biomarker testing-results from TIME-CHF in context of predictive, preventive, and personalized medicine. EPMA J 2018 Jun;9(2):161-173 [FREE Full text] [doi: 10.1007/s13167-018-0137-7] [Medline: 29896315]

XSL•FO
RenderX

44. del Valle KL, McDonnell ME. Chronic care management services for complex diabetes management: a practical overview. Curr Diab Rep 2018 Oct 20;18(12):135. [doi: 10.1007/s11892-018-1118-x] [Medline: 30343337]

Abbreviations

ACEI: angiotensin-converting enzyme inhibitor
ARB: angiotensin receptor blocker
BP: blood pressure
CMS: Centers for Medicare and Medicaid Services
COPD: chronic obstructive pulmonary disease
DM: diabetes mellitus
EHR: electronic health record
HF: heart failure
PCI: percutaneous coronary intervention
PRO: patient-reported outcome
QI: quality improvement

Edited by G Eysenbach; submitted 31.01.19; peer-reviewed by P Athilingam, A Miranda; comments to author 27.04.19; revised version received 22.06.19; accepted 19.08.19; published 15.11.19.

Please cite as:
Park C, Otobo E, Ullman J, Rogers J, Fasihuddin F, Garg S, Kakkar S, Goldstein M, Chandrasekhar SV, Pinney S, Atreja A
Impact on Readmission Reduction Among Heart Failure Patients Using Digital Health Monitoring: Feasibility and Adoptability Study
JMIR Med Inform 2019;7(4):e13353
URL: https://medinform.jmir.org/2019/4/e13353
doi:10.2196/13353
PMID:31730039

©Christopher Park, Emamuzo Otobo, Jennifer Ullman, Jason Rogers, Farah Fasihuddin, Shashank Garg, Sarthak Kakkar, Marni Goldstein, Sai Vishudhi Chandrasekhar, Sean Pinney, Ashish Atreja. Originally published in JMIR Medical Informatics (http://medinform.jmir.org), 15.11.2019. 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 http://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

XSL•FO
RenderX