JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 6 (2018), Issue 1   ISSN 2291-9694   Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Reviews

Original Papers

Review

The Use of Technology in Identifying Hospital Malnutrition: Scoping Review

Dino Trtovac[1], BA, BSc; Joon Lee[1], PhD

[1] Health Data Science Lab, School of Public Health and Health Systems, University of Waterloo, Waterloo, ON, Canada

Corresponding Author:
Joon Lee, PhD
Health Data Science Lab
School of Public Health and Health Systems
University of Waterloo
200 University Avenue West
Waterloo, ON, N2L 3G1
Canada
Phone: 1 519 888 4567 ext 31567
Email: joon.lee@uwaterloo.ca

Abstract

Background: Malnutrition is a condition most commonly arising from the inadequate consumption of nutrients necessary to maintain physiological health and is associated with the development of cardiovascular disease, osteoporosis, and sarcopenia. Malnutrition occurring in the hospital setting is caused by insufficient monitoring, identification, and assessment efforts. Furthermore, the ability of health care workers to identify and recognize malnourished patients is suboptimal. Therefore, interventions focusing on the identification and treatment of malnutrition are valuable, as they reduce the risks and rates of malnutrition within hospitals. Technology may be a particularly useful ally in identifying malnutrition due to scalability, timeliness, and effectiveness. In an effort to explore the issue, this scoping review synthesized the availability of technological tools to detect and identify hospital malnutrition.

Objective: Our objective was to conduct a scoping review of the different forms of technology used in addressing malnutrition among adults admitted to hospital to (1) identify the extent of the published literature on this topic, (2) describe key findings, and (3) identify outcomes.

Methods: We designed and implemented a search strategy in 3 databases (PubMed, Scopus, and CINAHL). We completed a descriptive numerical summary and analyzed study characteristics. One reviewer independently extracted data from the databases.

Results: We retrieved and reviewed a total of 21 articles. We categorized articles by the computerized tool or app type: malnutrition assessment (n=15), food intake monitoring (n=5), or both (n=1). Within those categories, we subcategorized the different technologies as either hardware (n=4), software (n=13), or both (n=4). An additional subcategory under software was cloud-based apps (n=1). Malnutrition in the acute hospital setting was largely an unrecognized problem, owing to insufficient monitoring, identification, and initial assessments of identifying both patients who are already malnourished and those who are at risk of malnourishment. Studies went on to examine the effectiveness of health care workers (nurses and doctors) with a knowledge base focused on clinical care and their ability to accurately and consistently identify malnourished geriatric patients within that setting.

Conclusions: Most articles reported effectiveness in accurately increasing malnutrition detection and awareness. Computerized tools and apps may also help reduce health care workers' workload and time spent assessing patients for malnutrition. Hospitals may also benefit from implementing malnutrition technology through observing decreased length of stay, along with decreased foregone costs related to missing malnutrition diagnoses. It is beneficial to study the impact of these technologies to examine possible areas of improvement. A future systematic review would further contribute to the evidence and effectiveness of the use of technologies in assessing and monitoring hospital malnutrition.

(JMIR Med Inform 2018;6(1):e4) doi:10.2196/medinform.7601

KEYWORDS

hospital malnutrition; technology-driven health care; malnutrition detection; nutrition diagnosis; malnutrition assessment; food-intake monitoring; automated data; malnutrition; nutritional status; nutrition assessment

Introduction

An inadequate diet can cause malnutrition, a condition where an individual is not consuming sufficient amounts of nutrients necessary to maintain their current level of physiological health [1].

Malnutrition increases the risk of developing certain chronic diseases, such as cardiovascular disease, osteoporosis (a debilitating loss of bone density), and sarcopenia (a debilitating loss of muscle tissue, mass, and function) [1,2]. Additionally, in hospitalized patients, malnutrition is associated with a diverse range of unfavorable outcomes, such as lowered immune function, higher infection rates, higher surgical complication rates, increased muscle loss, impaired wound healing, and overall increased morbidity and mortality [2].

Hospital malnutrition has been a historical problem as well as a current troubling issue, with documented prevalence rates as high as 60% in some hospitals [2]. Malnutrition in the acute hospital setting is suggested to be an unrecognized problem, owing to insufficient monitoring, identification, and initial assessments of patients who are already malnourished and those who are at risk [3-6]. Results suggest that health care workers' ability to accurately and consistently identify malnourished geriatric patients is suboptimal in recognizing and monitoring malnourished patients and those at risk of malnutrition [3].

Addressing, identifying, and monitoring malnutrition properly may improve the outlook for older hospitalized patients. Interventions that focus on identification and treatment of malnutrition have positive influences on decreasing the risks and rates of malnutrition in hospitals [7]. Furthermore, there is a demand for enhanced nutrition monitoring and for standardizing food intake processes early and systematically [8-10]. Other studies have found a need for development and innovation in the area of nutrition informatics [11,12]. With the current trend of technological advancement, implementation of electronic health records, and the evolving use of health data, the field is positioned to advance best practices of nutrition care to primary care settings and specifically for patients [13,14]. Consequently, how and what types of technologies are being used in this field needs to be understood in depth.

There are no known scoping reviews, to the best of our knowledge, that have synthesized the availability of technological tools to detect and identify hospital malnutrition. Therefore, we undertook a scoping review of the literature to identify the extent of the published literature on this topic, describe key findings, and identify outcomes.

Methods

We followed Arksey and O'Malley's [15] scoping review methodological framework (which is generally recognized as best practice for scoping reviews), with the following stages.

Development of Research Questions

The research questions were as follows: What is the extent of the published literature on using technology to monitor and assess malnutrition within the hospital setting? What is known from the existing literature about the impact and outcomes of implementing technology in such a manner?

Search Strategy Development

We initially developed the complete search in PubMed (Textbox 1) and then adapted it for the Scopus and CINAHL databases.

Multimedia Appendix 1 shows the search terms used in the search strategies in PubMed, Scopus, and CINAHL. The searches were conducted between August 9 and 15, 2017. Following the search of the 3 databases and article selection, we also reviewed references in the included studies to ensure that we considered all possible relevant articles.

Selection Criteria

We selected primary articles for inclusion in the scoping review if they discussed applicable technology-based approaches to monitor or assess malnutrition within the hospital or primary care setting. Articles not written in or translated into English were not included.

The definition of "applicable" included systems that were described, designed, or in current use and application at the hospital or primary care setting. We screened articles to determine relevancy of the systems.

Textbox 1. Keyword search strategy for PubMed (MeSH: Medical Subject Heading; tw: text word).

(Nutrition [tw] OR Malnutrition [tw] OR Nutritional [tw] OR Hospital malnutrition [tw] OR Dietary assessment [tw] OR Food habits [tw] OR Eating [tw] OR Diet records [tw] OR Nutritional assessment [tw] OR nutrition support [tw] OR food habits [MeSH] OR eating [MeSH] OR diet records [MeSH] OR nutritional assessment [MeSH])

AND

(Monitoring [tw] OR Screening [tw] OR Technology-based dietary assessment [tw] OR Food record [tw] OR Recording [tw] OR Assessment [tw])

AND

(Device [tw] OR informatics [tw] OR technology [tw] OR computer [tw] OR Web based [tw] OR image based [tw] OR image retrieval [tw] OR picture [tw] OR digital photography [tw] OR mobile device [tw] OR mobile technology [tw] OR smartphone [tw] OR technology assist [tw] OR multimedia tool [tw] OR electronic [tw] OR wearable [tw] OR signal processing, Computer-Assisted/instrumentation* [MeSH] OR software [MeSH] OR wireless technology/instrumentation* [MeSH])

AND

(hospital [tw] OR primary care [tw] OR care [tw])

Technology-based approaches referred to any tool or system that was technology driven or computer based, and that comprised a hardware component or a software component (including Web- and cloud-based apps). Essentially, technology-based approaches excluded any conventional nutrition support tools that only included paper-based methods traditionally used for nutrition assessment.

We excluded studies that primarily considered nontechnologically driven or noncomputer-based solutions to address hospital malnutrition. We also excluded technologically driven or computer-based solutions based in hospitals, such as hospital accounting systems or inventory systems, that had no relevance to nutrition. For the purposes of our review, we excluded studies that primarily dealt with nutrition management systems focusing on prenatal and neonatal patients. The primary aim of this scoping review was to identify technologically relevant systems for use in hospitals for adults.

Article Selection and Data Extraction

We selected articles in the following 2 phases: (1) title, abstract, and keyword review phase, and (2) full-text review phase.

Phase 1 review considered all search results. We developed a relevance form adapted from the template used by Griebel et al [16] to aid the phase 1 review (Multimedia Appendix 1). In the phase 2 review, we included articles that met the inclusion criteria from the phase 1 review. We screened and subsequently removed duplicate articles. Titles for which an abstract was not available were included for phase 2 review if the title was enough to indicate that technologically driven nutrition management tools or systems were being discussed. For those that we deemed relevant after phase 1 and 2 reviews, we obtained the full-text articles and included them in this scoping review.

Data Charting

We organized the articles according to what the tool or app measured: malnutrition assessment, food intake monitoring, or both. Within these 3 categories, we subcategorized the articles by the type of technology used: software (including cloud-based apps), hardware based (including portable devices), or both.

Collation and Summary of the Results

The goal of this scoping review was to analyze eligible articles to obtain an overview of the scientific literature on technological and computer-based approaches to nutrition monitoring and assessment in hospitals. With this goal in mind, we summarized and present the collection of key messages and concepts from eligible publications. We developed a data synthesis and characterization form to include the following study characteristics: authors; year of publication; country of origin; publication type; aims and purpose; description of the patients or participants; description of the article; type of technology used (hardware, software, or both); availability of the tool or app (theoretic, prototypic, in use, or validated); outcomes of the intervention; and cost implications (if available).

Results

Descriptive Summary

The search returned 5444 articles. We deemed a total of 21 of those articles to be relevant after phase 1 and 2 reviews. Malnutrition assessment was the most frequent aim (n=15, 71%), followed by nutrition intake monitoring (n=5, 24%), followed by an app that measured both components (n=1, 5%) (Figure 1).

Most of the malnutrition assessment articles discussed technology based on software created and incorporated within the hospital’s own computer system (n=12, 80%). Subcategorized within the software category was an article that proposed a cloud-based system for malnutrition assessment (n=1, 7%). Some studies directly used a hardware component (portable device) that assessed malnutrition (n=2, 13%).

Among the food intake monitoring articles, 2 (40%) studies contained a hardware component and 3 (60%) used technology that contained both a hardware and a software component.

One article mentioned an inclusive nutrition system that contained technology for both malnutrition assessment and food intake monitoring. Multimedia Appendix 2 details all of the primary articles that we retrieved and characterized for this review. Most of the articles were published from 2012 to 2016 (Figure 2).

The largest share of publications on the topic originated from the United States (n=7, 33%), followed by the United Kingdom (n=4, 19%), Spain (n=2, 10%), Taiwan, Philippines, the Netherlands, France, China, Australia, Argentina, and Israel, each with 1 article (n=1, 4.8%).

Most of the articles reported semiexperimental studies (n=16), followed by descriptive studies (n=3), a randomized controlled trial (RCT; n=1), and a retrospective study (n=1). Semiexperimental studies may dominate the field because they all introduced a novel malnutrition assessment or nutrition intake monitoring tool, which was primarily being tested for accuracy and effectiveness (on patients, participants, or extracted data) against universally accepted and standardized malnutrition assessment tools (eg, Mini Nutritional Assessment-Short Form [MNA-SF], Subjective Global Assessment [SGA], or Full Nutrition Assessment [FNA]) or expert opinions.

Most of the studies tested their tools and apps on patients or participants (n=14). Additionally, most of the studies reported that their tool or app was able to increase malnutrition detection or awareness (n=14). The same articles reported a degree of accuracy respective to their tool or app as determined by using validated tools (eg, MNA-SF, SGA, or FNA) or clinical nutrition expert and dietician consultations or testing (n=15).

Figure 1. Flowchart summarizing the results.

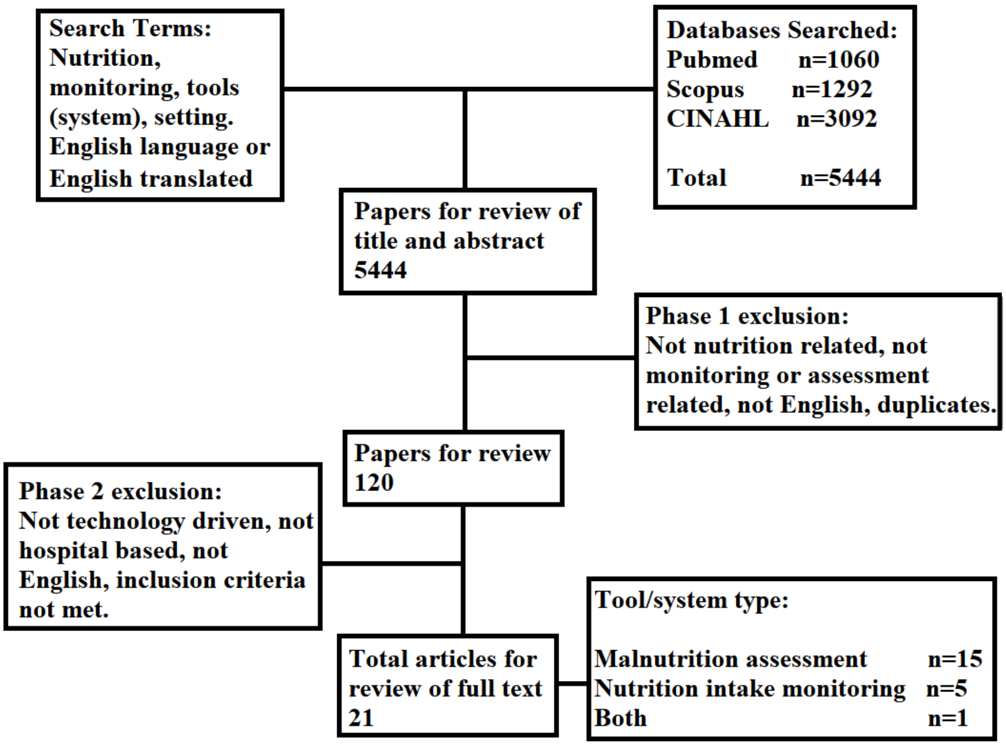
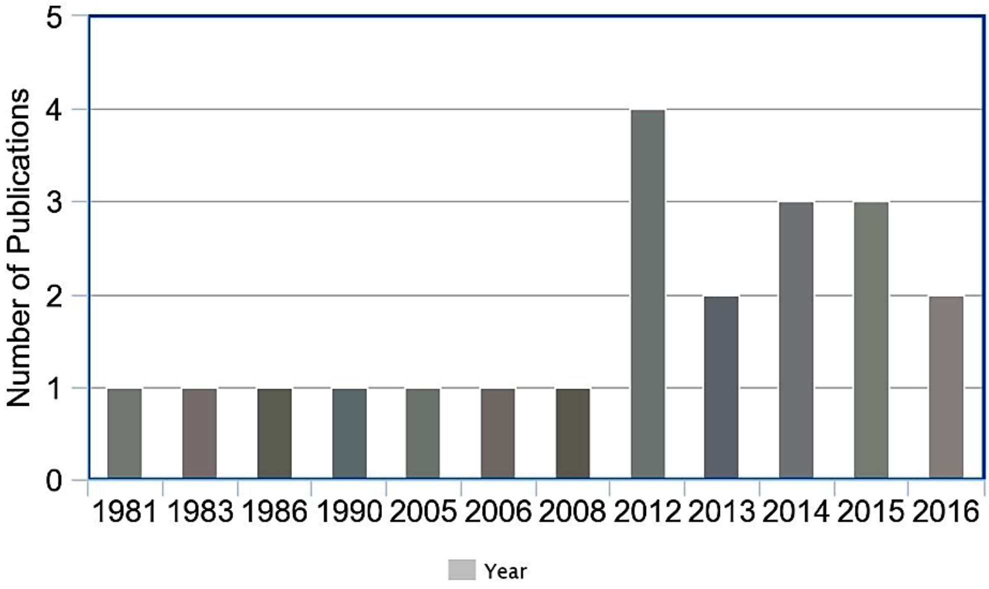

Search Terms:
Nutrition, monitoring, tools (system), setting. English language or English translated
Databases Searched:
Pubmed n=1060
Scopus n=1292
CINAHL n=3092
Total n=5444
Papers for review of title and abstract 5444
Phase 1 exclusion:
Not nutrition related, not monitoring or assessment related, not English, duplicates.
Papers for review 120
Phase 2 exclusion:
Not technology driven, not hospital based, not English, inclusion criteria not met.
Total articles for review of full text 21
Tool/system type:
Malnutrition assessment n=15
Nutrition intake monitoring n=5
Both n=1

Figure 2. Number of articles by year of publication.



Study Characteristic Analysis

Technological Interventions

A key concept mentioned in some of the articles that included a software component is automated data collection. For example, some software systems included features that allowed other departments within the hospital, such as the biochemical laboratory, to update a patient's blood test results automatically [17-19]. In addition to data obtained from laboratory results, bioelectric impedance results were automatically updated and used to assess malnourishment [20-22]. (Bioelectric impedance is a noninvasive way to measure body fluid composition using electrodes that are placed on the wrist and ankle. The results of the measurement can be used as a parameter—in conjunction with other parameters, such as height and weight—to assess whether an individual is malnourished.) The corresponding data were then added into the patient's file on the computer app, which in conjunction with biochemical markers could be used to accurately assess malnutrition.

Another feature of software apps was the ability to display a nutrition diagnosis (eg, undernourished, malnourished, low protein energy metabolism), an automated warning, or an alert once relevant data parameters were assessed. Some apps provided automated alerts that informed health care workers when a patient was assessed as being at risk for malnutrition [12,17,19,21]. Other software apps provided a malnutrition assessment when prompted, which was done by manual data entry of relevant parameters [23-25].

Only 1 study had cloud-based technological components. It allowed health care workers to create an account, log in, and begin manually entering patient data, such as biochemical markers [26]. Afterward, the system provided a corresponding nutrition diagnosis (underweight, malnourished, and 48 other diagnoses).

For the hardware component, portable devices were used to assess malnutrition risk. Largely based in the 1980s, these portable devices were pocket computers, where physical, anthropometric, and laboratory results could be manually entered to produce a malnutrition assessment [27,28]. Included in the hardware component was a wearable device, eButton, designed to record real-time food intake and to assess the malnutrition risk via direct monitoring [29]. Another study discussed the use of hardware components (an electronic weight scale, a height scale, and a bioelectric impedance spectroscopy device) that, in conjunction with a computer program, could predict a patient's likelihood of becoming malnourished [30]. Other articles discussing systems that encompassed both hardware and software components mentioned the use of a tablet that contained a "wipe-away" feature for monitoring food intake [10,30]. This concept enabled more accurate food estimations. It allowed health practitioners to wipe away corresponding food items on a tablet depiction of the food a patient was served. The software component could accurately estimate the amount of food eaten by the corresponding amount that had been erased.

Overall Outcomes

Most of the interventions had successful outcomes (n=16, 76%) [12,17-23,25-27,29,31-34]. One outcome that was mentioned in several of the articles was an improvement in time efficiency, which was facilitated by the technological implementation of the process [19,23,25-27,31,33,34]. Other articles reported that implementing a technological component improved detection rates of malnutrition within the hospital [12,17,20,22,28, 30,31,34]. Earlier detection allowed for a faster response to those patients with a diagnosis of malnutrition.

In addition to saving time and improving detection rates, 4 articles discussed the economic benefits and potential costs saved from incorporating the mentioned technology into practice [19,23,31,34]. Giovannelli et al [31] discussed how an automated email alert system for readmitted patients with previously known malnutrition decreased hospital length of stay (LOS). Rossi et al [19] expected their computer program to save an annual Aus $6500 to $10,000 in hospital costs related to malnutrition. McGurk et al [23] argued that their unique self-screening malnutrition assessment process would have the potential to reduce the workload of health care workers, thereby potentially cutting costs. Hershkovich et al [34], discussed how their hospital experienced funding cutbacks that led to a downsizing in the clinical nutrition department, which gave way to the development of the Rambam Automated Nutrition Computerized Screening (RANCS) tool. They argued that RANCS is a cost-effective option when downsizing in the clinical nutrition department is an issue.

Availability

Of the tools and apps reviewed, less than half (n=9) were available for further clinical testing and implementation within other hospital systems [17,20,23,25,27,30,31,33,34]. The remaining tools or apps were in the prototype or further testing stage [18,21,22,24,26,29,35] or were a theoretical concept [10,28], or no information was available on the tool's or app's availability [12,19,32].

Discussion

The 2 research questions formulated at the beginning of this scoping review were as follows: What is the extensiveness of the existing literature? What are the key impacts and outcomes? We discuss the outcomes in terms of effectiveness and accuracy of tools and apps; financial implications; efficiency and educational implications; and future and ethical implications.

Effectiveness and Accuracy of Tools and Apps

A total of 15 of the 21 tools and apps examined in this scoping review were effective in accurately increasing malnutrition detection and awareness using either malnutrition assessment apps or nutrition intake monitoring tools [12,19-23,26,27, 29,31-34,36].

Effectiveness was tested in a variety of ways, including evaluation studies, an RCT, and prototype testing conducted on a sample of patients or participants. The malnutrition assessment apps' effectiveness and accuracy were tested against common practice-validated nutrition screening tools used globally by hospitals, such as the MNA-SF, SGA, and FNA tools; or they were directly assessed based on expert opinions of clinical dieticians and clinical nutrition experts. The nutrition intake monitoring tools' accuracy and effectiveness were tested by comparing results from the estimated amount of food consumed with the actual amount of food consumed [29,33,36]. These results suggest that the applicability of the technological tools and apps may not *only* be effective, but they may also have a degree of accuracy similar to current and widely used conventional malnutrition assessment tools. However, only Brieux et al [12] conducted an RCT that directly assessed the effectiveness of using a software system to identify cases of malnutrition by comparing it with conventional paper-based methods. Therefore, reported improvements in accuracy and effectiveness over conventional methods from the other articles are limited by the degree of evidence due to study design methodology.

The results show that malnutrition assessment tools and apps were the dominant forms of technology used to measure malnutrition outcomes, while nutrition intake monitoring tools were used to a lesser extent. Malnutrition assessment tools and apps may dominate the literature because they may offer greater incentives and benefits than nutrition intake monitoring tools, which by their nature may require additional hospital expenditures on infrastructure (such as recording hardware or cameras). One possible incentive for using malnutrition assessment tools and apps may be their relative ease of implementation compared with nutrition intake monitoring tools. This may be due to the fact that most malnutrition assessment apps incorporated software programs, which may easily be adapted into a hospital's existing electronic infrastructure [12,17,19-23,26,31,32,34].

Some of the most simplified apps used Excel files that were tied into an algorithm developed by the authors [23,31]. McGurk et al [23] described and tested their self-screening tool that allowed all hospital inpatients to measure and enter their own anthropometric parameters (weight and height) into the Excel file. This was designed to ensure that all possible patients would be screened for malnutrition. Favorably, the results showed that there was no significant difference or errors in measurement and data entry between health care workers and patients who self-screened. Giovannelli et al [31] described an Excel spreadsheet containing a list of all previous patients who were identified as being malnourished or at risk, and another Excel spreadsheet containing a list of all inpatients admitted within the last 24 hours. The authors developed an algorithm that automatically cross-referenced the 2 spreadsheets to potentially identify a matching name that appeared on both, which automatically alerted the nutrition department in the hospital. The effort required to introduce and build new hardware components may seem discouraging; however, studies such as the latter two may seem relatively appealing to hospitals should they chose to adopt and develop similar apps. In this scoping review, it was not possible to compare the effectiveness of malnutrition assessment apps versus nutrition intake monitoring apps because the studies' sizes differed and because they measured different parameters.

Financial Implications

From an economic perspective, malnutrition is associated with increased LOS and readmission rates within hospitals [35]. Paradoxically, prolonged LOS is associated with further decline in nutritional status, a sort of malicious cycle that occurs between the two elements [8]. Therefore, a decrease in the incidence of malnutrition may decrease the economic burden incurred by hospitals. Correia and Waitzberg [37] compared and evaluated hospital-related costs between malnourished patients and nourished patients; they found that hospital-related costs were 3 times as high for the malnourished patients. Similarly, other studies had comparable findings that malnourished patients had higher hospital-related costs than did nourished patients; even patients who were at risk of malnutrition had higher associated hospital-related costs [38-40].

A total of 3 of the reviewed articles mentioned economic incentives associated with implementing technology-based approaches for malnutrition assessment [19,23,31]. Although there was no direct monetary transaction, authors argued that economic incentives were in the form of reduced future hospital costs related to malnutrition. Such savings were argued to come from reduced health care worker hours through improved computer-assisted efficiency in processes and decreased LOS experienced by patients [19,23,31].

As health care costs in North America are on the rise and since the costs relating to malnutrition account for some of that expenditure, the results suggest that implementing such technology in the hospital setting could save costs [41]. Additionally, this type of cost savings may be appealing to hospitals because it differs from cost-cutting efforts that rely on laying off health care workers.

Because older patients experience increased rates of deconditioning within the hospital, hospitals may find it beneficial to undertake efforts that identify malnutrition faster than conventional methods. Giovannelli et al [31] suggested that detecting malnutrition sooner may elicit a faster response, thereby possibly reducing hospital LOS. Furthermore, their software program showed decreased rates of LOS and improved financial impacts on the hospital's budget. Therefore, such technology may be of interest to hospitals, as it may mitigate the negative consequences associated with deconditioning and increased LOS.

The economic incentive is also extended to patients. They benefit not only physiologically from earlier malnutrition detection but also financially through decreased spending related to deconditioning outside of the hospital for homecare or nursing care [42].

In hospitals and other medical settings where there is a "pay-for-performance" infrastructure, the detection of malnutrition may result in *significantly* greater financial reimbursements. Gout et al [5] brought together multiple cases from hospitals in different countries that implemented a pay-for-performance infrastructure and described the foregone costs related to undetected malnutrition: 2 Australian studies reported an annual financial loss of Aus $1,850,540 and Aus $1,677,235, respectively, relating directly to undiagnosed and undocumented malnutrition [5]; 1 German study reported a €35,280 financial loss due to unrecognized and unidentified malnutrition [43]; and 1 US study reported a US $86,000 financial loss for a hospital in 1 year [44]. The evidence suggests that the implementation of such technology may add to the potential of hospitals to recoup any forgone revenues or reimbursements relating to unidentified cases of malnutrition.

Efficiency and Educational Implications

The results suggest that technological nutrition support tools and apps may help reduce health care workers' workload and time spent assessing patients for malnutrition [19,23,27,32-34]. This may be beneficial because these technologies can facilitate malnutrition assessment, monitoring, and awareness efforts with a smaller health care workforce or by encouraging the patients themselves to self-assess for malnutrition. Technological nutrition support tools and apps may also have an edge on conventional hospital malnutrition monitoring methods. Some of the studies showed that technological nutrition support tools and apps had the potential to improve the hospitals' efficiency and effectiveness in identifying malnutrition [19]; detect *more* cases of malnutrition [12,34]; and improve the accuracy of clinical nutritional diagnoses [26].

The results also showed that certain tools and apps can collect relevant data either automatically or by being entered manually into the system. In our review, more articles discussed tools or apps that were able to automatically collect data. Some of the studies described their nutrition support technologies as being able to monitor, assess, and identify hospital malnutrition cases relatively independently or with minimal input from health care workers [12,17,19-22,29,31,33].

Technology that has the ability to run independently or with minimal input may be beneficial in hospital areas where noise pollution is abundant, or in areas where there is relatively little time for health care workers to manually collect and analyze malnutrition information (eg, in the intensive or critical care unit). Additionally, some hospitals choose to lay off health care workers, which increases the workload and patient to nurse ratio, adding further strain on malnutrition monitoring and assessment efforts [45]. Therefore, technology that can function independently may be beneficial in these areas to mitigate the negative associations relating to an increase in patient to nurse ratios.

Furthermore, nutrition support tools and apps may have educational implications for health care workers, including nurses and doctors. Dieticians and clinical nutrition experts have more years of formal schooling and experience relating to malnutrition assessment and identification than do nurses and doctors [46]. Nutrition education has a small role in medical schools and nursing schools and is further shown to be inadequate [46]. Few tools and resources are made available to nurses or doctors to expand their practical knowledge in this area. Therefore, tools and apps that have the ability to assess and monitor patient malnutrition with a user-friendly interface may be beneficial to these health care workers. All of the articles we reviewed tested or described a nutrition support tool or app that could provide a warning, alert, risk, or predictive outcome based on its respective form of data entry. These tools may have the potential to operate in a way that's analogous to using a calculator to solve a complex math problem, where extensive nutrition knowledge is not necessary.

However, none of the articles discussed the ease of use from the user's perspective. The importance of establishing a usability rating may help determine the tool's or app's *actual* usefulness in practical settings.

Future and Ethical Implications

It is common practice that scoping reviews do not address issues of quality appraisal relating to the sources of literature used. In this scoping review, we tried to use and capture peer-based research articles that directly tested their tool or app on a cohort of patients or participants. A few articles in this scoping review contained technology in the *theoretical concept* stage [10,28]. As such, their findings are used to outline possible future directions of research within that respective domain.

Additionally, each of the reviewed research articles included different and unique types of software and hardware technologies respective to their study. Future research should seek to test the efficacy of one type of computer tool or app across different hospitals in the form of RCTs in order to contribute to a higher level of evidence for the field. It would also be interesting for future efforts to compare and test the outcomes of using malnutrition assessment tools and apps versus nutrition intake monitoring tools.

Limitations

It is likely that we excluded some relevant research articles from this review, as we screened only English-language publications. Some studies (n=8) did not pass phase 2 relevance screening because of the language restriction. These studies might have added further applications and outcomes in this scoping review. Perhaps future efforts to translate these publications will add value and knowledge to the field. Lastly, because this was a review study, we based our analysis solely on the information found in each included article, which was up-to-date only at the time of publication.

Conclusion

The use of technologies for monitoring food intake and assessing malnutrition are beginning to be considered for future hospital processes that aim to identify, diagnose, and assess hospital malnutrition. Many computerized tools and apps are being developed worldwide to address the problem of hospital malnutrition. It is beneficial to study the impact of these technologies to examine their applicability and possible areas of improvement.

Acknowledgments

This work was supported by a Discovery Grant (RGPIN-2014- 04743) from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Authors' Contributions

DT and JL designed the research; DT conducted the research; JL provided essential materials; DT analyzed the data; DT and JL wrote the paper; and DT and JL had primary responsibility for the final content. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search terms by database.

[PDF File (Adobe PDF File), 65KB - medinform_v6i1e4_app1.pdf]

Multimedia Appendix 2

Adapted data characterization form.

[PDF File (Adobe PDF File), 388KB - medinform_v6i1e4_app2.pdf]

References

1. Berner Y. Nutrition in the elderly. In: Malnick SDH, Melzer E, Tal S, editors. Gastrointestinal Tract in the Aged. Hauppauge, NY: Nova Science Publishers Inc; 2013:43-64.
2. Public Health Agency of Canada. The Chief Public Health Officer's Report on The State of Public Health in Canada 2010: Chapter 3: The Health and Well-being of Canadian Seniors. 2010 Oct 28. URL: http://www.phac-aspc.gc.ca/cphorsphc-respcacsp/2010/fr-rc/cphorsphc-respcacsp-06-eng.php [accessed 2017-02-01] [WebCite Cache ID 6nxAkN1Xe]
3. Adams NE, Bowie AJ, Simmance N, Murray M, Crowe TC. Recognition by medical and nursing professionals of malnutrition and risk of malnutrition in elderly hospitalised patients. Nutr Diet 2008 Jun;65(2):144-150. [doi: 10.1111/j.1747-0080.2008.00226.x]
4. Suominen MH, Sandelin E, Soini H, Pitkala KH. How well do nurses recognize malnutrition in elderly patients? Eur J Clin Nutr 2009 Feb;63(2):292-296. [doi: 10.1038/sj.ejcn.1602916] [Medline: 17882130]
5. Gout BS, Barker LA, Crowe TC. Malnutrition identification, diagnosis and dietetic referrals: Are we doing a good enough job? Nutrition & Dietetics 2009;66(4):206-211. [doi: 10.1111/j.1747-0080.2009.01372.x] [Medline: 22489612]
6. McWhirter JP, Pennington CR. Incidence and recognition of malnutrition in hospital. BMJ 1994 Apr 09;308(6934):945-948 [FREE Full text] [Medline: 8173401]
7. O'Flynn J, Peake H, Hickson M, Foster D, Frost G. The prevalence of malnutrition in hospitals can be reduced: results from three consecutive cross-sectional studies. Clin Nutr 2005 Dec;24(6):1078-1088. [doi: 10.1016/j.clnu.2005.08.012] [Medline: 16219393]
8. Allard JP, Keller H, Jeejeebhoy KN, Laporte M, Duerksen DR, Gramlich L, et al. Decline in nutritional status is associated with prolonged length of stay in hospitalized patients admitted for 7 days or more: a prospective cohort study. Clin Nutr 2016 Feb;35(1):144-152. [doi: 10.1016/j.clnu.2015.01.009] [Medline: 25660316]
9. Laur C, McCullough J, Davidson B, Keller H. Becoming food aware in hospital: a narrative review to advance the culture of nutrition care in hospitals. Healthcare (Basel) 2015 Jun 01;3(2):393-407 [FREE Full text] [doi: 10.3390/healthcare3020393] [Medline: 27417769]
10. Macdonald AS, Teal G, Bamford C, Moynihan PJ. Hospitalfoodie: an interprofessional case study of the redesign of the nutritional management and monitoring system for vulnerable older hospital patients. Qual Prim Care 2012;20(3):169-177. [Medline: 22828671]
11. Ayres E, Hoggle L. Advancing practice: using nutrition information and technology to improve health-the nutrition informatics global challenge. Nutr Diet 2012;69(3):195-197. [doi: 10.1111/j.1747-0080.2012.01616.x]
12. Brieux HFM, Kaminker D, Campos F, Guillen S, Alejandris J, Luna D, et al. Nutritional Alert in hospitalized patients. Stud Health Technol Inform 2014;205:697-701. [Medline: 25160276]
13. Maunder K, Williams P, Walton K, Ferguson M, Beck E, Probst Y. Introduction to nutrition informatics in Australia. Nutr Diet 2014 Oct 23;71(4):289-294. [doi: 10.1111/1747-0080.12138]
14. Probst Y. Dietitians in the electronic age: progressing towards e-health. Nutr Diet 2011;68(3):177-178. [doi: 10.1111/j.1747-0080.2011.01542.x]
15. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
16. Griebel L, Prokosch H, Köpcke F, Toddenroth D, Christoph J, Leb I, et al. A scoping review of cloud computing in healthcare. BMC Med Inform Decis Mak 2015;15:17 [FREE Full text] [doi: 10.1186/s12911-015-0145-7] [Medline: 25888747]
17. Ignacio de Ulibarri J, González-Madroño A, de Villar NGP, González P, González B, Mancha A, et al. CONUT: a tool for controlling nutritional status. First validation in a hospital population. Nutr Hosp 2005;20(1):38-45. [Medline: 15762418]
18. Siquier Homar P, Pinteño Blanco M, Calleja Hernandez MA, Fernández Cortes F, Martínez Sotelo J. [Development of integrated support software for clinical nutrition]. Farm Hosp 2015 Sep 01;39(5):240-268 [FREE Full text] [doi: 10.7399/fh.2015.39.5.8807] [Medline: 26546938]
19. Rossi M, Campbell KL, Ferguson M. Implementation of the Nutrition Care Process and International Dietetics and Nutrition Terminology in a single-center hemodialysis unit: comparing paper vs electronic records. J Acad Nutr Diet 2014 Jan;114(1):124-130. [doi: 10.1016/j.jand.2013.07.033] [Medline: 24161368]
20. Slee A, Birc D, Stokoe D. Bioelectrical impedance vector analysis, phase-angle assessment and relationship with malnutrition risk in a cohort of frail older hospital patients in the United Kingdom. Nutrition 2015 Jan;31(1):132-137. [doi: 10.1016/j.nut.2014.06.002] [Medline: 25466657]
21. Wieskotten S, Heinke S, Wabel P, Moissl U, Becker J, Pirlich M, et al. Bioimpedance-based identification of malnutrition using fuzzy logic. Physiol Meas 2008 May;29(5):639-654. [doi: 10.1088/0967-3334/29/5/009] [Medline: 18460765]
22. Zhang G, Huo X, Wu C, Zhang C, Duan Z. A bioelectrical impedance phase angle measuring system for assessment of nutritional status. Biomed Mater Eng 2014;24(6):3657-3664. [doi: 10.3233/BME-141193] [Medline: 25227080]

23. McGurk P, Jackson JM, Elia M. Rapid and reliable self-screening for nutritional risk in hospital outpatients using an electronic system. Nutrition 2013 Apr;29(4):693-696. [doi: 10.1016/j.nut.2012.12.020] [Medline: 23466054]
24. Fraser RB, Turney SZ. An expert system for the nutritional management of the critically ill. Comput Methods Programs Biomed 1990 Nov;33(3):175-180. [Medline: 2279390]
25. Vanderveen TW, Groves WE. Computerized system for a nutritional support service. Comput Methods Programs Biomed 1986 Apr;22(2):189-197. [Medline: 3635458]
26. Chen Y, Hsu C, Liu L, Yang S. Constructing a nutrition diagnosis expert system. Expert Syst Appl 2012 Feb;39(2):2132-2156. [doi: 10.1016/j.eswa.2011.07.069]
27. Rich AJ. A programable calculator system for the estimation of nutritional intake of hospital patients. Am J Clin Nutr 1981 Oct;34(10):2276-2279. [Medline: 6794348]
28. Krenning LE. Pocket computer program for performing nutritional assessment on hospitalized patients. Comput Biol Med 1983;13(4):303-308. [Medline: 6689284]
29. Sun M, Burke LE, Mao Z, Chen Y, Chen H, Bai Y, et al. eButton: a wearable computer for health monitoring and personal assistance. Proc Des Autom Conf 2014;2014:1-6 [FREE Full text] [doi: 10.1145/2593069.2596678] [Medline: 25340176]
30. Visser M, van Venrooij LMW, Wanders DCM, de Vos R, Wisselink W, van Leeuwen PAM, et al. The bioelectrical impedance phase angle as an indicator of undernutrition and adverse clinical outcome in cardiac surgical patients. Clin Nutr 2012 Dec;31(6):981-986. [doi: 10.1016/j.clnu.2012.05.002] [Medline: 22640476]
31. Giovannelli J, Coevoet V, Vasseur C, Gheysens A, Basse B, Houyengah F. How can screening for malnutrition among hospitalized patients be improved? An automatic e-mail alert system when admitting previously malnourished patients. Clin Nutr 2015 Oct;34(5):868-873. [doi: 10.1016/j.clnu.2014.09.008] [Medline: 25277380]
32. Llido LO. The impact of computerization of the nutrition support process on the nutrition support program in a tertiary care hospital in the Philippines: report for the years 2000-2003. Clin Nutr 2006 Feb;25(1):91-101. [doi: 10.1016/j.clnu.2005.08.006] [Medline: 16198450]
33. Cox Sullivan S, Bopp MM, Roberson PK, Lensing S, Sullivan DH. Evaluation of an innovative method for calculating energy intake of hospitalized patients. Nutrients 2016 Sep 09;8(9):e557 [FREE Full text] [doi: 10.3390/nu8090557] [Medline: 27618096]
34. Hershkovich S, Stark AH, Levi CS, Weiner D, Gur O, Rozen GS. A tailored automated nutrition screening tool for rapid identification of risk in acute-care hospital settings. Eur J Clin Nutr 2017 Feb;71(2):284-286. [doi: 10.1038/ejcn.2016.150] [Medline: 27507071]
35. Freijer K, Tan SS, Koopmanschap MA, Meijers JMM, Halfens RJG, Nuijten MJC. The economic costs of disease related malnutrition. Clin Nutr 2013 Feb;32(1):136-141. [doi: 10.1016/j.clnu.2012.06.009] [Medline: 22789931]
36. Comber R, Weeden J, Hoare J, Lindsay S, Teal G, Macdonald A, et al. Supporting visual assessment of foodnutrient intake in a clinical care setting. 2012 Presented at: CHI 2012. SIGCHI Conference on Human Factors in Computing Systems; May 5–10, 2012; Austin, TX p. 919-922. [doi: 10.1145/2207676.2208534]
37. Correia MITD, Waitzberg DL. The impact of malnutrition on morbidity, mortality, length of hospital stay and costs evaluated through a multivariate model analysis. Clin Nutr 2003 Jun;22(3):235-239. [Medline: 12765661]
38. León-Sanz M, Brosa M, Planas M, García-de-Lorenzo A, Celaya-Pérez S, Hernández JA. PREDyCES study: the cost of hospital malnutrition in Spain. Nutrition 2015 Sep;31(9):1096-1102. [doi: 10.1016/j.nut.2015.03.009] [Medline: 26233866]
39. Lim SL, Ong KCB, Chan YH, Loke WC, Ferguson M, Daniels L. Malnutrition and its impact on cost of hospitalization, length of stay, readmission and 3-year mortality. Clin Nutr 2012 Jun;31(3):345-350. [doi: 10.1016/j.clnu.2011.11.001] [Medline: 22122869]
40. Löser C. Malnutrition in hospital: the clinical and economic implications. Dtsch Arztebl Int 2010 Dec;107(51-52):911-917 [FREE Full text] [doi: 10.3238/arztebl.2010.0911] [Medline: 21249138]
41. Auerbach DI, Kellermann AL. A decade of health care cost growth has wiped out real income gains for an average US family. Health Aff (Millwood) 2011 Sep;30(9):1630-1636 [FREE Full text] [doi: 10.1377/hlthaff.2011.0585] [Medline: 21900652]
42. Falvey JR, Mangione KK, Stevens-Lapsley JE. Rethinking hospital-associated deconditioning: proposed paradigm shift. Phys Ther 2015 Sep;95(9):1307-1315 [FREE Full text] [doi: 10.2522/ptj.20140511] [Medline: 25908526]
43. Ockenga J, Freudenreich M, Zakonsky R, Norman K, Pirlich M, Lochs H. Nutritional assessment and management in hospitalised patients: implication for DRG-based reimbursement and health care quality. Clin Nutr 2005 Dec;24(6):913-919. [doi: 10.1016/j.clnu.2005.05.019] [Medline: 16046034]
44. Funk KL, Ayton CM. Improving malnutrition documentation enhances reimbursement. J Am Diet Assoc 1995 Apr;95(4):468-475. [doi: 10.1016/S0002-8223(95)00123-9] [Medline: 7699190]
45. Alameddine M, Baumann A, Laporte A, Deber R. A narrative review on the effect of economic downturns on the nursing labour market: implications for policy and planning. Hum Resour Health 2012 Aug 20;10:23 [FREE Full text] [doi: 10.1186/1478-4491-10-23] [Medline: 22905739]
46. Henning M. Nursing's role in nutrition. Comput Inform Nurs 2009;27(5):301-306. [doi: 10.1097/NCN.0b013e31819f7ca8] [Medline: 19726924]

Abbreviations

FNA: Full Nutrition Assessment
LOS: length of stay
MNA-SF: Mini Nutritional Assessment-Short Form
RANCS: Rambam Automated Nutrition Computerized Screening
RCT: randomized controlled trial
SGA: Subjective Global Assessment

Edited by G Eysenbach; submitted 27.02.17; peer-reviewed by J Baxter, J Albuquerque; comments to author 05.06.17; revised version received 06.10.17; accepted 30.10.17; published 19.01.18.

Please cite as:
Trtovac D, Lee J
The Use of Technology in Identifying Hospital Malnutrition: Scoping Review
JMIR Med Inform 2018;6(1):e4
URL: https://medinform.jmir.org/2018/1/e4/
doi:10.2196/medinform.7601
PMID:29351894

Original Paper

Stage-Based Mobile Intervention for Substance Use Disorders in Primary Care: Development and Test of Acceptability

Deborah Levesque[1], BA, MA, PhD; Cindy Umanzor[1], BA, MPH; Emma de Aguiar[1], BA, MPH

Pro-Change Behavior Systems, Inc, South Kingstown, RI, United States

Corresponding Author:
Deborah Levesque, BA, MA, PhD
Pro-Change Behavior Systems, Inc
1174 Kingstown Road, Suite 101
South Kingstown, RI, 02879
United States
Phone: 1 401 360 2975
Fax: 1 401 360 2975
Email: dlevesque@prochange.com

Abstract

Background: In 2016, 21 million Americans aged 12 years and older needed treatment for a substance use disorder (SUD). However, only 10% to 11% of individuals requiring SUD treatment received it. Given their access to patients, primary care providers are in a unique position to perform universal Screening, Brief Intervention, and Referral to Treatment (SBIRT) to identify individuals at risk, fill gaps in services, and make referrals to specialty treatment when indicated. Major barriers to SBIRT include limited time among providers and low motivation to change among many patients.

Objective: The objective of this study was to develop and test the acceptability of a prototype of a mobile-delivered substance use risk intervention (SURI) for primary care patients and a clinical dashboard for providers that can address major barriers to SBIRT for risky drug use. The SURI delivers screening and feedback on SUD risk via mobile tools to patients at home or in the waiting room; for patients at risk, it also delivers a brief intervention based on the transtheoretical model of behavior change (TTM) to facilitate progress through the stages of change for quitting the most problematic drug and for seeking treatment if indicated. The prototype also delivers 30 days of stage-matched text messages and 4 Web-based activities addressing key topics. For providers, the clinical dashboard summarizes the patient’s SUD risk scores and stage of change data, and provides stage-matched scripts to guide in-person sessions.

Methods: A total of 4 providers from 2 federally qualified health centers (FQHCs) were recruited for the pilot test, and they in turn recruited 5 patients with a known SUD. Furthermore, 3 providers delivered dashboard-guided SBIRT sessions and completed a brief acceptability survey. A total of 4 patients completed a Web-based SURI session and in-person SBIRT session, accessed other program components, and completed 3 acceptability surveys over 30 days. Questions in the surveys were adapted from the National Cancer Institute’s Education Materials Review Form. Response options ranged from 1=strongly disagree to 5=strongly agree. The criterion for establishing acceptability was an overall rating of 4.0 or higher across items.

Results: For providers, the overall mean acceptability rating was 4.4 (standard deviation [SD] 0.4). Notably, all providers gave a rating of 5.0 for the item, “The program can give me helpful information about my patient.” For patients, the overall mean acceptability rating was 4.5 (SD 0.3) for the mobile- and provider-delivered SBIRT sessions and 4.0 (SD 0.4) for the text messages and Web-based activities. One highly rated item was “The program could help me make some positive changes” (4.5).

Conclusions: The SURI program and clinical dashboard, developed to reduce barriers to SBIRT in primary care, were well received by providers and patients.

(JMIR Med Inform 2018;6(1):e1) doi:10.2196/medinform.7355

KEYWORDS

pilot projects; substance use disorders; primary care; behavioral medicine; expert system

Introduction

Substance Use Disorders in Primary Care

Data from the Substance Abuse and Mental Health Services Administration’s 2016 National Survey on Drug Use and Health indicate that 21.0 million Americans aged 12 years and older (8.1%) needed treatment for a substance use disorder (SUD) in the past year [1]. The annual economic costs associated with SUD are estimated at US $193 billion for illicit substance use [2], US $78.5 billion for prescription opioid misuse [3], and US $249 billion for excessive alcohol use [4] because of lost productivity, health care costs, and criminal justice costs. SUD is under-recognized and under-treated; in 2016, only 10.6% of individuals requiring treatment for an SUD received it [1]. Although only a minority of individuals with an addiction seek specialty treatment [5], an estimated two-thirds see a primary care or urgent care provider every 6 months [6]. Given their access to patients, primary care providers are in a unique position to perform Screening, Brief Intervention, and Referral to Treatment (SBIRT) to fill gaps in services and make referrals to specialty treatment when indicated [7]. SBIRT begins with universal screening using a validated screening measure to identify the level of SUD risk. For at-risk patients, screening is followed by a brief intervention tailored to the level of risk with the goal of increasing patient motivation or skills required to avoid substance use. When appropriate, brief intervention is followed by a referral to specialty care.

SBIRT has been found effective for tobacco use [8] and risky drinking [9]. However, the data on SBIRT for dependent alcohol use and for drug use are inconsistent [10]. Although 1 study found prepost reductions in alcohol and illicit drug use following SBIRT [11], a National Institute on Drug Abuse (NIDA)-funded randomized clinical trial of an SBIRT intervention—Assessing Screening Plus Brief Intervention’s Resulting Efficacy (ASPIRE) to Stop Drug Use—found no effects on any of the outcomes examined [12]. A separate study found positive effects for SBIRT in 3 countries, and a negative effect in the United States [13].

Barriers to Screening, Brief Intervention, and Referral to Treatment

Barriers to delivering SBIRT in primary care may account for some of the negative outcomes regarding its efficacy. Barriers to screening and brief intervention for SUD include time constraints [14], fear of alienating the patients [15], and the challenge of working with patients with SUD and pain [16]. The barriers to referring patients for additional evaluation or specialty treatment include patient resistance [17,18], the stigma attached to treatment [19], and limited treatment resources [16]. There are additional challenges to implementing SBIRT for drug use as opposed to alcohol use. For example, the illegal nature of drug use raises concerns by patients and providers about privacy, and a brief intervention for drug use is more complicated than one for alcohol, as different drugs and patterns of use require different types of approaches to intervention [20]. Another challenge to making SBIRT work is ensuring, postvisit, that at-risk patients engage in appropriate self-management and adhere to treatment plans and referrals. A review of studies on dropout from SUD treatment programs revealed rates of dropout ranging from 21% to 43% for detoxification, 23% to 50% for outpatient, 17% to 57% for inpatient, and 32% to 68% for substitution (eg, methadone) treatment [21]. Among individuals who initially experience progress in treatment, relapse is common [22].

A Stage-Based Mobile Intervention to Address Barriers to Screening, Brief Intervention, and Referral to Treatment

To integrate best practices and reduce barriers to SBIRT, a mobile-delivered substance use risk intervention (SURI) was developed. To address *patient barriers to SBIRT*, it was decided at the outset that SURI would be based on the transtheoretical model of behavior change (TTM), an empirically validated framework for matching interventions to readiness along a continuum of change. Behavior change involves progress through the following 5 stages: (1) precontemplation—not intending to make the behavior change in the next 6 months, (2) contemplation—intending to make the change in the next 6 months, (3) preparation—intending to make the change in the next 30 days, (4) action—made the change less than 6 months ago, and (5) maintenance—made the change more than 6 months ago. The TTM includes the following additional constructs central to change: (1) decisional balance—the pros and cons of changing [23], (2) self-efficacy—confidence to make and sustain the change in difficult situations [24], and (3) processes of change—10 cognitive, affective, and behavioral activities that facilitate progress through the stages [25,26]. More than 35 years of research on the TTM has identified particular principles and processes of change that work best in each stage to facilitate progress. The relationships between stage of change and these behavior change constructs provide an evidence-based framework for developing and delivering tailored feedback that is more likely to be remembered [27,28], considered personally relevant and credible [28-30], and to change behavior [28-30]. A meta-analysis found that health interventions tailored to stage produced significantly greater effects than those not tailored to stage [31]. A TTM approach can help facilitate progress through the stages of change for ending or reducing substance use, and for following through with treatment recommendations.

To address *system barriers to SBIRT*, it was also decided at the outset that SURI would rely on expert system technology, which could carry a significant part of the load in delivering SBIRT. Computer-tailored interventions (CTIs) based on the TTM have been found effective across a range of behaviors and populations, including smoking cessation [32], stress management [33], and depression management [34]. A recent trial of a TTM-based CTI and text messages for risky drinking found a strong effect on adherence to low-risk drinking limits (Levesque D et al, unpublished data, 2017). A CTI that shares information with the provider has the potential to also reduce barriers to communication, as individuals are more likely to disclose sensitive information to computers than to human clinicians [35,36]. The SURI prototype would include a risk assessment; a TTM-based CTI, text messages, and Web-based activities; and a clinical dashboard that summarizes the patient’s risk scores and stage of change data and provides stage-matched scripts to guide a brief in-person intervention session.

Existing Web-based and digital tools for SBIRT include provider-facing mobile apps that lead providers through an SBIRT screening [37] and, more recently, SBIRT screening tools embedded in the electronic health record (EHR) [38]. Although digital provider-facing screeners have the potential to increase provider confidence and reduce measurement error, they are time-consuming and do not address other barriers to SBIRT, such as discomfort in talking about substance use or patient resistance to change. A number of patient-facing Web-based programs and mobile apps have been developed for SUDs—most notably: (1) the *Alcohol Comprehensive Health Enhancement Support System* [39], a smartphone-based relapse prevention program that offers access to peer and professional support, reminders, education, and a Global Positioning System that identifies risky situations; (2) the *Therapeutic Education System* [40] , an interactive, Web-based psychosocial intervention with 65 interactive modules focusing on skills training; and (3) *Seva* [41], which combines the 2 programs above, and also includes a provider dashboard to help with patient monitoring. Although impressive and likely to have an impact on addictions, all 3 programs are designed for patients in recovery and are not appropriate for SBIRT.

The remainder of this manuscript describes the following steps taken to develop the stage-based SURI prototype:

1. Formative research—conducting a literature review and semistructured interviews with experts to provide guidance on the design specifications for the SURI tools
2. Intervention development—developing the intervention prototype based on the design specifications
3. Pilot testing—assessing the acceptability of the SURI tools in a pilot test involving providers and patients recruited from federally qualified health centers (FQHCs).

Formative Research

A literature review and semistructured interviews with expert consultants provided guidance on the development of the design specifications for the patient- and provider-facing SURI tools. A total of 5 experts brought expertise on SBIRT research, program development, and training; 2 experts—an SUD treatment agency chief executive officer and a health home team coordinator and peer counselor—brought expertise on the delivery of SUD specialty treatment; 1 expert was the director of a National Research Network and brought expertise on health information technology; and 1 expert brought expertise on mobile apps for substance use recovery.

Questions for the literature review and expert interviews included the following:

1. What are the barriers to delivering SBIRT in primary care?
2. How effective is SBIRT for drug use?
3. Are there any clues about “what works”?
4. *For screening*, which measures and which drugs to target?
5. *For brief intervention*, what content and what structure?
6. *For referral to treatment*, when to refer and what does referral entail?

Interviews, which lasted about 1.5 hours, were conducted by phone with 8 experts and in person with 1 expert. Examples of two key findings from the formative research and how they informed the design specifications for SURI development are as follows:

How to Select the Target Drug?

Findings from our review of 7 SBIRT outcome studies focusing on illicit drug use suggested that strategies for selecting the drug targeted in the SBIRT intervention may have an impact on outcomes. Studies that used a “flexible approach” or that allowed the patients to identify the target drug yielded negative or negligible effects [12,42,43], whereas studies that targeted a specific drug or class of drugs [44,45] or that relied at least in part on a validated risk assessment to identify the patient’s most problematic drug [13,46] yielded more positive effects. Although it is customary in SBIRT and other motivational enhancement interventions to invite patients to identify the behavior that concerns them most or that they are most ready to change, we may have a greater impact by focusing instead on the behavior causing the most harm.

For the SURI prototype, with expert guidance, we chose the World Health Organization’s (WHO) Alcohol, Smoking and Substance Involvement Screening Test (ASSIST) [47] to measure SUD risk and identify the patient’s most problematic drug. Up to 7 items are administered for each of 9 substances, with skip patterns for drugs that were never used or were not used in the past 3 months. For each substance, the ASSIST yields a risk score. Risk categories based on the scores—low (0-3), moderate (4-26), and high (27+)—have been extensively validated [47,48]. Studies have demonstrated the reliability and validity of the ASSIST administered by computer in a safety net population [49,50]—a method of administration used in a separate SBIRT intervention study that showed positive outcomes [46]. In the SURI prototype, the drug with the highest ASSIST score would be targeted in the intervention; to deal with ties, the program would use tie-breaker rules based on experts’ mean rankings of the drugs based on risks to health and well-being (opiates at the top and marijuana at the bottom).

What to do About the Highest-Risk Patients?

For individuals with a substance-specific ASSIST score in the high-risk range (27+), referral for further evaluation or specialty SUD treatment is indicated. Our review of SBIRT outcome studies revealed positive effects in 2 studies that excluded patients deemed high-risk based on the ASSIST or some other risk assessment [13,46], and negative or negligible effects in studies that included them [12,42-45]. However, among the studies that included them, protocols for treatment referral were not described at all [43,45] or were woefully inadequate, consisting only of providing the patients with a list of resources [12,42,44].

A TTM approach is ideally suited to increasing patient readiness to seek treatment, given the data on low treatment uptake [1] and high rates of dropout and drug relapse [21,22]. However, experts stressed that even as patients move forward in their readiness for change, they may not have the wherewithal to progress to action without additional help with understanding and weighing their treatment options, setting up the first appointment, and sticking to it. To address these needs, the SURI prototype would provide high-risk patients in the early stages of change for seeking treatment with information on different types of treatment and encourage patients and providers to discuss those options. For patients in the preparation stage, the program would encourage a “warm hand-off,” in which the patient and provider call the receiving agency to set up the first appointment. The SURI’s stage-matched text messages would include reminder messages in the days leading up to an appointment, which is an effective, low-cost method for increasing treatment attendance [51].

A second round of interviews with SBIRT and SUD experts provided specific ideas and language for the intervention content for each of the TTM modules. For example, for individuals in the precontemplation and contemplation stages, TTM interventions generally include a module designed to increase the “pros” or benefits of making the change—a concept that is consistent with motivational interviewing for SBIRT. SBIRT and SUD experts helped to identify the key pros of quitting a drug (eg, *so I can be a better parent* and *so I can feel more in control of my life*). Some pros were drug-specific and others were specific to the level of use.

Intervention Development

TTM-based CTIs tested in randomized trials generally include 3 CTI sessions delivered over 3 to 6 months [33,52-54] and text messages up to 6 months (Levesque D et al, unpublished data, 2017). However, the SURI prototype developed here included only the baseline CTI session and 30 days of text messages. SURI development required documenting measures and decision rules for scoring the measures and delivering tailored feedback, writing intervention content, programming the decision rules, developing the look and feel, and testing and debugging. SURI prototype components include the following:

1. *Patient-facing computer-tailored intervention*: The SURI CTI session was accessible via an Internet-enabled smartphone, computer, or tablet computer, and could be completed at home or in the primary care clinic. The session’s general session flow, designed to address all components of SBIRT—screening, brief intervention, and referral to treatment–was as follows:
    - First, assess SUD risk using the ASSIST [47] and present a chart showing the patient’s level of risk (none, low, moderate, or high) for health and other problems associated with each class of drugs assessed.
    - Second, assign the patient to an intervention track based on the most problematic drug and level of risk (moderate vs high). Screen out low-risk patients.
    - Third, inform the patient about his or her most problematic drug, and provide education on the specific health risks associated with that drug following procedures outlined in the WHO’s ASSIST manual [55].
    - Fourth, assess readiness to quit the most problematic drug and deliver a brief stage-matched intervention representing 3 key processes and principles of change for that stage, encourage the patients to set at least one stage-matched goal from the list provided, and assist in making a simple plan for goal implementation [56].
    - Finally, assess readiness to seek treatment if the ASSIST score is ≥27 for the most problematic drug, and deliver a brief stage-matched intervention that includes information about treatment options.
2. *Stage-matched text messages*: Short message service messages for 30 days were tailored to the patient’s most problematic drug, level of risk, and stage of change for quitting and for seeking treatment, if indicated. Text messages were delivered every 1 to 3 days, depending on the stage of change. One text message each week contained a link to an interactive Personal Activity Center (PAC) activity. Sample text messages for a high-risk stimulant user in the contemplation stage for quitting included the following: (1) *Is the thought of having cravings keeping you from cutting back on your stimulant use? Learn how to deal with cravings at [link to PAC activity]*, and (2) *How much do you know about stimulants? Check out drugabuse.gov. Once you learn more, you can decide if you want to cut back or stop using them.*
3. *PAC activities:* This included brief interactive activities, accessible via text messages and email, which focused on key topics (eg, dealing with cravings and working on negative thinking) for making positive changes in substance use behavior.
4. *Clinical dashboard:* This included a provider-facing tool that displayed an overview of the patient’s CTI session data and provided scripts for a brief in-person intervention session matched to the patient’s stage of change for quitting the most problematic drug and for seeking treatment, if indicated. In the SURI prototype, the dashboard was accessible via a link from the clinic EHR. Providers entered the patient’s name and date of birth to retrieve the patient’s dashboard.
5. *Printable dashboard summary*: This included a portable document format (PDF) summary of the dashboard content, along with a list of local referral resources.
6. *Patient report:* This included a PDF of all the feedback the patient received during the SURI session, along with additional questions and resources.

Program screenshots are provided in Multimedia Appendix 1. The Flesch-Kincaid reading level for patient-facing content is 5.0. All decision rules, content, and the final working prototype were reviewed by experts, revised, and rereviewed. No subjects were recruited to provide feedback during the formative research or intervention development phases of the study. Given the funding source for this study, research involving the collection of data from more than 9 respondents required clearance by the US Office of Management and Budget, and it was not practical to seek clearance in the 6-month project period. All 9 subjects, which included providers, were reserved for the pilot test described below.

Pilot Test

The remainder of this report describes a pilot test conducted to gather preliminary data on the acceptability of the mobile-delivered SURI program and clinical dashboard on a sample of 9 providers and patients. Acceptability would be established if providers and patients perceived the mobile tools as acceptable and useful—as evidenced by overall mean ratings of at least 4.0 on 5-point acceptability measures.

Providers and patients provided written informed consent for the pilot test. The Pro-Change Institutional Review Board approved the study.

Methods

Participants

Providers

As the sample size was limited to 9 subjects, including providers, it was not possible to implement universal SURI screening as a method for identifying patients at risk for SUD. Instead, 4 providers—1 from a FQHC in Georgia and 3 from a FQHC in Rhode Island—were recruited to participate in the study, and each was to recruit 1 or 2 patients. The 4 providers included a physician, 2 physician's assistants, and 1 family nurse practitioner. They had an average age of 40.5 years (standard deviation [SD] 6.8); 3 were female, 3 were white non-Hispanic, and 1 was Hispanic. The Georgia provider was unsuccessful in recruiting patients, and attributed her difficulty to her recent arrival at the FQHC; she had not yet had the opportunity to build the necessary rapport with patients. The Georgia provider participated in other study activities that did not involve patients. The 3 Rhode Island providers recruited 5 patients to yield a total study sample of n=9. The Rhode Island FQHC is a federally qualified, Joint Commission-accredited Level 3 Patient-Centered Medical Home, which offers a full range of clinical services to over 13,000 culturally diverse patients per year. Providers were offered US $450 for participating in 2 interviews, completing a brief training on the dashboard, recruiting patients, and delivering an in-person dashboard-guided session to the patients they recruited.

Patients

Providers reached out to patients by phone or during a scheduled office visit to describe the study and invite them to participate. Those interested called the study team using a toll-free number provided. The 5 patients had an average age of 41.8 years (SD 13.6), 4 were male and 1 was male-to-female transgender, 3 were white non-Hispanic and 2 were Hispanic, and all were unemployed. Patients received a total of US $110 for participating in 4 interviews, a Web-based SURI session, and an in-person dashboard-guided session with their provider.

Procedure

Providers

All 4 providers took part in an initial 30-min interview asking about barriers and facilitators to SBIRT, current clinic policies, and personal opinions and practices regarding SBIRT. They also participated in a 30-min webinar training. The 3 Rhode Island providers recruited 5 patients and completed a dashboard-guided SBIRT session with 4 of them, as 1 patient dropped out before initiating his SURI CTI session. The 3 providers also participated in a final interview that included a 6-item acceptability measure containing questions adapted from NCI's Education Materials Review Form [57] and a 5-item measure [58] that has been used to evaluate other tailored and stage-matched intervention materials [59,60]. Response options for the acceptability measure ranged from 1 (strongly disagree) to 5 (strongly agree). The criterion for establishing acceptability in the pilot test was an overall rating of 4.0 or higher across items. Providers also answered follow-up questions on what they liked most and least about the dashboard, and about EHR and clinical flow integration and training needs. NVivo software (QSR International) was used for the qualitative analysis of the interview content using node reports to identify themes and patterns in provider responses.

Patients

Before completing the SURI CTI session, the 5 patients met with a member of the project team at the clinic or by phone to answer questions on demographics and prior experience with SBIRT in primary care, and to give feedback on the intervention title, logo, and introduction screen. One patient chose to discontinue his involvement in the study before completing the interview or starting the SURI CTI session. He said that he felt the interview questions were too personal. Data from the SURI session show that 3 of the 4 patients who completed the session were polysubstance users. For 2 patients, the most problematic drug was opioids (ASSIST scores of 21 and 30), and for 2, it was cocaine (ASSIST scores of 27 and 29). Furthermore, 2 patients were in the contemplation stage, 1 was in action, and 1 in maintenance.

After completing their SURI CTI session, patients participated in an in-person SBIRT session with their provider and accessed other program components during the next 30 days. They also completed 3 acceptability surveys following the same format as the provider surveys. The first survey, administered after the SURI CTI and in-person SBIRT session, included 10 questions assessing the acceptability of the SURI CTI session. The second and third surveys, administered 2 and 4 weeks later, included 8 questions assessing the acceptability of the text messages. Patients were also asked to report what they liked most and least about the various program components and what they found most helpful about the one-on-one meeting with their provider.

Results

Providers

Providers delivered dashboard-guided intervention sessions to all pilot test participants who completed a SURI CTI session. All elements of the dashboard functioned as intended.

Acceptability of the Clinical Dashboard

On the basis of self-report, providers spent an average of 11.2 min (SD 9.5) discussing the dashboard with their patients. Table 1 shows providers' mean ratings on the 6 dashboard acceptability dimensions. The overall mean rating across items was 4.4 (SD 0.4), which exceeded the study benchmark of 4.0 for acceptability. When asked what they *liked most* about the dashboard, common themes emerged:

- All used the ASSIST scores and agreed with the program’s decision regarding the patient’s most problematic drug.
- All mentioned that the dashboard taught or gave them something new to discuss with their patient.
- Two providers mentioned that their patient was thinking about the process of quitting differently. One provider said that the CTI session had planted a seed and the text messages were helping it to germinate.
- Two providers used the dashboard as a visual aid to facilitate communication with their patient. They said it allowed the patient and the provider to start on a common ground and work toward a mutual goal.
- All stated that the dashboard gave them a clear, concise, attractive visual representation of the patient’s data, which allowed for a structured conversation focusing on the most problematic drug.

When asked what they *liked least*, 1 provider commented that the dashboard did not accurately capture his patient’s stage of change. Upon further discussion, we realized that the patient had placed herself in maintenance because she had quit using cocaine on weekdays (though still used on weekends).

Clinical Flow Integration

The dashboard was easily integrated into clinical flow at the FQHC during the pilot test, particularly for the 3 patients who had completed the SURI CTI session at home. Providers supported the idea of patients completing the CTI session at home. However, they also felt that the SURI CTI session could be administered in the waiting room on a mobile device that could be carried into the exam room, if needed. They stated that incorporating other staff to facilitate the use of the tool would be imperative.

Electronic Health Record Integration

Providers stated that the need to search for the patient’s dashboard would be a major barrier to implementation. When prompted to identify what dashboard-EHR integration would ideally look like, providers recommended the following: (1) alerts in the EHR when a new dashboard becomes available, or when an existing dashboard is updated; (2) once in a patient’s EHR, the ability to gain access to that patient’s dashboard with a single click; (3) the ability to set and track patient goals or action steps in the dashboard; and (4) the ability to save the following data to the EHR: ASSIST drug risk scores, stage of change, any action steps, and the fact that a brief intervention focusing on substance use was conducted. These recommendations aligned with those of SBIRT experts who reviewed the intervention.

Training

To use the dashboard effectively and comfortably, providers agreed that a training session is needed and suggested a 30-min session like the one they had received.

Table 1. Acceptability of the dashboard among providers.

Acceptability dimension (n=3)	Mean rating (SD[a])
The program was easy to use	4.3 (0.6)
The data were easy to understand	4.7 (0.6)
I like the way the program looked	4.3 (0.6)
The program could help my patient make some positive changes	3.7 (1.5)
The program can give me helpful information about my patient	5.0 (0.0)
I would be willing to use this program again	4.3 (0.6)

[a] SD: standard deviation.

Table 2. Acceptability of the substance use risk intervention computer-tailored intervention session among patients.

Acceptability dimension (n=4)	Mean rating (SD[a])
The program was easy to use	5.0 (0.0)
The questions were easy to understand	4.8 (0.5)
The personal feedback was easy to understand	4.5 (0.6)
I like the way the program looked	4.3 (0.5)
I felt the program respected my thoughts and point of view	4.0 (0.0)
The program gave me new things to think about	4.5 (1.0)
The program could help me make some positive changes	4.5 (0.6)
The program can give my provider helpful information about me	4.5 (0.6)
I would be willing to use this program again	4.5 (0.6)
I know someone else who could benefit from this program	4.3 (1.0)

[a] SD: standard deviation.

Patients

On the basis of the self-report and program utilization data, all 4 pilot test participants accessed all program components (SURI CTI session, text messages, and PAC activities) during the first 2 weeks of the intervention period; 3 of 4 participants accessed all available program components (text messages and PAC activities) during the final 2 weeks. All patient-facing SURI components functioned as intended.

Acceptability of Substance Use Risk Intervention Computer-Tailored Intervention Session

Table 2 shows the 10 SURI acceptability dimensions and their mean ratings among patients. The overall mean rating was 4.5 (SD 0.3), which exceeded the benchmark for acceptability. These ratings are impressive, particularly in light of the fact that there were no opportunities to elicit patient feedback on the intervention during development. All 4 patients responded to the question asking what they *liked most* about their Web-based session. For example:

> *I could do it at my own time. I could think about my answers and there is no person giving you feedback right away. There was no judgment.*

Furthermore, 2 patients responded to the question asking what they *liked least*:

> *It was little bit long. Not too many questions, but it just seemed to take a while.*

Time to Complete Substance Use Risk Intervention Computer-Tailored Intervention Session

We had expected SURI CTI sessions, such as sessions for other TTM-based CTI programs, to take about 20 min to complete. However, Google Analytics showed that sessions took an average of 35.6 min (SD 14.1).

Helpfulness of In-Person Session

A total of 3 patients responded to the question regarding what they found most helpful about their in-person session. For example, one patient stated the following:

> *Well that she did review the online questionnaire I did, and the suggestions she had really hit home: I usually don't talk to people and she suggested that I need to talk to others about it. In a way it made me a little worried but towards the end I was more comfortable.*

Acceptability of Text Messages

All participants completed the 2-week assessment examining the acceptability of the text messages, and 3 of the 4 participants completed the 4-week assessment. Patients' 2- and 4-week acceptability ratings were nearly identical, so their ratings were averaged. Table 3 shows mean ratings for the 8 acceptability dimensions among patients. The lowest mean rating was for the item, *The text messages were easy to understand* (3.6), and highest was for the item, *Reading the text messages was worth the time it took* (4.8). The overall mean rating was 4.0 (SD 0.4), which met the benchmark for acceptability. When asked what they liked most about the text messages, 1 participant responded:

> *That it was a reminder that I am doing this. I tend to forget and it's nice to have this reminder so I don't forget. And it helps me to plan my days and avoid the temptations that I can.*

When asked what they liked least, 1 participant responded:

> *I have a disability and don't always understand the questions in the text messages. It's hard to get the point of it. I can't ask a phone to explain what you mean by a text.*

On the basis of these and other responses, we will pursue the following kinds of improvements to text messages in future work: (1) conduct focus group to help ensure that text messages are understood and interpreted as intended, (2) allow patients to specify the time of day that they receive text messages, (3) allow 2-way texting, (4) provide an email option, and (5) add more links to resources—for example, information and sources of help.

Table 3. Acceptability of the substance use risk intervention text messages.

Acceptability dimension (n=4)	Mean rating (SD[a])
The text messages were easy to understand	3.6 (1.8)
The text messages reinforced things I learned in the online session	4.1 (0.9)
The text messages were supportive	3.8 (1.0)
The text messages gave me new things to think about	4.5 (0.6)
The text messages could help me make some positive changes	3.6 (1.8)
Reading the text messages was worth the time it took	4.8 (0.5)
The text messages arrived at times when it was good for me to receive them	4.0 (1.2)
I know someone else who could benefit from messages like these	3.7 (1.5)

[a] SD: standard deviation.

A Follow-Up Message From a Participant

The patient who did not participate in the 4-week survey contacted the study's project manager approximately 4 months later and gave his consent for us to share his message:

> *I found your number in an email and I wanted to let you know that I have been sober for 7 weeks. It's the longest I've been off opioids besides the year I was in prison. It's amazing. I'm sorry I didn't complete the final activity. I wanted to know if I could still complete it. I met with [provider] today and she didn't think I could do it. But I did it. I can only move up from here.*

Although it is impossible to attribute this patient's positive changes to his involvement in the pilot test, it is interesting to note that he chose to share his success with both his provider and a member of the study team.

Discussion

Principal Findings

To help address the barriers to SBIRT, the SURI tools were designed to: (1) reduce provider time and burden, (2) facilitate patient-provider communication, (3) facilitate evidence- and risk-based decision making that accommodates a range of drugs and patterns of use, (4) increase provider adherence to best practices, (5) increase provider comfort and confidence, and (6) facilitate patient readiness to quit their most problematic drug and, if indicated, to follow through with treatment recommendations. All program components functioned as intended, and SURI program acceptability ratings from both providers and patients met or exceeded the criteria for establishing acceptability. Patient ratings are especially impressive, in light of the fact that there were no opportunities to elicit patient feedback on the intervention during development.

It is noteworthy that the provider who was unsuccessful in recruiting patients for the pilot test attributed the problem to her lack of rapport with patients, as she was relatively new to her clinic. And the patient who chose to discontinue his involvement during the first interview attributed his decision to discomfort with the study questions. No doubt, discussing substance use—and especially drug use—can be uncomfortable for both providers and patients. Although this discomfort may serve as a barrier to using the SURI tools, it is also possible that relying on SURI to introduce and deliver universal screening and a brief intervention via mobile tools, as a part of routine care, can reduce stigma and open a channel for patient-provider communication focused on patient health and well-being. A SURI demonstration project that allows universal screening is required to assess the program's acceptance and uptake among providers and patients under typical clinic conditions.

Substance Use Risk Intervention Program Enhancements

Only a prototype of the SURI program was developed here. Steps to complete intervention development would include the following: (1) revising intervention content and procedures based on the current findings and recommendations; (2) developing content for the second and third SURI CTI sessions; (3) writing an additional 5 months of text messages and additional PAC activities; and (4) evaluating all content for reading level (ensuring grade 5) and cultural sensitivity and revising as necessary. As Hispanic Americans comprise about 17% of the US population [61] and 34% of FQHC patients [62], translation of all patient-facing components into Spanish would be necessary to increase SURI program accessibility and disseminability.

The 2 follow-up CTI sessions would be similar in flow and structure to the baseline session described above. However, follow-up sessions would also inform participants on how they have changed on the following 4 key dimensions: (1) the drug identified as most problematic, (2) level of risk for that drug, (3) stage of change for quitting that drug, and (4) stage of change for seeking treatment, if indicated. Guidance delivered in the CTI sessions and text messages, and content in the PAC activities, would be matched to the updated data regarding the most problematic drug, risk level, and stage of change. These updated data would also be shared with the provider via the dashboard and trigger updated stage-matched scripts for new one-on-one sessions. The 3 SURI CTI sessions could be delivered over 3 or 6 months.

Moving forward, a more participatory approach will be required to refine and enhance the intervention based on the pilot test findings, and to develop the remainder of the intervention package. Patient focus groups in future research will review the operational definition of "quit," and will review all other intervention content to ensure it is interpreted as intended. Usability tests using a variety of devices (smartphone, tablet computer, and computer) will ensure that all program components are easily navigable, and will examine how long the SURI CTI session takes to complete and how participants spend their time. In addition, interviews with patients, experts, and providers can help to identify an acceptable session length for clinic administration (we suspect 20 min), and whether to scale back on intervention content to reduce session length—for example, by reserving some SURI content for PAC activities or follow-up SURI sessions.

In the research described above, a final round of interviews was conducted with 7 experts to outline plans for integrating the mobile tools with one or more EHR products and within clinical practice. Experts agreed that SURI integration with the EHR was essential, and their vision of what that should look like matched that of providers as described above. Experts also identified the following program features and functions deemed necessary to maximize the program's usability, impact, and likelihood of clinic-wide adoption in a large randomized trial and, eventually, under real-world conditions:

1. Allow the provider to *override SURI decision rules* regarding the most problematic drug and need for treatment. For example, a provider may want SURI to focus on a different drug that poses a great risk for the patient, given a cooccurring medical problem, or want a patient who falls below the ASSIST treatment cut point to seek treatment. We will also allow the provider to request that the patient's

stage of change be reevaluated by the SURI program. This request will trigger a text message linking the patient to a brief stage assessment.

2. Allow the provider to *select specific action steps* that could be communicated to SURI (eg, cut back on drug X by Y amount).
3. Program SURI to *monitor patients' progress on action steps* and *communicate that progress back to the clinic*. Adhering to patient privacy rules, SURI would not communicate with outside treatment providers or care organizations. Rather, it would rely on text messaging to elicit patient reports on their progress on action steps, and share patient responses with a designated patient care manager at the clinic, who could then take any appropriate steps required.

Features 2 and 3 above deepen SURI's integration with clinical practice and have the potential to increase SBIRT's impact, particularly among the highest-risk patients requiring further SUD evaluation or treatment.

Limitations

The pilot test was small and conducted under ideal conditions. It would be unrealistic to expect that level of enthusiasm and adherence when the intervention is rolled out in a large clinical trial to assess its efficacy or in the real world. Future research involving clinic-wide implementation would require a "make it happen" approach [63] to implementation. Making it happen would need to involve several best practices from implementation science, such as: (1) ensuring buy-in from leadership [64]; (2) assembling members of an implementation team [64] at each site to define site goals for screening and SBIRT delivery (percentage of patients screened and percentage of eligible patients who receive an in-person dashboard-guided session); (3) assembling members of an implementation team [64] at each site to outline practicable procedures that will support universal screening, timely in-person SBIRT sessions, and appropriate follow-up from a care manager; (4) identifying and training a champion at each site who can serve as a role model for the implementation [64]; (5) providing training to staff and providers; and (6) using plan-do-study-act cycles [65,66] once the implementation has started—in this case, using scores on key metrics to guide improvements in implementation procedures and outcomes over time, in an iterative fashion. Key metrics could include provider acceptability ratings; number of SBIRT screenings completed each week; number of moderate- and high-risk patients identified via screening; and among the patients identified, the percentage receiving an in-person SBIRT session. Low ratings at a given site could trigger an exploration of problems and barriers at the site, as well as efforts to find solutions.

Conclusions

This study represents a large step forward in the development of mobile tools that have the potential to reduce major barriers to SBIRT. The SURI program's particular combination of features, along with future enhancements and efforts to integrate SURI into clinical flow and the EHR, will be uniquely designed to help providers deliver all 3 elements of SBIRT—screening, brief intervention, and referral to treatment—with efficiency and adherence to evidence-based practices—within a busy primary care setting.

Acknowledgments

This study was funded by the National Institute on Drug Abuse (NIDA Contract No: N43DA-15-2242).

Conflicts of Interest

All 3 authors are employees of Pro-Change Behavior Systems, Inc.

Multimedia Appendix 1

SURI (Substance Use Risk Intervention) program screenshots.

[PDF File (Adobe PDF File), 580KB - medinform_v6i1e1_app1.pdf]

References

1. Substance Abuse and Mental Health Services Administration. SAMHSA. Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration; 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.pdf [accessed 2017-12-11] [WebCite Cache ID 6vdnDdag3]
2. National Drug Intelligence Center. Justice. Johnstown, PA: National Drug Intelligence Center; 2011. The economic impact of illicit drug use on American society URL: https://www.justice.gov/archive/ndic/pubs44/44731/44731p.pdf [accessed 2017-12-11] [WebCite Cache ID 6tJ8f9NZk]
3. Florence CS, Zhou C, Luo F, Xu L. The economic burden of prescription opioid overdose, abuse, and dependence in the United States, 2013. Med Care 2016 Oct;54(10):901-906. [doi: 10.1097/MLR.0000000000000625] [Medline: 27623005]
4. Sacks JJ, Gonzales KR, Bouchery EE, Tomedi LE, Brewer RD. 2010 national and state costs of excessive alcohol consumption. Am J Prev Med 2015;49(5):e73-e79. [doi: 10.1016/j.amepre.2015.05.031] [Medline: 26477807]

5. Cherpitel CJ, Ye Y. Drug use and problem drinking associated with primary care and emergency room utilization in the US general population: data from the 2005 national alcohol survey. Drug Alcohol Depend 2008 Oct 1;97(3):226-230 [FREE Full text] [doi: 10.1016/j.drugalcdep.2008.03.033] [Medline: 18499355]
6. Bowman S, Eiserman J, Beletsky L, Stancliff S, Bruce RD. Reducing the health consequences of opioid addiction in primary care. Am J Med 2013;126(7):565-571. [doi: 10.1016/j.amjmed.2012.11.031] [Medline: 23664112]
7. Ghitza UE, Tai B. Challenges and opportunities for integrating preventive substance-use-care services in primary care through the Affordable Care Act. J Health Care Poor Underserved 2014;25(1 Suppl):36-45 [FREE Full text] [doi: 10.1353/hpu.2014.0067] [Medline: 24583486]
8. Land TG, Rigotti NA, Levy DE, Schilling T, Warner D, Li W. The effect of systematic clinical interventions with cigarette smokers on quit status and the rates of smoking-related primary care office visits. PLoS One 2012;7(7):e41649 [FREE Full text] [doi: 10.1371/journal.pone.0041649] [Medline: 22911834]
9. Jonas D, Garbutt J, Amick H, Brown J, Brownley K, Council C, et al. Behavioral counseling after screening for alcohol misuse in primary care: a systematic review and meta-analysis for the U.S. Preventive Services Task Force. Ann Intern Med 2012;157(9):645-654. [Medline: 22876371]
10. Hingson R, Compton WM. Screening and brief intervention and referral to treatment for drug use in primary care: back to the drawing board. J Am Med Assoc 2014 Aug 06;312(5):488-489. [doi: 10.1001/jama.2014.7863] [Medline: 25096687]
11. Madras BK, Compton WM, Avula D, Stegbauer T, Stein JB, Clark HW. Screening, brief interventions, referral to treatment (SBIRT) for illicit drug and alcohol use at multiple healthcare sites: comparison at intake and 6 months later. Drug Alcohol Depend 2009 Jan 01;99(1-3):280-295 [FREE Full text] [doi: 10.1016/j.drugalcdep.2008.08.003] [Medline: 18929451]
12. Saitz R, Palfai TP, Cheng DM, Alford DP, Bernstein JA, Lloyd-Travaglini CA, et al. Screening and brief intervention for drug use in primary care: the ASPIRE randomized clinical trial. J Am Med Assoc 2014;312(5):502-513 [FREE Full text] [doi: 10.1001/jama.2014.7862] [Medline: 25096690]
13. Humeniuk R, Ali R, Babor T, Souza-Formigoni ML, de Lacerda RB, Ling W, et al. A randomized controlled trial of a brief intervention for illicit drugs linked to the Alcohol, Smoking and Substance Involvement Screening Test (ASSIST) in clients recruited from primary health-care settings in four countries. Addiction 2012;107(5):957-966. [doi: 10.1111/j.1360-0443.2011.03740.x] [Medline: 22126102]
14. Muench J, Jarvis K, Vandersloot D, Hayes M, Nash W, Hardman J, et al. Perceptions of clinical team members toward implementation of SBIRT processes. Alcohol Treat Q 2015;33(2):143-160. [doi: 10.1080/07347324.2015.1018775]
15. Friedmann PD, McCullough D, Saitz R. Screening and intervention for illicit drug abuse: a national survey of primary care physicians and psychiatrists. Arch Intern Med 2001;161(2):248-251. [Medline: 11176739]
16. Urada D, Teruya C, Gelberg L, Rawson R. Integration of substance use disorder services with primary care: health center surveys and qualitative interviews. Subst Abuse Treat Prev Policy 2014;9:15 [FREE Full text] [doi: 10.1186/1747-597X-9-15] [Medline: 24679108]
17. Rapp RC, Li L, Siegal HA, DeLiberty RN. Demographic and clinical correlates of client motivation among substance abusers. Health Soc Work 2003 May;28(2):107-115. [Medline: 12774532]
18. Boyle K, Polinsky ML, Hser Y. Resistance to drug abuse treatment: a comparison of drug users who accept or decline treatment referral assessment. J Drug Issues 2000 Jul;30(3):555-574. [doi: 10.1177/002204260003000304]
19. Rahm AK, Boggs JM, Martin C, Price DW, Beck A, Backer TE, et al. Facilitators and barriers to implementing screening, brief intervention, and referral to treatment (SBIRT) in primary care in integrated health care settings. Subst Abus 2015;36(3):281-288. [doi: 10.1080/08897077.2014.951140] [Medline: 25127073]
20. Saitz R, Alford DP, Bernstein J, Cheng DM, Samet J, Palfai T. Screening and brief intervention for unhealthy drug use in primary care settings: randomized clinical trials are needed. J Addict Med 2010 Sep;4(3):123-130 [FREE Full text] [doi: 10.1097/ADM.0b013e3181db6b67] [Medline: 20936079]
21. Brorson HH, Ajo AE, Rand-Hendriksen K, Duckert F. Drop-out from addiction treatment: a systematic review of risk factors. Clin Psychol Rev 2013 Dec;33(8):1010-1024. [doi: 10.1016/j.cpr.2013.07.007] [Medline: 24029221]
22. McLellan AT, Lewis DC, O'Brien CP, Kleber HD. Drug dependence, a chronic medical illness: implications for treatment, insurance, and outcomes evaluation. J Am Med Assoc 2000 Oct 04;284(13):1689-1695. [Medline: 11015800]
23. Janis I, Mann L. Decision making: A psychological analysis of conflict, choice, and commitment. New York: Free Press; 1977.
24. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev 1977 Mar;84(2):191-215. [Medline: 847061]
25. Prochaska J, DiClemente C. The transtheoretical approach: Crossing traditional boundaries of change. Homewood, IL: Dorsey Press; 1984.
26. Prochaska J, DiClemente C. Common processes of self-change in smoking, weight control, and psychological distress. In: Shiffman S, Wills T, editors. Coping and substance abuse: A conceptual framework. New York: Academic Press; 1985:345-363.
27. Brug J, Steenhuis I, van Assema P, Glanz K, de Vries H. Computer-tailored nutrition education: differences between two interventions. Health Educ Res 1999 Apr;14(2):249-256. [Medline: 10387504]

28. Kreuter MW, Strecher VJ, Glassman B. One size does not fit all: the case for tailoring print materials. Ann Behav Med 1999;21(4):276-283. [Medline: 10721433]
29. Johnson SS, Driskell M, Johnson JL, Dyment SJ, Prochaska JO, Prochaska JM, et al. Transtheoretical model intervention for adherence to lipid-lowering drugs. Dis Manag 2006 Apr;9(2):102-114. [doi: 10.1089/dis.2006.9.102] [Medline: 16620196]
30. Prochaska JO. Treating entire populations for behavior risks for chronic diseases. Homeost Health Dis 2003;42(1-2):1-12.
31. Noar SM, Benac CN, Harris MS. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. Psychol Bull 2007 Jul;133(4):673-693. [doi: 10.1037/0033-2909.133.4.673] [Medline: 17592961]
32. Prochaska JO, DiClemente CC, Velicer WF, Rossi JS. Standardized, individualized, interactive, and personalized self-help programs for smoking cessation. Health Psychol 1993 Sep;12(5):399-405. [Medline: 8223364]
33. Evers KE, Prochaska JO, Johnson JL, Mauriello LM, Padula JA, Prochaska JM. A randomized clinical trial of a population- and transtheoretical model-based stress-management intervention. Health Psychol 2006 Jul;25(4):521-529. [doi: 10.1037/0278-6133.25.4.521] [Medline: 16846327]
34. Levesque DA, Van Marter DF, Schneider RJ, Bauer MR, Goldberg DN, Prochaska JO, et al. Randomized trial of a computer-tailored intervention for patients with depression. Am J Health Promot 2011;26(2):77-89. [doi: 10.4278/ajhp.090123-QUAN-27] [Medline: 22040388]
35. Couper M, Tourangeau R, Marvin T. Taking the audio out of audio-CASI. Public Opin Q 2009;73(2):281-303. [doi: 10.1093/poq/nfp025]
36. Tourangeau R, Smith R. Asking sensitive questions: the impact of data collection mode, question format, and question format. Public Opin Q 1996;60(2):275-304. [doi: 10.1086/297751]
37. Bray J. itunes.apple. SBIRT (iTunes app) URL: https://itunes.apple.com/us/app/sbirt/id877624835?mt=8 [accessed 2017-09-07] [WebCite Cache ID 6tJ8u2xyE]
38. Press A, DeStio C, McCullagh L, Kapoor S, Morley J, SBIRT NY-II Team, et al. Usability testing of a National Substance Use Screening Tool embedded in electronic health records. JMIR Hum Factors 2016 Jul 08;3(2):e18 [FREE Full text] [doi: 10.2196/humanfactors.5820] [Medline: 27393643]
39. Gustafson DH, McTavish FM, Chih M, Atwood AK, Johnson RA, Boyle MG, et al. A smartphone application to support recovery from alcoholism: a randomized clinical trial. JAMA Psychiatry 2014 May;71(5):566-572 [FREE Full text] [doi: 10.1001/jamapsychiatry.2013.4642] [Medline: 24671165]
40. Marsch LA, Guarino H, Acosta M, Aponte-Melendez Y, Cleland C, Grabinski M, et al. Web-based behavioral treatment for substance use disorders as a partial replacement of standard methadone maintenance treatment. J Subst Abuse Treat 2014 Jan;46(1):43-51 [FREE Full text] [doi: 10.1016/j.jsat.2013.08.012] [Medline: 24060350]
41. Quanbeck AR, Gustafson DH, Marsch LA, McTavish F, Brown RT, Mares M, et al. Integrating addiction treatment into primary care using mobile health technology: protocol for an implementation research study. Implement Sci 2014 May 29;9(65) [FREE Full text] [doi: 10.1186/1748-5908-9-65] [Medline: 24884976]
42. Bogenschutz MP, Donovan DM, Mandler RN, Perl HI, Forcehimes AA, Crandall C, et al. Brief intervention for patients with problematic drug use presenting in emergency departments: a randomized clinical trial. JAMA Intern Med 2014 Nov;174(11):1736-1745 [FREE Full text] [doi: 10.1001/jamainternmed.2014.4052] [Medline: 25179753]
43. Roy-Byrne P, Bumgardner K, Krupski A, Dunn C, Ries R, Donovan D, et al. Brief intervention for problem drug use in safety-net primary care settings: a randomized clinical trial. J Am Med Assoc 2014 Aug 06;312(5):492-501 [FREE Full text] [doi: 10.1001/jama.2014.7860] [Medline: 25096689]
44. Bernstein J, Bernstein E, Tassiopoulos K, Heeren T, Levenson S, Hingson R. Brief motivational intervention at a clinic visit reduces cocaine and heroin use. Drug Alcohol Depend 2005 Jan 07;77(1):49-59. [doi: 10.1016/j.drugalcdep.2004.07.006] [Medline: 15607841]
45. Zahradnik A, Otto C, Crackau B, Löhrmann I, Bischof G, John U, et al. Randomized controlled trial of a brief intervention for problematic prescription drug use in non-treatment-seeking patients. Addiction 2009 Jan;104(1):109-117. [doi: 10.1111/j.1360-0443.2008.02421.x] [Medline: 19133895]
46. Gelberg L, Andersen RM, Afifi AA, Leake BD, Arangua L, Vahidi M, et al. Project QUIT (Quit Using Drugs Intervention Trial): a randomized controlled trial of a primary care-based multi-component brief intervention to reduce risky drug use. Addiction 2015 Nov;110(11):1777-1790 [FREE Full text] [doi: 10.1111/add.12993] [Medline: 26471159]
47. Humeniuk R, Ali R, Babor TF, Farrell M, Formigoni ML, Jittiwutikarn J, et al. Validation of the Alcohol, Smoking And Substance Involvement Screening Test (ASSIST). Addiction 2008 Jun;103(6):1039-1047. [doi: 10.1111/j.1360-0443.2007.02114.x] [Medline: 18373724]
48. Newcombe DA, Humeniuk RE, Ali R. Validation of the World Health Organization Alcohol, Smoking and Substance Involvement Screening Test (ASSIST): report of results from the Australian site. Drug Alcohol Rev 2005 May;24(3):217-226. [doi: 10.1080/09595230500170266] [Medline: 16096125]
49. McNeely J, Strauss SM, Wright S, Rotrosen J, Khan R, Lee JD, et al. Test-retest reliability of a self-administered Alcohol, Smoking and Substance Involvement Screening Test (ASSIST) in primary care patients. J Subst Abuse Treat 2014 Jul;47(1):93-101 [FREE Full text] [doi: 10.1016/j.jsat.2014.01.007] [Medline: 24629887]

50. McNeely J, Wright S, Matthews AG, Rotrosen J, Shelley D, Buchholz MP, et al. Substance-use screening and interventions in dental practices: survey of practice-based research network dentists regarding current practices, policies and barriers. J Am Dent Assoc 2013 Jun;144(6):627-638 [FREE Full text] [Medline: 23729460]
51. Gurol-Urganci I, de Jongh JT, Vodopivec-Jamsek V, Atun R, Car J. Mobile phone messaging reminders for attendance at healthcare appointments. Cochrane Database Syst Rev 2013(12):CD007458. [doi: 10.1002/14651858.CD007458.pub3] [Medline: 24310741]
52. Johnson SS, Driskell M, Johnson JL, Prochaska JM, Zwick W, Prochaska JO. Efficacy of a transtheoretical model-based expert system for antihypertensive adherence. Dis Manag 2006 Oct;9(5):291-301. [doi: 10.1089/dis.2006.9.291] [Medline: 17044763]
53. Levesque DA, Johnson JL, Welch CA, Prochaska JM, Paiva AL. Teen dating violence prevention: cluster-randomized trial of teen choices, an online, stage-based program for healthy, nonviolent relationships. Psychol Violence 2016 Jul;6(3):421-432 [FREE Full text] [doi: 10.1037/vio0000049] [Medline: 27482470]
54. Mauriello LM, Ciavatta MM, Paiva AL, Sherman KJ, Castle PH, Johnson JL, et al. Results of a multi-media multiple behavior obesity prevention program for adolescents. Prev Med 2010 Dec;51(6):451-456 [FREE Full text] [doi: 10.1016/j.ypmed.2010.08.004] [Medline: 20800079]
55. World Health Organization. In: Humeniuk R, Henry-Edwards S, Ali R, Poznyak V, Monteiro MG, editors. The Alcohol, Smoking and Substance Involvement Screening Test (ASSIST): Manual for use in primary care. Geneva, Switzerland: World Health Organization; 2010.
56. Gollwitzer PM. Implementation intentions: strong effects of simple plans. Am Psychol 1999;54(7):493-503. [doi: 10.1037/0003-066X.54.7.493]
57. Cancer. 2003. Making health communication programs work: A planner's guide URL: https://www.cancer.gov/publications/health-communication/pink-book.pdf [accessed 2017-09-07] [WebCite Cache ID 6tJA4xYqN]
58. Rimer BK, Orleans CT, Fleisher L, Cristinzio S, Resch N, Telepchak J, et al. Does tailoring matter? The impact of a tailored guide on ratings and short-term smoking-related outcomes for older smokers. Health Educ Res 1994 Mar;9(1):69-84. [Medline: 10146734]
59. Cardinal BJ. Development and evaluation of stage-matched written materials about lifestyle and structured physical activity. Percept Mot Skills 1995 Apr;80(2):543-546. [doi: 10.2466/pms.1995.80.2.543] [Medline: 7675587]
60. Levesque DA, Johnson JL, Welch CA, Prochaska JM, Fernandez AC. Computer-tailored intervention for juvenile offenders. J Soc Work Pract Addict 2012 Jan 1;12(4):391-411 [FREE Full text] [doi: 10.1080/1533256X.2012.728107] [Medline: 23264754]
61. Pew Research Center. Pewhispanic.: Pew Research Center; 2016. Facts on U.S. Latinos, 2015: Statistical portrait of Hispanics in the United States URL: http://www.pewhispanic.org/2016/04/19/statistical-portrait-of-hispanics-in-the-united-states-key-charts/ [accessed 2017-12-11] [WebCite Cache ID 6tJA9UQeD]
62. National Association of Community Health Centers. NACHC.: National Association of Community Health Centers; 2016. America's health centers: Fact sheet March 2016 URL: http://www.nachc.org/wp-content/uploads/2015/06/Americas-Health-Centers-March-2016.pdf [accessed 2017-12-11] [WebCite Cache ID 6tJAFNhJG]
63. 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]
64. Fixen DL, Naoom SF, Blase KA, Friedman RM, Wallace F. nirn.fpg.unc. Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network; 2005. Implementation Research: A Synthesis of the Literature URL: http://nirn.fpg.unc.edu/sites/nirn.fpg.unc.edu/files/resources/NIRN-MonographFull-01-2005.pdf [accessed 2017-12-11] [WebCite Cache ID 6tJAm7uWk]
65. Berwick DM. Developing and testing changes in delivery of care. Ann Intern Med 1998 Apr 15;128(8):651-656. [Medline: 9537939]
66. Agency for Healthcare Research and Quality. Innovations.ahrq. 2013. Plan-Do-Study-Act (PDSA) Cycle URL: http://innovations.ahrq.gov/qualitytools/plan-do-study-act-pdsa-cycle [accessed 2017-12-11] [WebCite Cache ID 6nafAvHj0]

Abbreviations

ASSIST: Alcohol, Smoking and Substance Involvement Screening Test
ASPIRE: Assessing Screening Plus Brief Intervention’s Resulting Efficacy
CTIs: computer-tailored interventions
EHR: electronic health record
FQHC: federally qualified health center
NIDA: National Institute on Drug Abuse
PAC: Personal Activity Center
PDF: portable document format
SBIRT: Screening, Brief Intervention, and Referral to Treatment

SD: standard deviation
SUD: substance use disorder
SURI: substance use risk intervention
TTM: transtheoretical model of behavior change
WHO: World Health Organization

Edited by G Eysenbach; submitted 19.01.17; peer-reviewed by P Chai, B Curtis, J Thrul; comments to author 12.07.17; revised version received 08.09.17; accepted 23.09.17; published 02.01.18.

Please cite as:
Levesque D, Umanzor C, de Aguiar E
Stage-Based Mobile Intervention for Substance Use Disorders in Primary Care: Development and Test of Acceptability
JMIR Med Inform 2018;6(1):e1
URL: http://medinform.jmir.org/2018/1/e1/
doi:10.2196/medinform.7355
PMID:29295811

Original Paper

A Pilot Study of Biomedical Text Comprehension using an Attention-Based Deep Neural Reader: Design and Experimental Analysis

Seongsoon Kim[1*], PhD; Donghyeon Park[1*], MSc; Yonghwa Choi[1*], BS; Kyubum Lee[1], PhD; Byounggun Kim[2], BS; Minji Jeon[1], MSc; Jihye Kim[3], PhD; Aik Choon Tan[3], PhD; Jaewoo Kang[1], PhD

[1]Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul, Republic Of Korea

[2]Interdisciplinary Graduate Program in Bioinformatics, Korea University, Seoul, Republic Of Korea

[3]Division of Medical Oncology, Department of Medicine, Translational Bioinformatics and Cancer Systems Biology Laboratory, University of Colorado Anschutz Medical Campus, Aurora, CO, United States

[*]these authors contributed equally

Corresponding Author:
Jaewoo Kang, PhD
Department of Computer Science and Engineering
College of Informatics
Korea University
145 Anam-ro, Seongbuk-Gu
Seoul, 02841
Republic Of Korea
Phone: 82 02 3290 4840
Email: kangj@korea.ac.kr

Abstract

Background: With the development of artificial intelligence (AI) technology centered on deep-learning, the computer has evolved to a point where it can read a given text and answer a question based on the context of the text. Such a specific task is known as the task of machine comprehension. Existing machine comprehension tasks mostly use datasets of general texts, such as news articles or elementary school-level storybooks. However, no attempt has been made to determine whether an up-to-date deep learning-based machine comprehension model can also process scientific literature containing expert-level knowledge, especially in the biomedical domain.

Objective: This study aims to investigate whether a machine comprehension model can process biomedical articles as well as general texts. Since there is no dataset for the biomedical literature comprehension task, our work includes generating a large-scale question answering dataset using PubMed and manually evaluating the generated dataset.

Methods: We present an attention-based deep neural model tailored to the biomedical domain. To further enhance the performance of our model, we used a pretrained word vector and biomedical entity type embedding. We also developed an ensemble method of combining the results of several independent models to reduce the variance of the answers from the models.

Results: The experimental results showed that our proposed deep neural network model outperformed the baseline model by more than 7% on the new dataset. We also evaluated human performance on the new dataset. The human evaluation result showed that our deep neural model outperformed humans in comprehension by 22% on average.

Conclusions: In this work, we introduced a new task of machine comprehension in the biomedical domain using a deep neural model. Since there was no large-scale dataset for training deep neural models in the biomedical domain, we created the new cloze-style datasets Biomedical Knowledge Comprehension Title (BMKC_T) and Biomedical Knowledge Comprehension Last Sentence (BMKC_LS) (together referred to as BioMedical Knowledge Comprehension) using the PubMed corpus. The experimental results showed that the performance of our model is much higher than that of humans. We observed that our model performed consistently better regardless of the degree of difficulty of a text, whereas humans have difficulty when performing biomedical literature comprehension tasks that require expert level knowledge.

(JMIR Med Inform 2018;6(1):e2) doi:10.2196/medinform.8751

KEYWORDS

machine comprehension; biomedical text comprehension; deep learning; machine comprehension dataset

Introduction

The rate of discovering and accumulating new biomedical knowledge continues to increase rapidly due to technological advances. Most of the new findings are published in the form of biomedical literature. The rate of increase in PubMed volume reflects such a growth trend. On average, more than 3000 papers are newly added to PubMed every day. As the number of publications of biomedical research papers rapidly increases, it becomes more difficult for biomedical knowledge workers to collect and assemble information from the fast-growing literature to compose answers to biomedical questions [1]. To address this issue, automatic information-seeking and processing approaches such as information retrieval, biomedical text mining [2-5], and biomedical question answering (QA) systems [6-11] have been rigorously studied in recent years.

Recently, advances in artificial intelligence (AI) based on deep learning technology not only improved the performance of existing text mining models, but also reached a level where machines can read and comprehend texts so that they can respond to given questions. In the AI community, researchers have actively conducted studies to measure a machine's ability to understand text in reading comprehension tasks [12-17]. Machine comprehension tasks can be defined as testing the ability of a machine to answer a question based on context. Recent studies show that deep neural network-based models hold promise for performing reading comprehension tasks, and currently outperform all alternative models [12-14]. Several AI research groups, including Google, Facebook, and IBM Watson, developed new text comprehension models [12-15].

Deep learning-based approaches require a sufficient amount of data to train a model. Therefore, in addition to model architecture, methods that automatically generate a considerable amount of data (which can be used for training neural models) have been actively studied. One study used cloze-style [18] QA pairs that were employed to assess the learning ability of elementary school students. Several large cloze-style context-question-answer datasets have also been introduced. These datasets contain only general information from sources such as news articles (Cable News Network [CNN]/Daily Mail) and children's books, and not professional knowledge.

With a well-developed machine comprehension model, one can quickly and efficiently find the correct answer to a question using the given context. However, while machine comprehension is actively studied in the AI research field, recent machine comprehension technologies have not been applied to the biomedical domain, which requires information processing the most. Currently, there are no datasets for biomedical text comprehension tasks, and thus a computer's ability to comprehend biomedical domain knowledge has not yet been verified.

In this article, we propose a machine comprehension task on biomedical literature. We also provide a new and large cloze-style dataset called BioMedical Knowledge Comprehension (BMKC) which can be employed to train deep neural network models. Our goal was to test whether a machine can correctly comprehend scientific papers such as those in our dataset, since it has already been proven in previous research that it can comprehend general text such as storybooks. We demonstrate that our state-of-the-art deep learning model enhanced with biomedical domain-specific features can comprehend biomedical literature. Through a performance comparison with humans, we observed that the comprehension performance of humans varies depending on the degree of difficulty of a text, while machines perform consistently well.

This research offers three contributions to the field. First, to the best of our knowledge, this work is the first to propose a deep learning-based machine comprehension task in the biomedical domain. Second, we used the PubMed corpus to generate considerably large datasets for training deep neural machine comprehension models. The automatically generated datasets open huge opportunities for data-hungry techniques such as deep-learning and future QA systems. We made the datasets publicly available [19]. Third, we present methods that can improve the performance of existing machine comprehension models using pretrained Word2Vec and entity type embedding features. We employed an ensemble approach of combining multiple single models to produce improved answer prediction results. The experimental results showed that our proposed methods can help our model, based on the original text comprehension model developed for general text, to achieve state-of-the-art performance in biomedical literature.

Methods

In this section, we first explain the process of automatically creating a large-scale biomedical text dataset for machine comprehension tasks. We then describe the Attention Sum Reader (ASR) [15], a state-of-the-art deep neural model that is used for machine comprehension tasks. We propose two additional techniques utilizing pretrained word vector and entity type embeddings, both of which we used to build our text comprehension model tailored to the biomedical domain. To improve the prediction accuracy, we also applied ensemble learning in which the final answer prediction was obtained by integrating the output of several independent homogenous models.

Cloze-Style Biomedical Machine Comprehension Task Overview

A cloze-style question is formed by removing a phrase from a sentence; cloze-style questions are *fill-in-the-blank* type questions. The cloze-style dataset is in the form of context-question-answer triplets. From the perspective of machine learning, this task is easy to evaluate. The cloze-style text comprehension task can be defined as tuples of the form (d, q, a, A), where d is a document, q is a query, and a is the answer to query q, which comes from a set of candidate answers A. More specifically, given a document-query pair (d, q), we aim to find a A which answers q.

Cloze-Style Biomedical Machine Comprehension Dataset

Our BMKC datasets are in cloze-style form (context-question-answer) like other existing datasets. The main difference is that BMKC consists of scientific articles in the biomedical domain, which require expert knowledge for comprehension, while other existing datasets contain nonscientific, general texts such as news articles and children's storybooks [12,13,16].

We explain in detail the method for generating the dataset as follows. First, we needed a document for the context. We chose the abstract of a paper as the context *d* in our BMKC datasets. Unlike the CNN news dataset in which summaries are given, abstracts of research articles do not have such summaries. Hence, we took a different approach to automatically generating questions.

The question *q* is generated in two different ways. A question in Biomedical Knowledge Comprehension Title (BMKC_T) is constructed from the title of an academic paper because the title can be considered as a short summary of the abstract of the paper. Biomedical Knowledge Comprehension Last Sentence (BMKC_LS) uses the last sentence in the abstract of a paper as a question, inspired by Hill et al's work [13]. In short, the BMKC datasets (Table 1) can be defined as tuples of the form (*d*, *q*, *a*, *A*), where *d* is an abstract, *q* is a title (BMKC_T) or the last sentence in an abstract (BMKC_LS), and *a* is the answer to query *q*.

Data Generation Process

The process of generating the BMKC datasets consisted of the following three steps. First, we gathered biomedical research articles from PubMed. Having started in the 1960s, PubMed now provides more than 24 million references to biomedical and life science articles dating back as far as 1946. We downloaded a total of 200 MEDLINE files (medline16n0813-medline16n08131012) that contain approximately 2,200,000 biomedical papers that include titles, abstracts, keywords, published year, author information, and so on.

Of the 200 MEDLINE files, we used 196 files (medline16n0813-medline16n08131008) as our training set, two files (medline16n1009-1010) as our validation set, and the last two files as our test set (medline16n1011-1012). Table 2 shows the number of articles by published years in the 200 MEDLINE files. More than 95% (2,110,444/2,208,081) of the articles were published after 2010. Note that the publication dates of the journal papers were randomly distributed across the training set, validation set, and the test set.

The next step was extracting biomedical entities to generate candidate answers to cloze-style questions. We exploited the biomedical named entity extractor in Biomedical Entity Search Tool (BEST) [20]. To increase the coverage of biomedical entities, we added Medical Subject Headings (MeSH), a hierarchical biomedical vocabulary thesaurus, for our entity extraction process. One advantage of using MeSH is that it provides a kind of entity resolution function that groups several different biomedical entity names with the same meaning into one MeSH identification (ID). Next, we replaced all entity names with their unique entity IDs. Unlike the work of Herman et al [12], we did not randomly permute the entity ID for each context. Retaining unique entity IDs allows the model to acquire background knowledge during the training process, which will improve the performance of the biomedical knowledge-specific QA task.

Table 1. Example of BMKC_T (Title) and BMKC_LS (Last Sentence). In the BMKC_LS dataset, the last sentence of context is excluded in training as it is a question itself.

Parameter	BMKC_T (Title)	BMKC_LS (Last Sentence)
Context (abstract of a paper)	In breast cancer, overexpression of the nuclear coactivator NCOA1 (SRC-1) is associated with disease recurrence and resistance to endocrine therapy. To examine the impact of NCOA1 overexpression on morphogenesis and carcinogenesis in the mammary gland (MG), we generated MMTV-hNCOA1 transgenic [Tg(NCOA1)] mice. (...) In a cohort of 453 human breast tumors, NCOA1 and CSF1 levels correlated positively with disease recurrence, higher tumor grade, and poor prognosis. Together, our results define an NCOA1/AP-1/CSF1 regulatory axis that promotes breast cancer metastasis, offering a novel therapeutic target for impeding this process.	
Question	___?___ directly targets M-CSF1 expression to promote breast cancer metastasis.	Together, our results define an NCOA1/___?___/CSF1 regulatory axis that promotes breast cancer metastasis, offering a novel therapeutic target for impeding this process.
Answer Candidates (Biomedical Named Entities)	macrophage, carcinogenesis, morphogenesis, metastasis, disease, AP-1, tumor, lung, NCOA1, (therapy, therapeutic), recurrence, mammary gland, epithelial cells, cells, CSF1, SRC, mice, c-Fos, human, affect, (breast cancer, breast tumors), efficiency	

Table 2. Number of publications by years in the 200 MEDLINE files.

Year	Number of papers
1910 - 1959	12,178
1960 - 2009	85,459
2010 - 2016	2,110,444

Last, we filtered context-question pairs that did not meet the following two conditions: (1) the answer should appear at least once in both the context and the question to form a valid context-question pair, and (2) the total number of candidate answers should exceed 20 to ensure a certain level of difficulty and a fair comparison with other corpora. In the end, we obtained approximately one half million context-question pairs for both the BMKC_T and BMKC_LS datasets.

Attention Sum Reader

The Deep Long-Short Term Memory Reader [12] was first proposed to perform a machine comprehension task on a cloze-style dataset with a deep-learning model, and subsequent studies were also conducted. Recently, attention-based models have been actively studied among various deep learning models due to their high performance on various tasks [21-24]. Since the text comprehension task involves selecting one correct word in the context, the attention mechanism achieves superior performance on the task. Specifically, the ASR model [15] achieves state-of-the-art performance on the general text datasets (CNN and Daily Mail). Hence, we performed our task of biomedical literature comprehension based on ASR architecture. The overall ASR model works as follows.

The ASR model uses the word embedding function e, utilizing look-up matrix W_v, to convert words into low-dimensional vector representations whose rows are word indices from the vocabulary V (Figure 1 a).

The model has two encoders: a context encoder (Figure 1b) and a query encoder (Figure 1c). The encoders convert a context and a query into continuous vector representations. The context encoder *f* is implemented by a bidirectional Gated Recurrent Unit (GRU). Details of the answer calculation process are as follows:

The encoders receive word vectors by the word embedding function as an input. We denote the contextual embedding of the *i*-th word in *d* as f_i *(d)*=*(d)*||*(d)* where || denotes the vector concatenation of forward and backward contextual embeddings and. Then, a query is encoded by the query encoder *g* which is also implemented by another bidirectional GRU network such that *g(q)*=*(q)*||*(q)*. The parameters *f*, *g*, and are jointly optimized during the training phase.

Next, word attention (answer probability) *i* is calculated by the dot product between the encoders (Figure 1d) and passed to the soft-max layer as follows:

where <,> denotes the dot product between the vectors. Finally, the model calculates the scores of all possible answers based on their representations, and combines multiple mentions of the same candidate answer by adding up their answer probability (Figure 1e). The final answer token has the highest probability *P* (*a* |*d*, *q*) to answer question *q* over given document *d* such that:

where *I (a,d)* is a set of positions of the answer token in the document. The candidate answer with the maximum probability is then selected as the final answer.

As we described, the ASR model adopts an aggregation scheme known as pointer sum attention. Hence, the performance of the attention-based model is superior to that of the general deep learning models [12,13]. Since the attention-based model is suitable for focusing on a specific target, it can achieve high performance on the cloze-style QA task of selecting a specific word to answer a question using context.

Figure 1. The ASR model architecture adopted from the original paper.

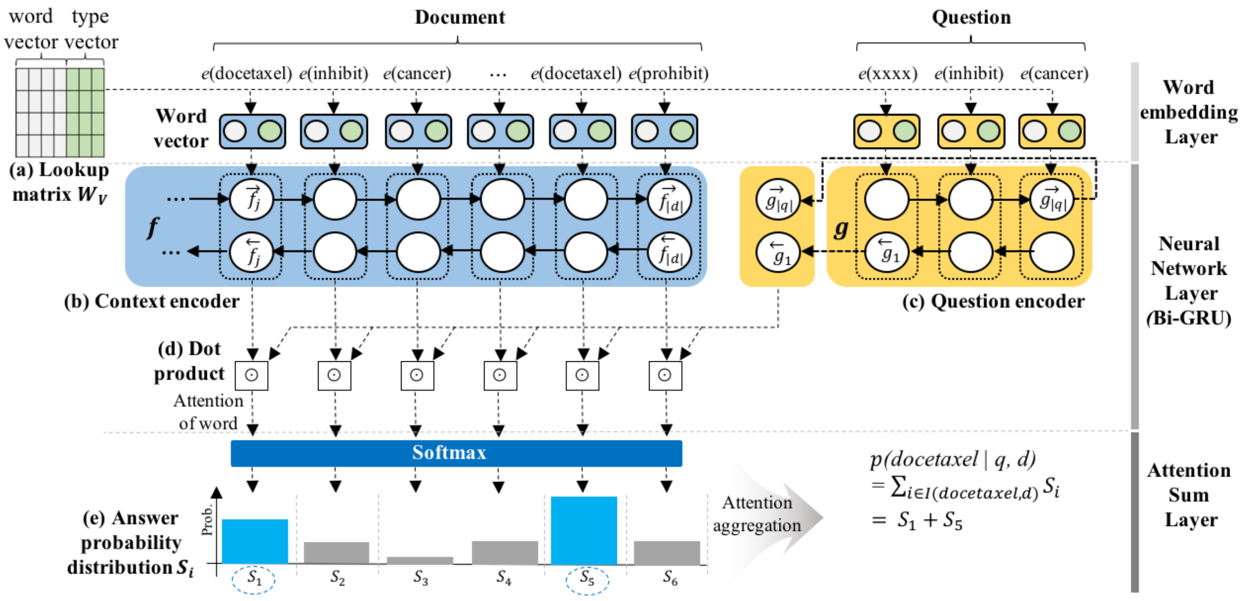


Table 3. The list of entity types from two information sources: BEST entity extractor [20] and MeSH tree structures.

Type Source	Entity Types
BEST	Gene, Drug, Chemical Compounds, Target, Disease, Toxin, Transcription Factor, miRNA, Pathway, Mutation
MeSH	Anatomy [A]; Organisms [B]; Diseases [C]; Chemicals and Drugs [D]; Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E]; Psychiatry and Psychology [F]; Phenomena and Processes [G]; Disciplines and Occupations [H]; Anthropology, Education, Sociology, and Social Phenomena [I]; Technology, Industry, and Agriculture [J]; Humanities [K]; Information Science [L]; Named Groups [M]; Health Care [N]; Publication Characteristics [V]; Geographicals [Z]

Improving Model Performance Using Pretrained Biomedical Word Embedding

Representing words as low-dimensional vectors is a key element of deep learning models used in natural language processing (NLP) tasks. As described in the previous section, the neural model selects the correct answer using the inner product between the vectors of the context and the query representation. Therefore, if the vector of the word that makes up the context and the query are well represented in the vector space, the probability that the chosen answer is correct will be higher.

It is known that word embeddings trained on an adequately large corpus capture latent semantic meanings and improve performance on nearly all NLP tasks. The openly available biomedical literature resources (eg, PubMed and PubMed Central Open Access) contain over 5.5 billion words in abstracts and full texts [25]. Using word embedding vectors trained on such a large amount of text can improve the performance of the model in our task. This is true because a vector representation learned on a large corpus captures more precise semantics of words. We therefore aimed to improve the performance of the original ASR model developed for general text (news) using a pretrained word vector instead of a randomly initialized word embedding. We downloaded the pretrained word vector from Pyysalo et al [26]. The details about bio-word vectors are as follows. The source data for training bio-word-vectors were derived from PubMed and all of the full-text documents obtained from the PubMed Central Open Access subset. The word vectors were generated by the Skip-Gram model with a window size of 5, hierarchical soft-max training, and a frequent word subsampling threshold of 0.001. We used 200-dimensional word vectors, as done in many previous NLP tasks. We compared the performance of each initialization of the lookup table in the *Experimental* section.

Improving Model Performance Using Entity Type Embedding

Adding entity type information can be helpful for understanding contexts. For example, when expressions such as “@entity1 expression” or “@entity2 expression” appear in the context and the model knows @entity1 and @entity2 are *Gene* type entities, the model can learn that the context is about gene expression. Also, when other expressions such as “@entity3 0.3%” or “@entity4 100mg” appear and information that @entity3 and @entity4 are *Drug* type entities is given, the model can learn that the context is about drug concentration.

To leverage type information of biomedical entities, we used entity types identified by the BEST entity extraction tool [20]. To improve recall, we additionally extracted MeSH terms and utilized the MeSH term hierarchy as each term’s entity type label. More specifically, the MeSH tree has a hierarchical structure similar to that of concept ontology. We used parent nodes in the MeSH tree as representative entity types. Finally, we selected 10 entity types from BEST and 16 types from MeSH (Table 3).

Next, we merged some entity types that share similar semantics. For example, *Gene*, *Target*, and *Transcription Factor* types can be merged into the type *Gene*. Similarly, *Drug*, *Toxin*, *Chemical Compounds*, and *Chemicals andDrugs* types are merged into the representative type *Chemicals andDrugs* [D]. We assigned *Unknown* if words did not have a specific type. We finally constructed 20-dimensional randomly initialized type embedding vectors and concatenated them to the original word vector (Figure 1a).

Improving Model Performance Using an Ensemble Model

A neural network ensemble approach combines the prediction results of individual models. This ensemble approach can lead to performance improvement based on its generalization capabilities [27]. Two considerations for a neural network ensemble approach are individual network generation and integrated output [28]. We adopted the ensemble averaging method in this study. An ensemble averaging consists of a set of independently trained neural network models which share the same training data, and whose individual outputs are linearly combined by averaging the results of the individual models to produce an overall prediction. Since the weights of each neural network model are randomly initialized, we can create an independent network with the same network structure. Although the resulting ensemble model has the same bias as the individual models, its variance is reduced, and thus it can achieve better prediction accuracy than a single model.

Results

Biomedical Knowledge Comprehension Dataset

Our BMKC datasets are the first large-scale datasets developed for biomedical machine comprehension tasks. We made our dataset publicly available for future research use [19]. Table 4 shows the statistical summaries of our dataset in comparison with four existing machine comprehension datasets.

Table 4. Statistics of BMKC datasets and other existing datasets. Note that the number of queries is equal to the number of documents since one query is generated per document.

Dataset	Number of Queries	Maximum number of options	Average number of options	Average number of tokens	Vocabulary Size (all)
BMKC_T					
Train	463,981	93	25.6	291	876,621
Validation	5278	66	25.4	291	
Test	3868	74	25.7	289	
BMKC_LS					
Train	362,439	90	25.3	270	714,751
Validation	4136	57	25.1	269	
Test	3205	74	25.4	271	
CNN [12]					
Train	380,298	527	26.4	762	118,497
Validation	3924	187	26.5	763	
Test	3198	396	24.5	716	
Daily Mail [12]					
Train	879,450	371	26.5	813	208,045
Validation	64,835	232	25.5	774	
Test	53,182	245	26.0	780	
CBT_NE [a] [13]					
Train	120,769	10	10	470	53,185
Validation	2000	10	10	448	
Test	2500	10	10	461	
CBT_Noun [b] [13]					
Train	180,719	10	10	433	53,063
Validation	2000	10	10	412	
Test	2500	10	10	424	

[a] CBT_NE is a dataset that uses the Children's Book Test Named Entity that appears in a context as a candidate answer

[b] CBT_Noun is a dataset that uses the Children's Book Test Noun phrase that appears in a context as a candidate answer

The CNN and Daily Mail datasets contain story-question pairs from CNN and Daily Mail news stories, respectively. The Children’s Book Test (CBT) dataset contains stories from children's books. A context consists of 20 consecutive sentences from children’s books and a question is made by removing a word from the 21st consecutive sentence. The detailed comparison of the datasets is given below. The dataset comparison is based on the training set that occupies the largest portion of each dataset.

Dataset Size

The size of the BMKC datasets (BMKC_T: 463,981, BMKC_LS: 362,439) is larger than that of all other datasets (CNN: 380,298, Children's Book Test Noun Phrase [CBT_Noun]: 180,719, Children's Book Test Named Entity [CBT_NE]: 120,769) except that of the Daily Mail dataset (879,450). Although the current BMKC dataset is large enough to train a reasonably complex deep neural reader, the size of the training set can easily be increased by adding articles from MEDLINE.

Query Length

As with the length of each query (a single context-question pair), the average number of tokens of our BMKC dataset (BMKC_T: 291, BMKC_LS: 270) is smaller than that of other datasets (CNN: 762, Daily Mail: 813, CBT_Noun: 470, CBT_NE: 433). The length of abstracts of academic papers is usually limited, while news articles can include lengthy context and have no length limit.

Number of Candidate Answers

The average number of options (which is the number of candidate answers to a question) of the BMKC dataset is comparable to that of the CNN and Daily Mail datasets, and larger than that of the CBT_Noun and CBT_NE datasets.

Table 5. Accuracies of the original ASR model and feature-enhanced models (ASR+BE, ASR+TE, ASR+BE+TE) on the BMKC_T and BMKC_LS datasets. The results of both the single and ensemble models are reported. The best scores are highlighted in italics.

Model	BMKC_T		BMKC_LS	
	Validation (%)	Test (%)	Validation (%)	Test (%)
Single				
ASR [15]	79.8	77.8	73.4	70.5
ASR+BE	81.0	*78.6*	74.6	71.4
ASR+TE	80.9	78.5	74.3	70.1
ASR+BE+TE	*81.4*	78.3	*74.8*	*72.0*
Ensemble				
ASR	83.7	81.4	77.6	75.8
ASR+BE	85.2	83.3	*80.1*	*77.7*
ASR+TE	85.2	*83.9*	79.5	76.6
ASR+BE+TE	*85.5*	83.6	*80.1*	77.3

Unique Vocabulary Size

The size of a unique vocabulary of the BMKC dataset exceeds that of all other datasets because academic articles contain considerably more domain-specific terms than general texts.

Deep Neural Model Performance

Performance Enhancement With Biomedical Domain Specific Features

The ASR model used stochastic gradient descent with the Adaptive Moment Estimation update rule and learning rates of 0.001 and 0.0005. The model used GRU for its Recurrent Neural Network. The initial weights in the word embedding matrix were randomly and uniformly drawn from the interval (-0.25, 0.25). We used a batch size of 32.

The performance of text comprehension models on the BMKC_T and BMKC_LS datasets is summarized in Table 5. We have created four single models and four ensemble models. The ASR model represents the basic implementation of the ASR model originally developed for general text comprehension tasks [15]. The ASR model uses all randomly initialized word vectors. The ASR+Bio-word Embedding (ASR+BE) model represents an ASR model that is initialized with word vectors pretrained on PubMed, whereas the ASR+Type Embedding (ASR+TE) model represents an ASR with type information embedding. The ASR+BE+TE model denotes ASR with bio-word vector embedding and type embedding.

Single Model

We report the performance of the single models on the validation and test sets. While the original ASR model achieved accuracies of 79.8% and 73.4% on the BMKC_T and BMKC_LS validation set (respectively), the ASR+BE single model featuring pretrained word embedding achieved accuracies of 81.0% and 74.6% on the BMKC_T and BMKC_LS datasets (respectively), and the ASR+TE single model with entity type information obtained accuracies of 80.9% and 74.3% on the BMKC_T and on BMKC_LS datasets (respectively). The single model with all features (ASR+BE+TE) achieved the highest validation accuracies of 81.4% and 74.8% on the BMKC_T and BMKC_LS datasets, respectively. The test set accuracy also increased when we used pretrained word vectors and type embedding. The ASR+BE single model achieved the best accuracy of 78.6% on the BMKC_T test set whereas the ASR+BE+TE single model achieved 72.0% on the BMKC_LS test set.

Ensemble Model

We also report the performance results of our ensemble models. For the ensemble method, we used the ensemble of eight models. Among all of the learned models, we selected the model that achieved an accuracy of at least 70% on the validation set as the ensemble member. Fusing multiple models significantly increased the validation and test accuracy on both the BMKC_T and BMKC_LS datasets. As in the case of the single models, the ensemble models trained with the biomedical-enhanced features ASR+BE+TE achieved the highest accuracies on both the BMKC_T and BMCK_LS validation sets. The ASR+BE+TE ensemble model performed 5.0% and 6.6% better than the ASR+BE+TE single models on the BMKC_T and BMKC_LS validation sets, respectively (from 81.4% to 85.5% on BKMC_T and from 74.8% to 80.1% on BKMC_LS with the ASR+BE+TE setting). When using the ASR+BE+TE ensemble model, performance on the test set improved considerably. The ASR+TE ensemble model achieved the best performance of 83.9% (6.9% improved from the ASR+TE single model) on the BMKC_T test set, and the ASR+BE ensemble model achieved the best performance of 77.7% (8.8% improved) on the BMKC_LS test set.

Improvements From the Original ASR Model

We augmented the original ASR model [15] with bio-word embedding, entity type embedding, and an ensemble model, each of which improved the performance of the original model. The ASR+BE+TE ensemble model outperformed the original ASR model by 7.1% (from 79.8% to 85.5%) and 9.1% (from 73.4% to 80.1%) on the BMKC_T and BMKC_LS validation sets, respectively. Similarly, the ASR+BE+TE ensemble model performed 7.5% (from 77.8% to 83.6%) and 9.6% (from 70.5%

to 77.3%) better than the original model on the BMKC_T and BMKC_LS test sets, respectively.

In addition, we report the top-*N* accuracy of our model's top-*N* predicted answers in Table 6. In top-*N* accuracy, if any of the top-*N* predicted answers match the correct answer, the model's output is considered correct. The ASR+BE+TE single model was used to compute the top-*N* accuracies. As demonstrated by the result, our model effectively puts correct answers in the top of the list of predicted answers. For example, on the BMKC_T test set, our model achieved a top-3 accuracy of 90.3%, which signifies that in over 90% of cases, users can find the correct answer in the top 3 of the outputs of the model.

Our Model and Human Performance Comparison

We made a test set for measuring a human's ability to read and comprehend biomedical literature, and compared the performance of humans on the test set with that of our neural model (Table 7). For the test set, we randomly selected 25 articles each from the BMKC_T and BMKC_LS datasets. We selected articles containing the terms "human" and "cancer" that were published between 01/01/2016 and 12/31/2016.

For the human evaluees, we hired six people from three different backgrounds. The first group consisted of two undergraduate students with a background in computer science. The second group consisted of two graduate students majoring in bioinformatics. The last group consisted of two bioinformatics professionals with at least eight years of post-doctoral experience in computational oncology. To measure the comprehension ability of a machine, we used the pretrained ASR+BE+TE single model.

To evaluate the performance of the machine comprehension model, which is given a certain amount of information, we report the global ID setting in which all contexts share the global entity ID set, and the local ID setting in which the entity ID is independently assigned for each context. We provided a human evaluee with a set of tests that did not anonymize the entity ID, which is equivalent to the global ID setting for the model.

The experimental results in Table 7 show that the machine outperformed the human groups in both accuracy and time. The machine performed at a similar level to that observed in Table 6. Even in the local ID setting, in which the information about the entity is hidden from the model, the model outperformed the human evaluees. Furthermore, the human groups had some difficulty answering the given test set. The group of graduate students with biomedical background knowledge performed better than the undergraduate student group, as we expected. Interestingly, the bioinformatician group took longer to answer questions in our BMKC datasets. We assume that bioinformaticians tend to exploit their knowledge to solve the problems, whereas students with no background knowledge in the biomedical domain tend to guess. A detailed description of the test questions and the responses of our model (ASR+BE+TE) and each human evaluee is provided in the Multimedia Appendix 2.

Our model's outperformance of humans is notable because humans have usually performed better on existing cloze-style datasets (as shown in Table 8). We present Table 8 to compare the comprehension performance of humans and the machine on the other general text domain datasets. Note that there were no human evaluation results reported for the CNN dataset when it was initially released. Hence, the CNN and CBT_NE datasets were manually evaluated by humans through the crowdsourcing platform CrowdFlower [29]. Details of the human evaluation results are provided in Multimedia Appendix 1. The results show that humans perform better than (or at least comparable to) the machine in the general text comprehension tasks.

Table 6. Top-*N* accuracy of the model on the BMKC test sets. The top-*N* accuracy is calculated using the ASR+BE+TE single model.

Dataset	Top-1 accuracy (%)	Top-2 accuracy (%)	Top-3 accuracy (%)	Top-5 accuracy (%)
BMKC_T-Test	78.3	86.8	90.3	93.5
BMKC_LS-Test	72.0	81.7	85.7	90.5

Table 7. Biomedical literature comprehension results of humans and our model on the BMKC datasets.

User	BMKC_T		BMKC_LS		Total		
	Number of problems	Accuracy (%)	Number of problems	Accuracy (%)	Number of problems	Accuracy (%)	Time (minutes)
Human							
Undergraduate	14.5/25	58.0	10.5/25	42.0	25/50	50.0	77.5
Graduate	18/25	72.0	14/25	56.0	32/50	64.0	117.5
Expert	16.5/25	66.0	13/25	52.0	29.5/50	59.0	115.5
Machine							
ASR+BE+TE_single (global ID)	23/25	92.0	19/25	76.0	42/50	84.0	0.001
ASR+BE+TE_single (local ID)	19/25	76.0	18/25	72.0	37/50	74.0	0.001

Table 8. Text comprehension results of humans and the text comprehension model on the CNN and CBT datasets. The machine comprehension results are obtained from Kadlec et al [15].

Model	Dataset, accuracy (%)	
	CNN	CBT_NE
Human	69.2	81.6
Machine (ASR-single)	69.5	68.6

Discussion

Deep Neural Models are Less Affected by the Difficulty of the Text Than Humans

The aim of this study was to evaluate the machine comprehension model's performance on biomedical literature datasets. In the performance evaluation on our new BMKC datasets and the existing general text datasets, our deep neural models achieved robust performance regardless of the degree of difficulty of the text, whereas humans found it difficult to solve the biomedical literature comprehension tasks that require expert knowledge. This result demonstrates that deep neural models are less affected by the difficulty of text than humans, and therefore may be used to assist human researchers when processing information in big data.

Error Analyses

In this section, we analyzed the errors in the machine comprehension results of our machine comprehension model. The QA results of the model are shown as an attention heatmap. We discuss the two representative error cases in detail below: *causal inference error* and *concept hierarchy error*.

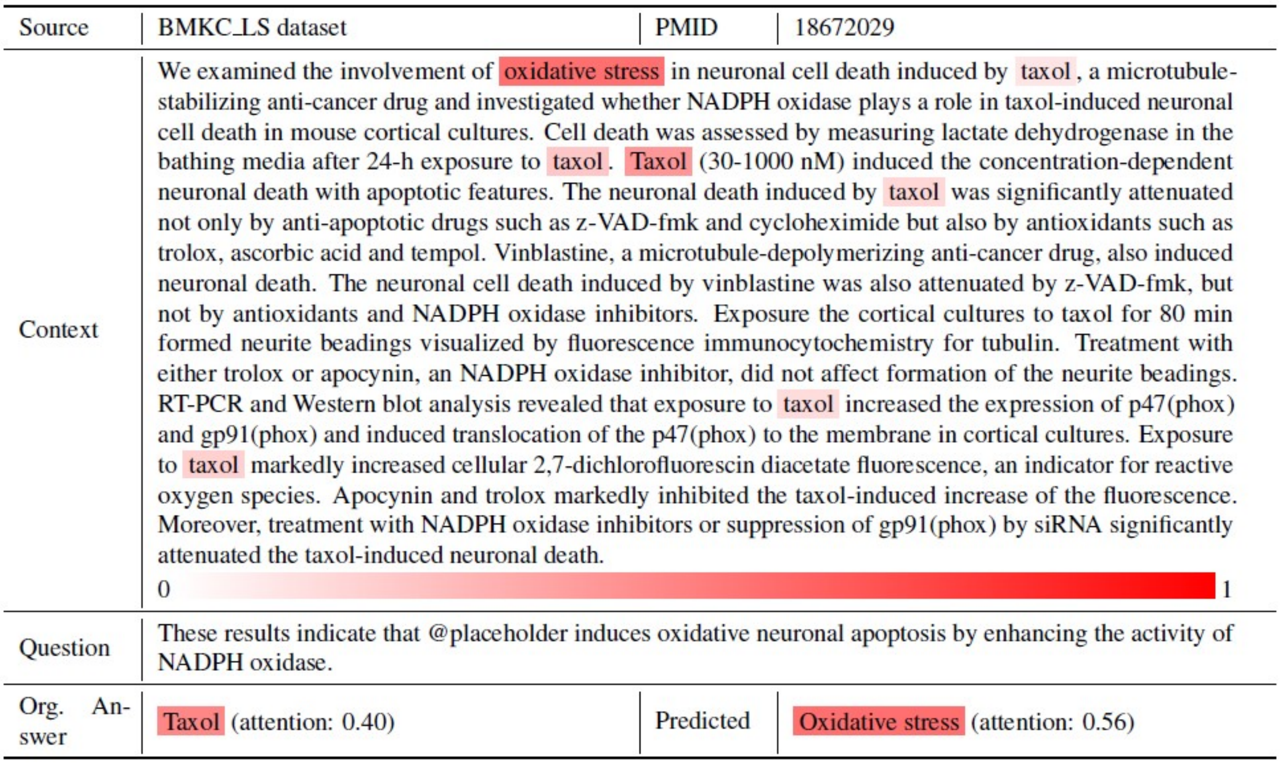
Causal Inference Error

We observed cases in which the model could not respond accurately to questions that required step-by-step reasoning, such as a time-order relationship with the cause preceding the effect. We explain such cases using the example in Figure 2. The example document includes the relationship between Taxol, oxidative stress, and cell death. According to the context, Taxol induces oxidative stress, which leads to neuronal apoptosis. The question asked for the cause of oxidative neuronal apoptosis or cell death. As observed in the attention heatmap, the model provided oxidative stress as the cause of cell death, but it is ultimately triggered by Taxol, which is the correct answer.

Concept Hierarchy Error

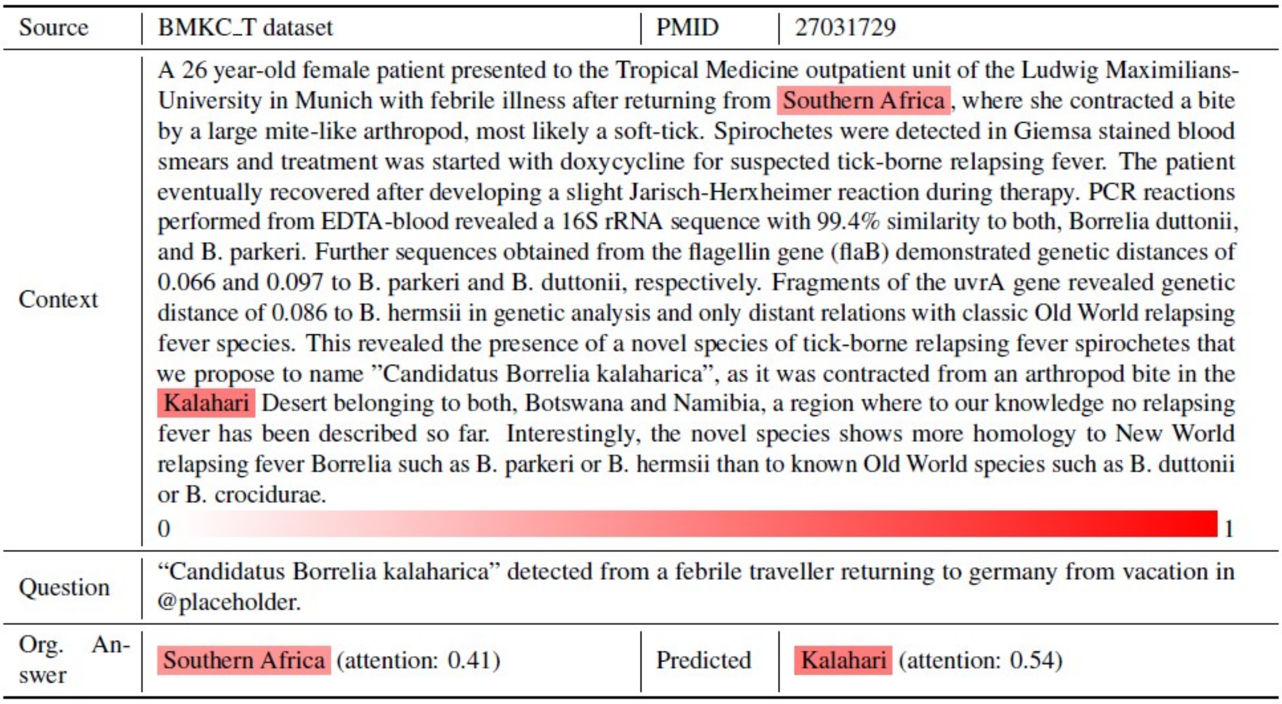
A concept hierarchy error refers to a situation in which the model selects an option that does not match the correct answer when considering entities in an inclusive relationship. The attention heatmap in Figure 3 shows examples of concept hierarchy errors. The question asks about geo-location and the answer is "South Africa." Interestingly, we observed that the model considers both "South Africa" and "Kalahari," which is the name of a desert located in South Africa, as candidate answers. However, the model gave "Kalahari" more weight, which is also correct.

Figure 2. Attention heatmap from the ASR model for case 1: causal inference problem.



Source	BMKC_LS dataset	PMID	18672029
Context	We examined the involvement of oxidative stress in neuronal cell death induced by taxol , a microtubule-stabilizing anti-cancer drug and investigated whether NADPH oxidase plays a role in taxol-induced neuronal cell death in mouse cortical cultures. Cell death was assessed by measuring lactate dehydrogenase in the bathing media after 24-h exposure to taxol . Taxol (30-1000 nM) induced the concentration-dependent neuronal death with apoptotic features. The neuronal death induced by taxol was significantly attenuated not only by anti-apoptotic drugs such as z-VAD-fmk and cycloheximide but also by antioxidants such as trolox, ascorbic acid and tempol. Vinblastine, a microtubule-depolymerizing anti-cancer drug, also induced neuronal death. The neuronal cell death induced by vinblastine was also attenuated by z-VAD-fmk, but not by antioxidants and NADPH oxidase inhibitors. Exposure the cortical cultures to taxol for 80 min formed neurite beadings visualized by fluorescence immunocytochemistry for tubulin. Treatment with either trolox or apocynin, an NADPH oxidase inhibitor, did not affect formation of the neurite beadings. RT-PCR and Western blot analysis revealed that exposure to taxol increased the expression of p47(phox) and gp91(phox) and induced translocation of the p47(phox) to the membrane in cortical cultures. Exposure to taxol markedly increased cellular 2,7-dichlorofluorescin diacetate fluorescence, an indicator for reactive oxygen species. Apocynin and trolox markedly inhibited the taxol-induced increase of the fluorescence. Moreover, treatment with NADPH oxidase inhibitors or suppression of gp91(phox) by siRNA significantly attenuated the taxol-induced neuronal death. 0 1			
Question	These results indicate that @placeholder induces oxidative neuronal apoptosis by enhancing the activity of NADPH oxidase.		
Org. Answer	Taxol (attention: 0.40)	Predicted	Oxidative stress (attention: 0.56)

Figure 3. Attention heatmap from the ASR model for case 2: concept hierarchy problem.



Source	BMKC_T dataset	PMID	27031729
Context	A 26 year-old female patient presented to the Tropical Medicine outpatient unit of the Ludwig Maximilians-University in Munich with febrile illness after returning from Southern Africa, where she contracted a bite by a large mite-like arthropod, most likely a soft-tick. Spirochetes were detected in Giemsa stained blood smears and treatment was started with doxycycline for suspected tick-borne relapsing fever. The patient eventually recovered after developing a slight Jarisch-Herxheimer reaction during therapy. PCR reactions performed from EDTA-blood revealed a 16S rRNA sequence with 99.4% similarity to both, Borrelia duttonii, and B. parkeri. Further sequences obtained from the flagellin gene (flaB) demonstrated genetic distances of 0.066 and 0.097 to B. parkeri and B. duttonii, respectively. Fragments of the uvrA gene revealed genetic distance of 0.086 to B. hermsii in genetic analysis and only distant relations with classic Old World relapsing fever species. This revealed the presence of a novel species of tick-borne relapsing fever spirochetes that we propose to name ”Candidatus Borrelia kalaharica”, as it was contracted from an arthropod bite in the Kalahari Desert belonging to both, Botswana and Namibia, a region where to our knowledge no relapsing fever has been described so far. Interestingly, the novel species shows more homology to New World relapsing fever Borrelia such as B. parkeri or B. hermsii than to known Old World species such as B. duttonii or B. crocidurae. 0 1		
Question	“Candidatus Borrelia kalaharica” detected from a febrile traveller returning to germany from vacation in @placeholder.		
Org. Answer	Southern Africa (attention: 0.41)	Predicted	Kalahari (attention: 0.54)

To summarize, the error cases discussed above can be regarded as structural limitations of the ASR model configured by the pointer-sum network method, which selects only one correct word as the final answer in the given context. The pointer-sum network structure is limited in solving questions that require an understanding of the inclusive relationship between step-wise reasoning and conceptual reasoning. Other recent deep-running models that currently perform machine comprehension tasks also do not consider such causal inference or concept hierarchy. These are fundamental limitations of the current deep learning models and should be improved with the advances of AI technology in the near future.

Limitations of Cloze-Style Question Answering and Future Direction

The final goal of biomedical knowledge QA is to help domain experts more quickly and efficiently discover knowledge from the vast amount of information in the literature. However, the knowledge obtained through QA systems is context-insensitive and thus is not directly applicable to individual patient care scenarios. The QA systems are more appropriate to be used as decision support systems for domain experts to help them quickly process information and make more educated decisions in a shorter time.

One limitation of our current QA system is that the candidate answers are limited to biomedical entities. Although the answer probabilities are calculated for all words in the input context, the system only considers as candidate answers the biomedical entities identified by the entity extraction module used in our preprocessing step. Extracting candidate answers from the input text and providing them along with the question is a common practice in cloze-style QA systems. However, it would improve the utility of the system if the system could answer questions without prespecified answer candidates and produce any word/phrase in the text as an answer.

Another limitation that is common to all of the ASR-based deep neural models described in this paper (and other similar existing machine comprehension models) is that they assume that a single context is given when performing a machine comprehension task. During the stages of developing and evaluating machine comprehension technologies, it may be necessary to use problems that are well-defined and simple (ie, one context per question). However, such models may have limited utility in practice. If a user has a question but does not know the context or article in which the answer can be found, the user may be unable to utilize these systems. In an ideal scenario, the user should be able to query the systems without prespecifying the contexts, and the systems should be able to infer the answer by analyzing the contents of all documents in the datasets.

To address the above issues, in our future work we will expand our QA system in the following direction. First, we will modify our QA system so that it accepts a question without prespecified context and searches the entire dataset to find a subset of relevant documents. This search process can be implemented using BEST [20], which is a fast and efficient biomedical entity search tool that we developed in our previous research. Second, we will extract partial answers from each relevant document using our proposed machine comprehension model. The improved system will not require prespecified answer candidates. Finally, we will combine the partial answers from relevant contexts to form a final answer to the original query. Although searching for informative sources and expanding the proposed model to consider multiple sources would be a challenging task, we believe that this expanded system will be a useful tool for assisting biomedical scientists and practitioners by providing knowledge QA functionality in the medical domain.

Conclusions

In this paper, we introduced a new task of machine comprehension in the biomedical domain using a deep neural model. To the best of our knowledge, our work is the first to apply the deep learning-based machine comprehension task to the biomedical domain. Since there was no large-scale dataset in the biomedical domain for training the deep neural model, we created the new cloze-style datasets BMKC_T and BMKC_LS using the PubMed corpus. To improve the comprehension performance of the existing deep neural models, we used pretrained word vectors, entity type embedding, and ensemble techniques. The experimental results show that our proposed model's performance on the comprehension task is much higher than that of humans, including domain experts. In future work, we will expand our machine comprehension model so that it considers causal inference, concept hierarchy, and multiple documents to effectively answer complex questions.

Acknowledgments

This work was supported by the National Research Foundation of Korea grant, which was funded by the government of Korea (MSIP; 2014R1A2A1A10051238, 2016M3A9A7916996).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Human evaluation details on the BMKC dataset.

[PDF File (Adobe PDF File), 273KB - medinform_v6i1e2_app1.pdf]

Multimedia Appendix 2

Human evaluation of the CNN and CBT dataset using crowdsourcing.

[PDF File (Adobe PDF File), 17KB - medinform_v6i1e2_app2.pdf]

References

1. Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics 2015 Apr 30;16:138 [FREE Full text] [doi: 10.1186/s12859-015-0564-6] [Medline: 25925131]
2. Huang C, Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform 2016 Jan;17(1):132-144 [FREE Full text] [doi: 10.1093/bib/bbv024] [Medline: 25935162]
3. Holzinger A, Schantl J, Schroettner M, Seifert C, Verspoor K. Biomedical text mining: state-of-the-art, open problems and future challenges. In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. Berlin, Germany: Springer Science; 2014:271-300.
4. Fleuren WW, Alkema W. Application of text mining in the biomedical domain. Methods 2015 Mar;74:97-106. [doi: 10.1016/j.ymeth.2015.01.015] [Medline: 25641519]
5. Duz M, Marshall JF, Parkin T. Validation of an improved computer-assisted technique for mining free-text electronic medical records. JMIR Med Inform 2017 Jun 29;5(2):e17 [FREE Full text] [doi: 10.2196/medinform.7123] [Medline: 28663163]
6. Wongchaisuwat P, Klabjan D, Jonnalagadda SR. A semi-supervised learning approach to enhance health care community-based question answering: a case study in alcoholism. JMIR Med Inform 2016 Aug 02;4(3):e24 [FREE Full text] [doi: 10.2196/medinform.5490] [Medline: 27485666]
7. Hristovski D, Dinevski D, Kastrin A, Rindflesch TC. Biomedical question answering using semantic relations. BMC Bioinformatics 2015 Jan 16;16:6 [FREE Full text] [doi: 10.1186/s12859-014-0365-3] [Medline: 25592675]
8. Balikas G, Krithara A, Partalas I, Paliouras G. Bioasq: a challenge on large-scale biomedical semantic indexing and question answering. In: Multimodal Retrieval in the Medical Domain. New York: Springer International Publishing; 2015:26-39.
9. Asiaee AH, Minning T, Doshi P, Tarleton RL. A framework for ontology-based question answering with application to parasite immunology. J Biomed Semantics 2015;6:31 [FREE Full text] [doi: 10.1186/s13326-015-0029-x] [Medline: 26185615]
10. Neves M, Leser U. Question answering for biology. Methods 2015 Mar;74:36-46. [doi: 10.1016/j.ymeth.2014.10.023] [Medline: 25448292]
11. Gobeill J, Gaudinat A, Pasche E, Vishnyakova D, Gaudet P, Bairoch A, et al. Deep question answering for protein annotation. Database (Oxford) 2015 [FREE Full text] [doi: 10.1093/database/bav081] [Medline: 26384372]

12. Hermann KM, Kocisky T, Grefenstette E, Espeholt L, Kay W, Suleyman M, et al. Teaching machines to read and comprehend. In: Advances in Neural Information Processing Systems 28.: Curran Associates, Inc; 2015 Presented at: Neural Information Processing Systems 28 (NIPS 2015); December 07-12, 2015; Montreal, Canada p. 1693-1701.
13. Hill F, Bordes A, Chopra S, Weston J. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. 2015. URL: http://arxiv.org/abs/1511.02301 [accessed 2017-12-08] [WebCite Cache ID 6vYPXLCYy]
14. Kobayashi S, Tian R, Okazaki N, Inui K. Dynamic entity representations with max-pooling improves machine reading. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.: Association for Computational Linguistics; 2016 Jun Presented at: HLT-NAACL; 2016; San Diego, California p. 850-855 URL: http://www.aclweb.org/anthology/N16-1099 [doi: 10.18653/v1/N16-1099]
15. Kadlec R, Schmid M, Bajgar O, Kleindienst J. Text understanding with the attention sum reader network. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.: Association for Computational Linguistics; 2016 Presented at: The 54th Annual Meeting of the Association for Computational Linguistics; August 7-12, 2016; Berlin, Germany p. 908-918. [doi: 10.18653/v1/P16-1086]
16. Onishi T, Wang H, Bansal M, Gimpel K, McAllester D. Who did what: a large-ccale person-centered cloze dataset. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.: Association for Computational Linguistics; 2016 Presented at: The 2016 Conference on Empirical Methods in Natural Language Processing; November 1-5, 2016; Austin, Texas p. 2230-2235.
17. Dhingra B, Liu H, Yang Z, Cohen W, Salakutdinov R. Gated-attention readers for text comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.: ACL; 2017 Presented at: The 55th Annual Meeting of the Association for Computational Linguistics; July 2017; Vancouver, Canada p. 1832-1846 URL: http://www.aclweb.org/anthology/P17-1168 [doi: 10.18653/v1/P17-1168]
18. Taylor WL. “Cloze Procedure”: a new tool for measuring readability. Journalism Bulletin 2016 Oct 28;30(4):415-433. [doi: 10.1177/107769905303000401]
19. Kang J. BMKC: A dataset for BioMedical Knowledge Comprehension. 2017 Feb 14. URL: http://infos.korea.ac.kr/bmkc/ [accessed 2018-01-02] [WebCite Cache ID 6wBRhfS7o]
20. Lee S, Kim D, Lee K, Choi J, Kim S, Jeon M, et al. BEST: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. PLoS One 2016;11(10):e0164680 [FREE Full text] [doi: 10.1371/journal.pone.0164680] [Medline: 27760149]
21. Mnih V, Heess N, Graves A, Kavukcuoglu K. Recurrent models of visual attention. In: Advances in Neural Information Processing Systems 27. 2014 Presented at: Neural Information Processing Systems Conference; December 08-13, 2014; Montreal, Canada p. 2204-2212.
22. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, et al. Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning.: PMLR; 2015 Presented at: International Conference on Machine Learning; July 7-9, 2015; Lille, France p. 2048-2057.
23. Minh L, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.: Association for Computational Linguistics; 2015 Presented at: The 2015 Conference on Empirical Methods in Natural Language Processing; September 17-21, 2015; Lisbon, Portugal p. 1412-1421.
24. Chorowski J, Bahdanau D, Serdyuk D, Cho K, Bengio Y. Attention-based models for speech recognition. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press; 2015 Presented at: NIPS'15; December 07-12, 2015; Montreal, Canada p. 577-585 URL: https://dl.acm.org/citation.cfm?id=2969304
25. 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: The 5th International Symposium on Languages in Biology and Medicine; December 12-13, 2013; Tokyo, Japan p. 39-44 URL: http://lbm2013.biopathway.org/lbm2013proceedings.pdf
26. Biomedical natural language processing. 2017. Tools and resources URL: http://bio.nlplab.org/ [accessed 2017-10-12] [WebCite Cache ID 6u9BPx1WV]
27. Hansen L, Salamon P. Neural network ensembles. IEEE Trans Pattern Anal Machine Intell 1990;12(10):993-1001. [doi: 10.1109/34.58871]
28. Li K, Liu W, Zhao K, Shao M, Liu L. A novel dynamic weight neural network ensemble model. Int J Distrib Sens Netw 2015 Jan;11(8):862056. [doi: 10.1155/2015/862056]
29. CrowdFlower. 2017. URL: https://www.crowdflower.com/ [accessed 2018-01-02] [WebCite Cache ID 6wBRqVs6V]

Abbreviations

AI: artificial intelligence
ASR: Attention Sum Reader
ASR+BE: Attention Sum Reader + Bio-word Embedding
ASR+TE: Attention Sum Reader + Type Embedding

BEST: Biomedical Entity Search Tool
BMKC: Biomedical Knowledge Comprehension
BMKC_T: Biomedical Knowledge Comprehension Title
BMKC_LS: Biomedical Knowledge Comprehension Last Sentence
CBT: Children's Book Test
CBT_NE: Children's Book Test Named Entity
CBT_Noun: Children's Book Test Noun Phrase
CNN: Cable News Network
GRU: Gated Recurrent Unit
ID: identification
MeSH: Medical Subject Headings
NLP: natural language processing
QA: question answering

Edited by G Eysenbach; submitted 15.08.17; peer-reviewed by M Roccetti, L Zhang, G Lim; comments to author 07.09.17; revised version received 25.10.17; accepted 16.11.17; published 05.01.18.

Please cite as:
Kim S, Park D, Choi Y, Lee K, Kim B, Jeon M, Kim J, Tan AC, Kang J
A Pilot Study of Biomedical Text Comprehension using an Attention-Based Deep Neural Reader: Design and Experimental Analysis
JMIR Med Inform 2018;6(1):e2
URL: http://medinform.jmir.org/2018/1/e2/
doi:10.2196/medinform.8751
PMID:29305341

Original Paper

Automated Information Extraction on Treatment and Prognosis for Non–Small Cell Lung Cancer Radiotherapy Patients: Clinical Study

Shuai Zheng[1], PhD; Salma K Jabbour[2], MD; Shannon E O'Reilly[3], PhD; James J Lu[4], PhD; Lihua Dong[5], MD; Lijuan Ding[5], MD; Ying Xiao[3], PhD; Ning Yue[2], PhD; Fusheng Wang[6,7*], PhD; Wei Zou[3*], PhD

[1]Department of Biomedical Informatics, Emory University, Atlanta, GA, United States

[2]Department of Radiation Oncology, Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, United States

[3]Penn Medicine, Department of Radiation Oncology, University of Pennsylvania, Philadelphia, PA, United States

[4]Department of Mathematics and Computer Science, Emory University, Atlanta, GA, United States

[5]Department of Radiation Oncology, The First Hospital, Changchun, China

[6]Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, United States

[7]Department of Computer Science, Stony Brook University, Stony Brook, NY, United States

[*]these authors contributed equally

Corresponding Author:
Wei Zou, PhD
Penn Medicine
Department of Radiation Oncology
University of Pennsylvania
3400 Civic Center Blvd
Philadelphia, PA, 19104
United States
Phone: 1 215 866 7087
Email: wei.zou@uphs.upenn.edu

Abstract

Background: In outcome studies of oncology patients undergoing radiation, researchers extract valuable information from medical records generated before, during, and after radiotherapy visits, such as survival data, toxicities, and complications. Clinical studies rely heavily on these data to correlate the treatment regimen with the prognosis to develop evidence-based radiation therapy paradigms. These data are available mainly in forms of narrative texts or table formats with heterogeneous vocabularies. Manual extraction of the related information from these data can be time consuming and labor intensive, which is not ideal for large studies.

Objective: The objective of this study was to adapt the interactive information extraction platform Information and Data Extraction using Adaptive Learning (IDEAL-X) to extract treatment and prognosis data for patients with locally advanced or inoperable non–small cell lung cancer (NSCLC).

Methods: We transformed patient treatment and prognosis documents into normalized structured forms using the IDEAL-X system for easy data navigation. The adaptive learning and user-customized controlled toxicity vocabularies were applied to extract categorized treatment and prognosis data, so as to generate structured output.

Results: In total, we extracted data from 261 treatment and prognosis documents relating to 50 patients, with overall precision and recall more than 93% and 83%, respectively. For toxicity information extractions, which are important to study patient posttreatment side effects and quality of life, the precision and recall achieved 95.7% and 94.5% respectively.

Conclusions: The IDEAL-X system is capable of extracting study data regarding NSCLC chemoradiation patients with significant accuracy and effectiveness, and therefore can be used in large-scale radiotherapy clinical data studies.

(JMIR Med Inform 2018;6(1):e8) doi:10.2196/medinform.8662

KEYWORDS

information extraction; oncology; chemoradiation treatment; prognosis; non–small cell lung; information storage and retrieval; natural language processing

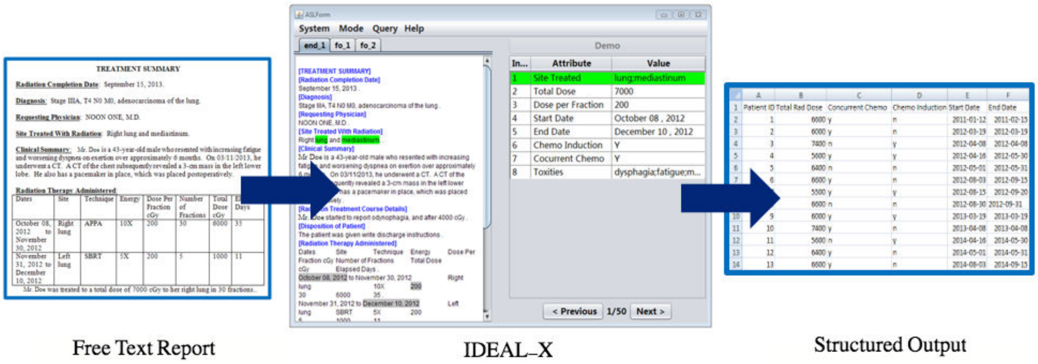
Introduction

Locally advanced or inoperable non–small cell lung cancer (NSCLC) occurs in approximately 20% to 30% of all cases of NSCLC [1] and may be treated with a combination of definitive concurrent chemotherapy and radiation. Modern radiotherapy has made great advances in the care of NSCLC patients, by reducing potential toxicities using involved field irradiation, while improving survival rates [2-4]. Assessing the effects of new developments in treatment techniques and regimens requires studies on the correlation between the treatment and prognosis [5-7]. Such studies involve extracting extensive patient information on chemoradiation treatments and follow-up assessments, including survival, tumor control, and toxicities.

Information about treatment and prognosis is embedded in treatment summaries and clinical encounter notes, which have various formats and diverse vocabularies. Manual extraction from large volumes of patient treatment summaries and records describing prognosis is time consuming and labor intensive. There is a need for an automated information system, as a natural language processing tool, to extract the needed patient treatment and prognosis data. During recent years, automated information systems have become widely used in medical and biomedical domains. The clinical Text Analysis and Knowledge Extraction System specializes in clinical information extraction [8]. The Cancer Tissue Information Extraction System focuses on annotating cancer text [9]. MedLEE supports connecting value to controlled vocabularies [10]. MedEx aims to extract medication-related information such as dosage and duration [11]. The Clinical Language Annotation, Modeling, and Processing toolkit integrates award-winning algorithms and, moreover, enables users to customize natural language processing components so as to encode clinical text automatically [12,13]. Medical text extraction processes pathology reports and uses rule-based methods to classify lung cancer stages [14]. A recent study also demonstrated that the metastatic site and status of lung cancer could be extracted from pathology reports using a pipeline [15]. Another study showed that cancer stage information could also be extracted with natural language processing [16]. Most traditional information extraction systems rely on batch training or predefined rules and were designed for only limited medical domains or tasks.

To support a retrospective study of NSCLC chemoradiotherapy patients, we adapted our in-house–developed information extraction platform, Information and Data Extraction using Adaptive Learning (IDEAL-X; X represents controlled vocabulary) system [17-19]. This information extraction system aims to transform free-text clinical documents into structured data and has been used by projects in cardiology and pathology. IDEAL-X possesses unique features different from the systems mentioned above: (1) users may freely customize attributes to be extracted; (2) the system extracts information from narrative medical documents and generates normalized values to populate output tables and assist manual annotation; (3) it requires no mandatory configurations or training before performing annotation and adaptive learning processes; and (4) the system learns from users' normal interactions transparently, and establishes and refines decision models incrementally, which further alleviates manual annotation efforts. Figure 1 shows how the IDEAL-X system processes the input from free-text reports generated during physician and patient encounters and delivers structured output.

Figure 1. Screenshot of the Information and Data Extraction using Adaptive Learning (IDEAL-X) platform, and example input and output.



TREATMENT SUMMARY

Radiation Completion Date: September 15, 2013.

Diagnosis: Stage IIIA, T4 N0 M0, adenocarcinoma of the lung.

Requesting Physician: NOON ONE, M.D.

Site Treated With Radiation: Right lung and mediastinum.

Clinical Summary: Mr. Doe is a 43-year-old male who resented with increasing fatigue and worsening dyspnea on exertion over approximately 6 months. On 03/11/2013, he underwent a CT. A CT of the chest subsequently revealed a 3-cm mass in the left lower lobe. He also has a pacemaker in place, which was placed postoperatively.

Radiation Therapy Administered

Dates	Site	Technique	Energy	Dose Per Fraction cGy	Number of Fractions	Total Dose cGy	Elapsed Days
October 08, 2012 to November 30, 2012	Right lung	APPA	10X	200	30	6000	35
November 31, 2012 to December 10, 2012	Left lung	SBRT	5X	200	5	1000	11

Mr. Doe was treated to a total dose of 7000 cGy to her right lung in 30 fractions.

[TREATMENT SUMMARY]
[Radiation Completion Date]
September 15, 2013 .
[Diagnosis]
Stage IIIA, T4 N0 M0, adenocarcinoma of the lung .
[Requesting Physician]
NOON ONE, M.D .
[Site Treated With Radiation]
Right lung and mediastinum .
[Clinical Summary]
Mr Doe is a 43-year-old male who resented with increasing fatig and worsening dyspnea on exertion over approximately 6 m On 03/11/2013, he underwent a CT. A CT of the quently revealed a 3-cm mass in the left lower has a pacemaker in place, which was placed ely .
[Ra on Treatment Course Details]
Mr. Doe started to report odynophagia, and after 4000 cGy .
[Disposition of Patient]
The patient was given write discharge instructions .
[Radiation Therapy Administered]
Dates Site Technique Energy Dose Per Fraction cGy Number of Fractions Total Dose cGy Elapsed Days .
October 08, 2012 to November 30, 2012 Right lung 10X 200 30 6000 35 .
November 31, 2012 to December 10, 2012 Left lung SBRT 5X 200 5 1000 11

In...	Attribute	Value
1	Site Treated	lung;mediastinum
2	Total Dose	7000
3	Dose per Fraction	200
4	Start Date	October 08 , 2012
5	End Date	December 10 , 2012
6	Chemo Induction	Y
7	Cocurrent Chemo	Y
8	Toxities	dysphagia;fatigue;m...

	A	B	C	D	E	F
1	Patient ID	Total Rad Dose	Concurrent Chemo	Chemo Induction	Start Date	End Date
2	1	6600	y	n	2011-01-12	2011-02-15
3	2	6000	y	n	2012-03-19	2012-03-19
4	3	7400	n	y	2012-04-08	2012-04-08
5	4	5600	y	y	2012-04-16	2012-05-30
6	5	6400	n	n	2012-05-01	2012-05-31
7	6	6600	y	n	2012-08-03	2012-09-15
		5500	y	y	2012-08-15	2012-09-20
		6600	n	n	2012-08-30	2012-09-31
10	9	6000	y	y	2013-03-19	2013-03-19
11	10	7400	y	n	2013-04-08	2013-04-08
12	11	5600	n	y	2014-04-16	2014-05-30
13	12	6400	y	n	2014-05-01	2014-05-31
14	13	6600	y	n	2014-08-03	2014-09-15

Methods

Patient Information

We collected NSCLC patient data to investigate the relationship between shrinkage of the treated tumor and each category of prognosis data: survival, tumor control, and toxicities. The patient treatment data we needed to identify included the chemoradiotherapy drugs used, dose, and treatment time frame. From the follow-up clinical notes, we needed to extract tumor control information diagnosed from the patient's follow-up computed tomography and positron emission tomography images, patient toxicities, and complication data, including skin, internal organ, blood, and overall body reactions to treatment. We further categorized toxicities into different toxicity grades [20]. After we extracted the information in a structured format, we intended to use it to statistically correlate treatment tumor shrinkage with survival time, disease control rate, and the toxicities.

From studies approved by the institutional review boards of both Rutgers University and Emory University, we retrospectively identified 50 patients who had primary unresectable, locally advanced, biopsy-proven stage II-III NSCLC, and who had received chemoradiotherapy with a median follow-up of 22 months. In total, we exported 261 treatment and patient follow-up documents from the patient electronic health record system ARIA (Varian Medical Systems, Inc, Palo Alto, CA, USA) and anonymized the data for this study.

IDEAL-X System Development

We adapted the IDEAL-X system to support automated information extraction from the NSCLC chemoradiation patients' documents. After a requirement analysis, we added new features, such as extracting timex and parsing tabular information, to enhance the original system. We also implemented corresponding feature extraction and machine learning processes for timex and tabular formats, and constructed the dictionary to assist toxicity data extraction. We extracted patient information, such as treatment time frame and chemoradiotherapy, from treatment records with an adaptive learning process (Table 1). In extracting this information, the system began without any prior training and created its machine learning model incrementally. During the information extraction of the toxicities, the adaptive learning process was disabled. We used the dictionary shown in Textbox 1 to aid in toxicities information extraction. Along with extracted values, the sentences where the values were embedded were also output in a spreadsheet, which could be used for further manual toxicity grade differentiation based on patient Common Terminology Criteria for Adverse Events guidelines v 4.0, which were designated previously in the patient charts [20].

In addition, to verify the extracted data, we asked 2 physicians to manually annotate these reports. We used the manually annotated ground truth to validate the automatically generated output from the IDEAL-X system. We used precision and recall results to estimate the effectiveness of extraction.

IDEAL-X Adaptive Learning Process

Through adaptive learning, IDEAL-X established its decision model through ordinary operations in manual annotation. First, the user designated the value to fill every attribute in the structured output form. After a few initial documents, the system quickly learned important and related information that the user sought and began to generate standardized values automatically in subsequent documents. The system continued to learn and update its knowledge, without special user intervention. This incremental learning process made the system domain agnostic and not limited to a specific medical report. When available, a user-defined controlled dictionary and other configurations could also be provided by the user to facilitate this learning process, but they were not mandatory.

System Data Flow

Figure 2 demonstrates the system's data flow. Each time that the system loaded a document, the system moved through the preprocessing phase and parsed the text to analyze and identify important linguistic features and natural language elements. These features and elements included (1) part of speech: the part-of-speech tag of each word, for example, noun and verb; (2): timex: the system relied on predefined regular expressions to identify timex, such as 2010-01-09 and Sep 13, 2013, and then indexed them based on their position in the text; (3) tabular information: the system identified and parsed tables in input text to comprehend underlying relations between values and the metadata in a table; (4) negation terms: the system detected negation terms and regions being affected, for example, in the case of "patient denies fever and fatigue," "fever" and "fatigue" were not extracted as part of the toxicities; and (5) uncertain terms: the system identified uncertain phrases and regions being governed, for example, "We explained to her that the risks of the treatment included dysphagia and pneumonitis" meant that dysphagia and pneumonitis had not appeared yet as symptoms. We used these features to mark the input text and provide detailed linguistic indications during extraction.

After preprocessing, the parsed text was investigated by the automated annotation component of the system to populate the output form automatically. First, sentences where possible values may be located were extracted based on text hierarchy, frequently co-occurring terms, previously extracted values, or user-customized vocabularies. The system then identified candidate phrases from located sentences using either a hidden Markov model [21] chunker or a dictionary chunker. Subsequently, candidate values were examined by various filters based on linguistic features such as part of speech, certainty, or negation collected during preprocessing. After filtering, the sentence score and the chunk score were combined, on the basis of which a classifier determined the overall confidence score of each candidate value and categorized it as "accept" or "reject."

We then reviewed the automatically extracted values manually for the purpose of adaptive learning. We considered positive and negative scenarios: if the user navigated to the next document without changing any values, we regarded the values generated by the system as positive training cases; if the user modified any values, we regarded the system-generated values as negative training cases and the manually updated values as positive ones. We used the results of the review to support further improvements in the automated annotation component. Difference feature extract procedures, which model the traits of numerical, nominal, timex, and tabular data elements, were applied to corresponding positive and negative instances. By repeating these steps, the system became intelligent incrementally and delivered more accurate results.

Table 1. Information extracted from treatment records of patients with non–small cell lung cancer.

Attributes	Text data type	Numbers of values	Dictionary	Adaptive learning
Treatment site	Nominal	68	N/A[a]	Yes
Chemotherapy information	Nominal	56	N/A	Yes
Treatment time frame	Date	92	N/A	Yes
Radiation therapy dose	Numerical	97	N/A	Yes
Toxicities	Nominal	331	Yes	N/A

[a] N/A: not applicable.

Textbox 1. Dictionary of toxicities.

Anemia

Lymphopenia

Anorexia

Dehydration

Dyspnea

Fatigue

Mucosal inflammation

Radiation esophagitis

Weight decrease

Cough

Febrile neutropenia

Neutropenia

Bronchitis

Diarrhea

Esophagitis

Hyponatremia

Nausea

Radiation pneumonitis

Dermatitis

Leukopenia

Thrombocytopenia

Decreased appetite

Dysphagia

Failure to thrive

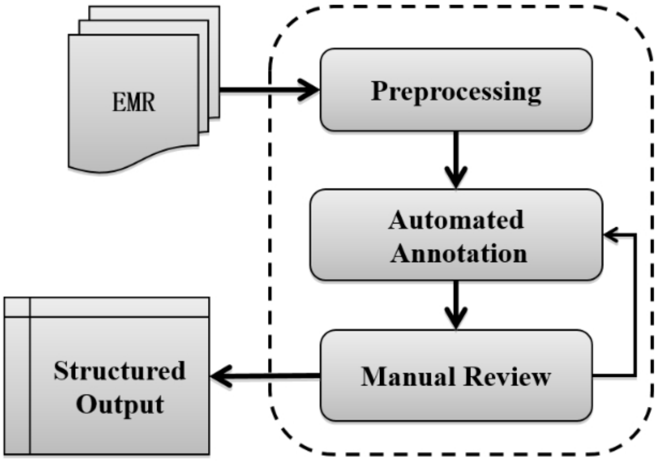
Localized infection

Pneumonia

Vomiting

Insomnia

Figure 2. Data flow in the Information and Data Extraction using Adaptive Learning (IDEAL-X) platform. EMR: electronic medical record.



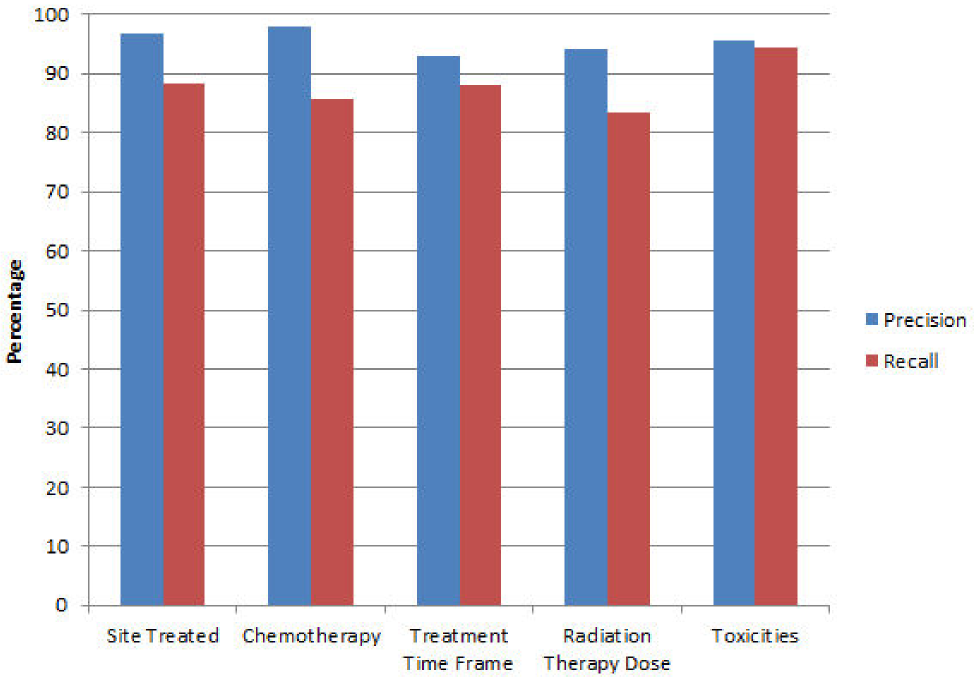
Results

Figure 3 shows the validation results against the manually annotated ground truth. In the validation for patient characteristics and tumor control, the system achieved an overall precision of over 93%. The recall values of all information were more than 83%. The recalls were lower than the precisions, as the recalls reflected the performance during the overall adaptive learning process—the system processed a few documents to construct and refine its decision model at its early stage in the adaptive learning process.

Especially in the extraction of the toxicities, the negation detection and certainty detection filters contributed directly to the accuracy of extraction. With the help of a controlled dictionary, the system achieved an overall precision of 95.7% and recall of 94.5%.

Within 1 second, a well-trained system can process patient documents of multiple pages and output the results in a predefined format. Compared with manual review, which requires reading through the entire document and manually annotating the notes on each patient, this system significantly improved the efficiency of information extraction.

Figure 3. Effectiveness of data extraction as estimated by precision and recall of automatically generated output compared with manually annotated ground truth.



Discussion

IDEAL-X employed adaptive learning and a controlled vocabulary to support information extraction, which alleviated both the training and the deployment processes that could be expensive in applying a traditional information extraction system. The various data types IDEAL-X supports cover the most important and common information in oncology reports, which delivers great usability to our use case. We have demonstrated the great advantage of this system in greatly improving information extraction effectiveness while maintaining high accuracy when applied to extracting NSCLC patient treatment and prognoses data from heterogeneous document formats. In addition, because the system improves its performance incrementally, its accuracy could be further improved with additional training documents. Once trained, the developed system was able to process further fed-in reports in batch mode without revision. Without an intervening regular manual reporting process that handles input documents in sequence, the system accumulates knowledge transparently to empower the task and, therefore, could be conveniently integrated into a regular clinical workflow. The technology it used was domain agnostic and, therefore, could be transformed to other disease sites and studies in radiation oncology.

Limitations

In the validation analysis, the system also revealed some unavoidable limitations. The system identified and comprehended information based on explicitly expressed keywords. For example, the phrases “neoadjuvant chemo” and “upfront chemotherapy” may be used as keywords to identify chemotherapy induction. However, in situations where relevant information is distributed across different regions in the text, more insightful comprehension becomes necessary. For example, in the case of “After 4 cycles of chemotherapy and abdomen...we began radiation...,” the system was not intelligent enough to interpret the meaning of “4 cycles” as “neoadjuvant chemotherapy” behind the narrations. In general, this sophisticated scenario reveals the limitation of this information extraction-based approach. The system requires explicit keywords or hints to determine an event; however, it cannot reason and analyze factors collected from different sources. Such cases resulted in lower recalls for chemotherapy than for other attributes and demanded a manual review. Therefore, to facilitate the manual review, we output the associated sentence with the extracted information together in tabular format for user manual review and validation at a later time.

Conclusion

We adapted the IDEAL-X system to automatically extract treatment and prognostic information for stage II and III NSCLC patients who had received chemoradiation. With this system, patient information was extracted efficiently from their medical documents in various formats. The system, together with minimized manual review efforts, generated outputs with high precision and recall. It significantly improved the effectiveness and can be easily applied to other radiation oncology patient studies at larger scales.

Conflicts of Interest

None declared.

References

1. Ramalingam S, Belani C. Systemic chemotherapy for advanced non-small cell lung cancer: recent advances and future directions. Oncologist 2008;13 Suppl 1:5-13 [FREE Full text] [doi: 10.1634/theoncologist.13-S1-5] [Medline: 18263769]
2. Furuse K, Fukuoka M, Kawahara M, Nishikawa H, Takada Y, Kudoh S, et al. Phase III study of concurrent versus sequential thoracic radiotherapy in combination with mitomycin, vindesine, and cisplatin in unresectable stage III non-small-cell lung cancer. J Clin Oncol 1999 Sep;17(9):2692-2699. [doi: 10.1200/JCO.1999.17.9.2692] [Medline: 10561343]
3. Belani CP, Choy H, Bonomi P, Scott C, Travis P, Haluschak J, et al. Combined chemoradiotherapy regimens of paclitaxel and carboplatin for locally advanced non-small-cell lung cancer: a randomized phase II locally advanced multi-modality protocol. J Clin Oncol 2005 Sep 01;23(25):5883-5891. [doi: 10.1200/JCO.2005.55.405] [Medline: 16087941]
4. Liao ZX, Komaki RR, Thames HD, Liu HH, Tucker SL, Mohan R, et al. Influence of technologic advances on outcomes in patients with unresectable, locally advanced non-small-cell lung cancer receiving concomitant chemoradiotherapy. Int J Radiat Oncol Biol Phys 2010 Mar 01;76(3):775-781. [doi: 10.1016/j.ijrobp.2009.02.032] [Medline: 19515503]
5. Bral S, De Ridder M, Duchateau M, Gevaert T, Engels B, Schallier D, et al. Daily megavoltage computed tomography in lung cancer radiotherapy: correlation between volumetric changes and local outcome. Int J Radiat Oncol Biol Phys 2011 Aug 01;80(5):1338-1342. [doi: 10.1016/j.ijrobp.2010.04.002] [Medline: 20638192]
6. Aupérin A, Le Pechoux C, Rolland E, Curran WJ, Furuse K, Fournel P, et al. Meta-analysis of concomitant versus sequential radiochemotherapy in locally advanced non-small-cell lung cancer. J Clin Oncol 2010 May 01;28(13):2181-2190. [doi: 10.1200/JCO.2009.26.2543] [Medline: 20351327]
7. Jabbour SK, Kim S, Haider SA, Xu X, Wu A, Surakanti S, et al. Reduction in tumor volume by cone-beam computed tomography predicts overall survival in non-small cell lung cancer treated with chemoradiation therapy. Int J Radiat Oncol Biol Phys 2015 Jul 01;92(3):627-633 [FREE Full text] [doi: 10.1016/j.ijrobp.2015.02.017] [Medline: 26068495]
8. 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]
9. Crowley RS, Castine M, Mitchell K, Chavan G, McSherry T, Feldman M. caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. J Am Med Inform Assoc 2010;17(3):253-264 [FREE Full text] [doi: 10.1136/jamia.2009.002295] [Medline: 20442142]

10. FierceBiotech. Columbia grants Health Fidelity exclusive license to MedLEE NLP. Newton, MA: Questex; 2012 Jan 11. URL: http://www.fiercebiotech.com/biotech/columbia-grants-health-fidelity-exclusive-license-to-medlee-nlp [accessed 2017-08-07] [WebCite Cache ID 6sXdtEJnH]
11. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc 2010;17(1):19-24 [FREE Full text] [doi: 10.1197/jamia.M3378] [Medline: 20064797]
12. CLAMP: Clinical Language Annotation, Modeling, and Processing Toolkit. Houston, TX: The University of Texas Health Science Center at Houston; 2018. URL: http://clamp.uth.edu/index.php [accessed 2018-01-18] [WebCite Cache ID 6wZIVTGfK]
13. Lee H, Zhang Y, Xu J, Moon S, Wang J, Wu Y, et al. UTHealth at SemEval-2016 Task 12: an end-to-end system for temporal information extraction from clinical notes. 2016 Presented at: 10th International Workshop on Semantic Evaluation (SemEval-2016); June 16-17, 2016; San Diego, CA, USA.
14. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc 2010;17(4):440-445 [FREE Full text] [doi: 10.1136/jamia.2010.003707] [Medline: 20595312]
15. Soysal E, Warner JL, Denny JC, Xu H. Identifying metastases-related information from pathology reports of lung cancer patients. AMIA Jt Summits Transl Sci Proc 2017;2017:268-277 [FREE Full text] [Medline: 28815141]
16. Warner JL, Levy MA, Neuss MN, Warner JL, Levy MA, Neuss MN. ReCAP: feasibility and accuracy of extracting cancer stage information from narrative electronic health record data. J Oncol Pract 2016 Feb;12(2):157-8; e169. [doi: 10.1200/JOP.2015.004622] [Medline: 26306621]
17. Zheng S, Lu JJ, Ghasemzadeh N, Hayek SS, Quyyumi AA, Wang F. Effective information extraction framework for heterogeneous clinical reports using online machine learning and controlled vocabularies. JMIR Med Inform 2017 May 09;5(2):e12 [FREE Full text] [doi: 10.2196/medinform.7235] [Medline: 28487265]
18. Zheng S, Lu JJ, Appin C, Brat D, Wang F. Support patient search on pathology reports with interactive online learning based data extraction. J Pathol Inform 2015;6:51 [FREE Full text] [doi: 10.4103/2153-3539.166012] [Medline: 26605116]
19. Zheng S, Wang F, Lu JJ. ASLForm: an adaptive self learning medical form generating system. AMIA Annu Symp Proc 2013;2013:1590-1599 [FREE Full text] [Medline: 24551429]
20. Common Terminology Criteria for Adverse Events (CTCAE) version 4.03. Washington, DC: U.S. Department of Health and Human Services, National Institutes of Health, and National Cancer Institute; 2010 Jun 14. URL: https://evs.nci.nih.gov/ftp1/CTCAE/CTCAE_4.03_2010-06-14_QuickReference_5x7.pdf [accessed 2018-01-18] [WebCite Cache ID 6wZIjlhp3]
21. Elliott RJ, Aggoun L, Moore JB. Hidden Markov Models: Estimation and Control. New York, NY: Springer; 1994.

Abbreviations

EMR: electronic medical record
IDEAL-X: Information and Data Extraction using Adaptive Learning
NSCLC: non–small cell lung cancer

Edited by G Eysenbach; submitted 07.08.17; peer-reviewed by C Tao, P Zarogoulidis; comments to author 21.09.17; revised version received 15.11.17; accepted 01.12.17; published 01.02.18.

Please cite as:
Zheng S, Jabbour SK, O'Reilly SE, Lu JJ, Dong L, Ding L, Xiao Y, Yue N, Wang F, Zou W
Automated Information Extraction on Treatment and Prognosis for Non–Small Cell Lung Cancer Radiotherapy Patients: Clinical Study
JMIR Med Inform 2018;6(1):e8
URL: http://medinform.jmir.org/2018/1/e8/
doi:10.2196/medinform.8662
PMID:29391345

Original Paper

A Clinical Decision Support Engine Based on a National Medication Repository for the Detection of Potential Duplicate Medications: Design and Evaluation

Cheng-Yi Yang[1,2*], PhD; Yu-Sheng Lo[1*], PhD; Ray-Jade Chen[3,4], MSc, MD; Chien-Tsai Liu[1], PhD

[1]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan

[2]Department of Medical Informatics, Industrial Technology Research Institute, Hsinchu, Taiwan

[3]Department of Surgery, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan

[4]Taipei Medical University Hospital, Taipei, Taiwan

[*]these authors contributed equally

Corresponding Author:
Chien-Tsai Liu, PhD
Graduate Institute of Biomedical Informatics
College of Medical Science and Technology
Taipei Medical University
250 Wuxing St
Taipei, 11030
Taiwan
Phone: 886 266382736 ext 1509
Email: ctliu@tmu.edu.tw

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2019/3/e15063/

Abstract

Background: A computerized physician order entry (CPOE) system combined with a clinical decision support system can reduce duplication of medications and thus adverse drug reactions. However, without infrastructure that supports patients' integrated medication history across health care facilities nationwide, duplication of medication can still occur. In Taiwan, the National Health Insurance Administration has implemented a national medication repository and Web-based query system known as the PharmaCloud, which allows physicians to access their patients' medication records prescribed by different health care facilities across Taiwan.

Objective: This study aimed to develop a scalable, flexible, and thematic design-based clinical decision support (CDS) engine, which integrates a national medication repository to support CPOE systems in the detection of potential duplication of medication across health care facilities, as well as to analyze its impact on clinical encounters.

Methods: A CDS engine was developed that can download patients' up-to-date medication history from the PharmaCloud and support a CPOE system in the detection of potential duplicate medications. When prescribing a medication order using the CPOE system, a physician receives an alert if there is a potential duplicate medication. To investigate the impact of the CDS engine on clinical encounters in outpatient services, a clinical encounter log was created to collect information about time, prescribed drugs, and physicians' responses to handling the alerts for each encounter.

Results: The CDS engine was installed in a teaching affiliate hospital, and the clinical encounter log collected information for 3 months, during which a total of 178,300 prescriptions were prescribed in the outpatient departments. In all, 43,844/178,300 (24.59%) patients signed the PharmaCloud consent form allowing their physicians to access their medication history in the PharmaCloud. The rate of duplicate medication was 5.83% (1843/31,614) of prescriptions. When prescribing using the CDS engine, the median encounter time was 4.3 (IQR 2.3-7.3) min, longer than that without using the CDS engine (median 3.6, IQR 2.0-6.3 min). From the physicians' responses, we found that 42.06% (1908/4536) of the potential duplicate medications were recognized by the physicians and the medication orders were canceled.

Conclusions: The CDS engine could easily extend functions for detection of adverse drug reactions when more and more electronic health record systems are adopted. Moreover, the CDS engine can retrieve more updated and completed medication histories in the PharmaCloud, so it can have better performance for detection of duplicate medications. Although our CDS engine approach could enhance medication safety, it would make for a longer encounter time. This problem can be mitigated by careful evaluation of adopted solutions for implementation of the CDS engine. The successful key component of a CDS engine is the completeness of the patient's medication history, thus further research to assess the factors in increasing the PharmaCloud consent rate is required.

(JMIR Med Inform 2018;6(1):e6) doi:10.2196/medinform.9064

KEYWORDS

duplicate medication; adverse drug reaction; clinical decision support system; PharmaCloud

Introduction

Duplication of medication can be defined as a patient being prescribed more than two medications of the same therapeutic class (including different doses, forms, frequencies, or routes) within an overlapping period, with one of the prescriptions being clinically redundant [1-3]. The duplication of medication orders is a critical issue that can result in some patients being affected by adverse drug reactions (ADRs) [4-6]. The potential for duplication of medications has increased, with patients visiting a greater number of different hospitals and following more extensive medication regimens. This issue particularly affects elderly patients and those suffering from chronic diseases [7-9]. A previous study indicated that physicians and pharmacists can help reduce unnecessary prescriptions and optimize a patient's drug therapy regimen by examining a patient's full medication record [1]. Reducing duplicate medications and treatment can contribute significantly to preventing ADRs.

In addition, duplication of medications increases overall medical expenditures [4-6], causes serious environmental pollution, and wastes medical and social resources [10,11]. Each year in Taiwan, more than 3 tons of prescribed medications go unused [11]. A study also indicated that 8.8% of outpatients received duplicate medications across different health care facilities in Japan [12]. In the United Kingdom, approximately £300 million of medicines prescribed by the National Health Service are wasted each year [10]. The issue of duplicate medications and its impact on patient safety have received attention in several countries.

More and more information and communication technologies, such as clinical decision support systems (CDSSs), have been proposed as a solution for improving medication safety. Many studies have suggested that a computerized physician order entry (CPOE) system combined with a CDSS could help in preventing ADRs [5,6,13], thereby reducing medication expenditure [14]. However, even when using such a system, duplication of medication can still occur due to a lack of an infrastructure supporting the integration of patients' medication records prescribed by different health care facilities in general, including clinics, doctor offices, medical centers, or large hospitals. Thus, when a patient is transferred from one hospital to another, or visits more than one hospital for the same condition, physicians may not be aware of the medication prescribed at other hospitals and may prescribe duplicate medications despite using a CPOE system with a CDSS.

In Taiwan, the National Health Insurance Administration (NHIA) has implemented approaches for sharing patients' medical care information nationwide, including health smart cards [4] and a Web-based medication query system based on a national medication repository known as the PharmaCloud [4,15]. The PharmaCloud contains the most complete and up-to-date version of a patient's medication history. According to NHIA policy, health care facilities must upload a patient's prescribed medications to the PharmaCloud within 24 hours after the patient's visit [15,16].

Currently, the PharmaCloud stores the latest 3 months of each patient's prescribed medication records. It supports two access modes for authorized clinical professionals. One is an online query through a Web browser interface; the other is a batch download. By using the online query mode, an authorized physician can access patients' medication histories in the PharmaCloud through a Web browser. When the physician wants to prescribe medication orders for a patient, he or she can use the Web browser to submit queries about the patient's medication history prior to ordering a prescription. A physician can check that information on the browser manually and then use that information to make independent decisions about the prescription to avoid prescribing duplicate medications. Most health care facilities encourage their physicians to use the online query mode to access the PharmaCloud. However, physicians are usually very busy and so this approach may not be feasible.

In the batch download mode, a patient's medication history can be downloaded from the PharmaCloud provided the patient has signed an informed consent form allowing the authorized physicians access and has made an appointment at least one day in advance. We will refer to the informed consent form as the PharmaCloud consent form. In this approach, the downloaded patients' medication histories have to be integrated into a CPOE system so that the CPOE can verify the prescription to see if there is any potential duplicate medication and other ADRs. However, CPOE systems are complex because they must access data from various systems within a hospital. Furthermore, electronic health record (EHR) systems are usually adopted incrementally [17,18]. Thus, a new approach to design a flexible and scalable decision support system that integrates the PharmaCloud and a CPOE system to prevent duplicate medications and other ADR events is needed.

In this study, we developed a modularized clinical decision support (CDS) engine that can support duplicate medication checks based on the PharmaCloud. We also analyzed the impact of the CDS engine on patient encounter time and physicians' responses to handling potential duplicate medication alerts. These results could provide insights to adopt the CDS engine and recommendations to improve the efficiency in medication safety checks.

Methods

Settings

For this study, the CDS engine was developed and installed at Taipei Medical University Hospital, a teaching hospital with nearly 800 beds. The hospital has a highly informative infrastructure and is a certified Healthcare Information and Management Systems Society EHR Adoption Model stage 6 hospital [19]. At this hospital, the backend repositories and databases of CPOE, online registration/appointment, and drug information management systems have been integrated. Although a CDS engine may perform many decision support functions, at this stage the implemented CDS engine supported the function of duplicate medication checking only.

Framework of the Clinical Decision Support Engine and its Interactions With the PharmaCloud and Computerized Physician Order Entry Systems

The framework of the CDS engine and interactions with the PharmaCloud and a CPOE system are presented in Figure 1. The implemented CDS engine consisted of four major components: the PharmaCloud adapter, CDS engine local repository, the duplicate medication checker, and the CDS engine adapter as described subsequently.

The PharmaCloud Adapter

The PharmaCloud adapter is used to access a patient's visit appointment information registered in the patient appointment system and to verify whether the patient signed the PharmaCloud consent form for PharmaCloud access. If so, the PharmaCloud adapter retrieves the patient's last 3 months of medication records from the PharmaCloud via batch download over the National Health Insurance (NHI) virtual private network (VPN).

Clinical Decision Support Engine Local Repository

The CDS engine local repository was implemented using the PostgreSQL relational database system [20] to store the patients' medication history data retrieved from the PharmaCloud. The medication history contains all medication records prescribed in the last 3 months by the health care facilities in Taiwan. The medication record contains information including the Anatomical Therapeutic Chemical (ATC) Classification name, NHI drug code, drug ingredients, drug name, prescribing date, number of days it was prescribed for, and the number of days of drug treatment remaining.

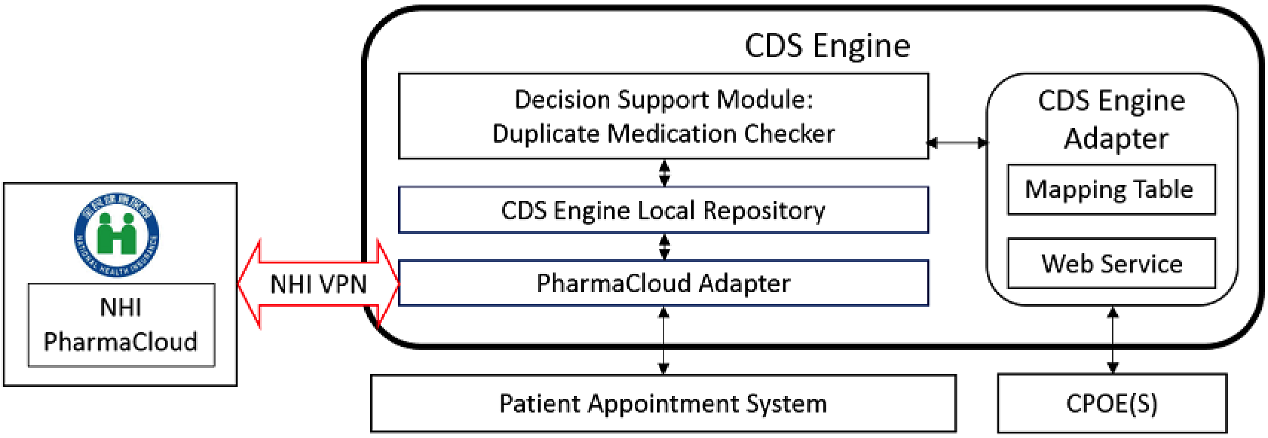
Decision Support Module: Duplicate Medication Checker

The decision support module is a decoupled, thematic design approach that allows health care facilities to add, update, and delete customized medication verification modules (eg, duplicate medication, drug-drug interaction, and maximum dosage). The duplicate medication checker was one of the verification modules used in this study. It consists of a set of logic and rules for detecting duplicate medications. Duplicate medication is primarily identified using the ATC system [2,3]. We defined potential duplicate medication as two prescribed drugs (not necessary in the same prescription, but with an overlap between their start date and stop date) that had the same ATC level 4 codes (ie, the first four digits of the ATC codes are identical) [15,21-24]. The NHIA has released a cross-mapping table between NHI drug codes and ATC codes [15].

Clinical Decision Support Engine Adapter

The CDS engine adapter is an interface between the CDS engine and a CPOE system that allows the CPOE system to initiate the duplicate medication checker. It performs mapping functions between the hospital's drug codes, NHI drug codes, and ATC codes. When a physician wants to prescribe a drug for a patient, the CPOE system sends the drug details, including the patient's identification, drug code, start date, and stop date, to the duplicate medication checker via the CDS engine adapter, which then converts the drug details into a form that can be interpreted by the duplicate medication checker. After checking for duplicate medication, the duplicate medication checker returns the result to the CPOE system via the CDS engine adopter.

Figure 1. The integrated computerized physician order entry (CPOE) system and clinical decision support (CDS) engine for detecting potential duplicate medications. NHI: National Health Insurance; VPN: virtual private network.


CDS Engine
Decision Support Module: Duplicate Medication Checker
CDS Engine Local Repository
PharmaCloud Adapter
CDS Engine Adapter
Mapping Table
Web Service
NHI PharmaCloud
NHI VPN
Patient Appointment System
CPOE(S)

Extension of Decision Support Function

The CDS engine is different from the traditional CDSS coupled with CPOE. The CDS engine has a decoupled decision support module from the hard-coded rules in CPOE. The module provides a thematic decision rules design approach. Health care facilities are able to maintain several independent thematic CDS modules for different CDS applications. These independent configurable knowledge rule modules allow CPOE to invoke few configurations and decrease the code change of the original CPOE. The scalable and flexible nature of the framework facilitates health care facilities to integrate the CDS function into their existing CPOE system. The steps involved in the extension of the CDS engine rule module are as follows:

1. Defining the theme of the decision support module. In this study, we defined a duplicate medication checker to distinguish duplicate medications. The theme of the decision support module can be extended using a drool file [25]. Creating a new decision support module or modifying the original decision support module is possible in the CDS engine.
2. Defining the input and output parameters and the decision support logics in the decision support module. Health care facilities must select the input parameters from the CPOE and local repository for the decision support logics, and those that are to be returned to the CPOE. The design of the decision support logics may be based on relevant clinical guidelines, regulations, protocols, or medication knowledge.
3. Retrieving the EHRs from the local repository. In our study, the EHRs were retrieved from PharmaCloud through the PharmaCloud adapter. In other scenarios, health care facilities could add other EHR adapters to retrieve different EHR sources.
4. Adding a Web service path to the CDS engine adapter. A health care facility can add a Web service URL for CPOE to invoke the added decision support module.
5. CPOE has an AJAX [26] Web service call from the CDS engine adapter to invoke the decision support module in a CPOE textbox; thus, physicians are alerted when prescribing medications.

Information Security Framework

To ensure a certain level of safety in storing medical information, we adopted some information security assumptions for both the EHR repositories and CPOE, such as secure tunnel, access control, and privacy control protection. In the secure tunnel, as PharmaCloud is deployed in the NHI VPN environment, the CDS engine must access the PharmaCloud through the NHI VPN. In the access control, we must have both the physician’s Healthcare Certification Authority card and the patient’s health smart card simultaneously inserted into the card reader to verify that the physician has the authority to access the patient’s medication history. Finally, the patient must sign the PharmaCloud consent form before the CDS engine batch downloads their medication history; if not, the CDS engine would not retrieve the patient’s medication history.

Workflow for Detecting Potential Duplicate Medication Across Health Care Facilities With the Clinical Decision Support Engine

Patients who wish to allow their physicians at a health care facility to access their medication history in the PharmaCloud must complete the PharmaCloud consent form and submit it to the health care facility. When a patient wants to visit a doctor, he or she makes an appointment and registers in advance by using the patient appointment system of the health facility. If the patient’s consent is in effect at the time of the visit, the CDS engine retrieves the patient’s medication history for the past 3 months from the PharmaCloud and stores it into the CDS engine local repository. To evaluate the impact of the CDS engine on an outpatient clinical encounter, a clinical encounter log iss created to collect information about the patient and physician, the start and end time of the clinical encounter, the drugs prescribed by the physician, and the physician’s responses to potential duplicate medication alerts, if any.

Figure 2 shows the prescription workflow of a clinical encounter using the CDS engine. First, the physician’s Healthcare Certification Authority card and the patient’s health smart card are simultaneously inserted into a card reader to initiate the clinical encounter. The CPOE system reads the physician’s and patient’s information. This information and the start time of the encounter are recorded into the clinical encounter log. The physician then conducts the patient assessment and diagnosis for the patient. If the patient does not require any medication, the workflow ends. If the patient requires medication, the physician prescribes a drug via the CPOE system. The CPOE system verifies whether the patient has signed the PharmaCloud consent form. If not, the physician simply uses the CPOE system to prescribe the drug without using the CDS engine. If the patient has signed the PharmaCloud consent form, the CPOE invokes the CDS engine duplicate medication checker to perform a duplicate medication check. The prescribed drug and the check result are also recorded into the clinical encounter log.

If the duplicate medication checker detects a potential duplicate medication (ie, the prescribed drug's ATC level 4 code is the same as the one stored in the CDS engine local repository), it sends an alert to the CPOE system. The alert information, including drug name, ATC code, and start and stop dates, is then displayed on a pop-up screen (Figure 3, upper panel). Our hospital requires the physician to provide a reason for prescribing the duplicate drug in order to meet the NHI payment policy. Thus, the physician can click on one of the check buttons (Figure 3, middle) and then proceed to prescribe the subsequent drug by clicking on the "Continue" button. If the physician does not select a reason, he or she has to click on the "Cancel" button to revoke the prescribed drug. The reason for prescribing the duplicate drug and the physician's response are recorded in the clinical encounter log. If no duplicate medication is found, the physician can continue prescribing drugs until no further drug prescription is required. Finally, the physician withdraws the patient's health smart card from the card reader to end the clinical encounter. The ending time is also recorded in the clinical encounter log.

Figure 2. The prescription workflow using the clinical decision support (CDS) engine for detection of potential duplicate medication.

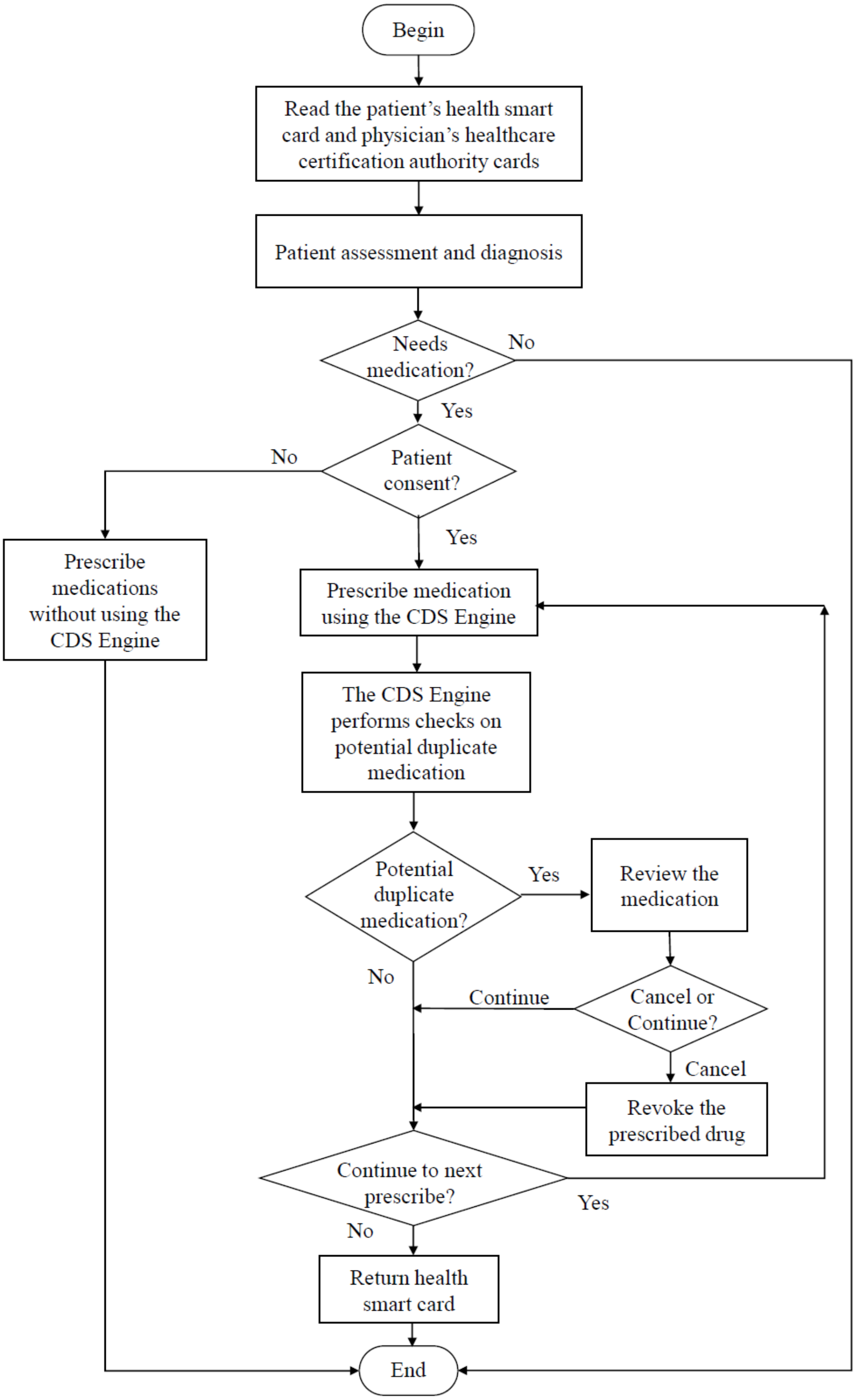
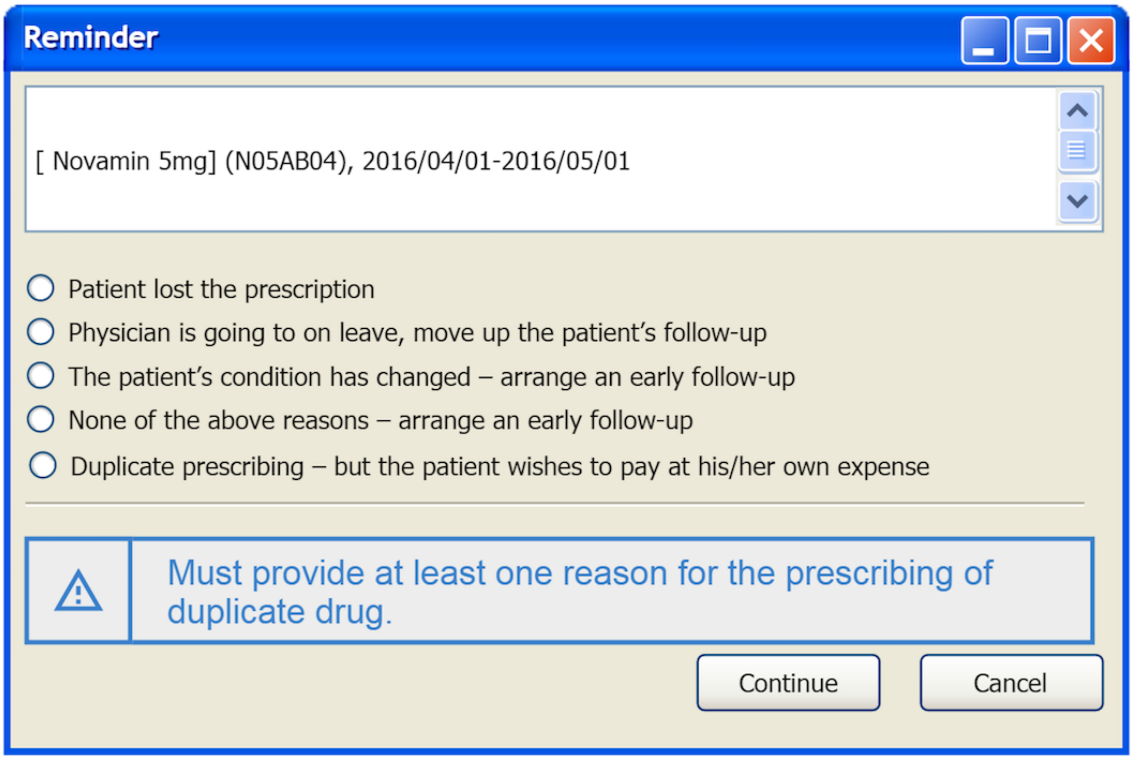

Begin
Read the patient's health smart card and physician's healthcare certification authority cards
Patient assessment and diagnosis
Needs medication?
No
Yes
No
Patient consent?
Yes
Prescribe medications without using the CDS Engine
Prescribe medication using the CDS Engine
The CDS Engine performs checks on potential duplicate medication
Potential duplicate medication?
Yes
Review the medication
No
Continue
Cancel or Continue?
Cancel
Revoke the prescribed drug
Continue to next prescribe?
Yes
No
Return health smart card
End

Figure 3. A screenshot of a pop-up screen showing an alert message that appears when a potential duplicate medication is detected. The screen presents the information about the duplicate drug (upper panel), response options for reasons for prescribing the medication (middle), and action to take (lower panel).



Reminder

[Novamin 5mg] (N05AB04), 2016/04/01-2016/05/01

○ Patient lost the prescription
○ Physician is going to on leave, move up the patient’s follow-up
○ The patient’s condition has changed – arrange an early follow-up
○ None of the above reasons – arrange an early follow-up
○ Duplicate prescribing – but the patient wishes to pay at his/her own expense

⚠ Must provide at least one reason for the prescribing of duplicate drug.

Analysis of Impacts of the Clinical Decision Support Engine on Outpatient Services

The CDS engine has been integrated into a CPOE system of the hospital to support order entry processes, and this combined system has operated in four outpatient departments: medicine, surgery, gynecology-pediatrics, and “other” departments. The clinical encounter log was initiated to collect information about clinical encounters from April 1 to June 30, 2016. During this period, for each patient’s clinical encounter, the log collected the starting and ending time of the encounter, prescribed medication, and the physician’s response(s) to a potential duplicate medication alert, if any. The log could be used to analyze how the CDS engine affected the encounter time and to determine physicians’ responses to potential duplicate medication alerts in different outpatient departments.

To investigate clinical encounter time, we divided clinical encounters into two groups based on the clinical encounter log: “with CDS engine” and “without CDS engine.” Patients in the without CDS engine group were those who did not sign a PharmaCloud consent form; patients in the with CDS engine group were those who signed the PharmaCloud consent form. We also analyzed the characteristics of the consent rate, potential duplicate medication rate, and physicians’ response(s) to any potential duplicate medication alerts. We used the statistical software R version 3.3.1 [27] to perform a Wilcoxon rank sum test with a 5% level of significance to assess the difference in encounter time between the without CDS engine and the with CDS engine groups.

Results

Overall, there were 178,300 patient visits to the four outpatient departments during the 3-month period, as shown in Table 1. There were 43,844 (24.59%) of patient visits in which the patients signed the PharmaCloud consent form, allowing their physicians to access their medication history stored in the PharmaCloud. That is, there were 43,844 and 134,456 patient visits in the with CDS engine and without CDS engine groups, respectively. In the without CDS engine group, there were 96,714 (71.93%) patient visits in which at least one medication order was prescribed. In the with CDS engine group, there were 31,614 (72.11%) patient visits in which at least one medication order was prescribed. Among these, 4227 (13.37%) prescriptions resulted in potential duplicate medication.

Table 2 shows the clinical encounter time in the with CDS engine and without CDS engine groups. A Wilcoxon rank sum test showed that the clinical encounter time in the with CDS engine group (median 4.3, IQR 2.3-7.3 min) was significantly longer than that in the without CDS engine group (median 3.6, IQR 2.0-6.3 min). A similar pattern was observed in the medicine, surgery, and other departments, but not in the gynecology-pediatrics department, where there was no significant difference between the groups. This might be because there are usually more medical checkups and procedures than medication treatments in gynecology-pediatrics outpatient services.

Table 1. Analysis of clinical encounters information during the 3-month data collection period.

Clinical encounters	With CDS[a] engine (n=43,844)	Without CDS engine (n=134,456)	Total (N=178,300)
No medicine prescribed, n (%)	12,230 (27.89)	37,742 (28.07)	49,972 (28.03)
Medicine prescribed, n (%)	31,614 (72.11)	96,714 (71.93)	128,328 (71.97)
Potential duplicate medication, n	4227	—	
No potential duplicate medication, n	27,387	—	

[a] CDS: clinical decision support.

Table 2. Differences in clinical encounter time between with clinical decision support (CDS) engine and without CDS engine groups.

Departments	Without CDS engine		With CDS engine		*P*
	n	Median time in minutes, (IQR)	n	Median time in minutes, (IQR)	
Medicine	49,310	3.7 (2.2-6.2)	17,189	4.5 (2.6-7.4)	<.01
Surgery	26,885	3.1 (1.5-5.8)	9031	3.7 (1.8-7.0)	<.01
Gynecology-Pediatrics	9129	4.5 (2.5-7.7)	2209	4.6 (2.3-8.0)	.92
Other	11,390	3.8 (2.2-6.4)	3185	4.3 (2.4-7.2)	<.01
Total	96,714	3.6 (2.0-6.3)	31,614	4.3 (2.3-7.3)	<.01

Table 3. Physicians’ responses to potential duplicate medications.

Department	Physician response, n (%)						
	Cancel[a]	Lost prescription[b]	Physician on leave[c]	Condition change[d]	Other[e]	Self-pay[f]	Total
Medicine	1049 (36.32)	87 (3.01)	91 (3.15)	905 (31.34)	528 (18.28)	228 (7.89)	2888 (100)
Surgery	603 (55.89)	38 (3.52)	9 (0.83)	226 (20.95)	155 (14.37)	48 (4.45)	1079 (100)
Gynecology-Pediatrics	88 (40.18)	11 (5.02)	0 (0)	88 (40.18)	16 (7.31)	16 (7.31)	219 (100)
Other	168 (48.00)	5 (1.43)	5 (1.43)	82 (23.43)	21 (6.00)	69 (19.71)	350 (100)
Total	1908 (42.06)	141 (3.11)	105 (2.31)	1301 (28.68)	720 (15.87)	361 (7.96)	4536 (100)

[a] Cancel: confirmed as a duplicate drug—cancel this drug.

[2] Lost prescription: the patient lost the prescription.

[c] Physician on leave: the physician is going on leave, plan earlier patient follow-up.

[d] Condition change: the patient’s condition has changed—arrange an early follow-up.

[e] Other: none of the above reasons—arrange an early follow-up.

[f] Self-pay: duplicate prescribing, but the patient wishes to pay at his or her own expense.

An alert was trigged by a drug in a prescription if it was detected as a potential duplicate medication. Among the 4227 potential duplicate medication prescriptions (Table 1, 13.37% of 31,614 prescriptions), more than two potential duplicate medication alerts occurred for 170 (totaling 479 potential duplicate medication alerts); therefore, a total of 4536 potential duplicate medication alerts were responded to by physicians in Table 3. In summary, 42.06 % (1908) of the alerts led to cancelation of the duplicate drugs to be prescribed (ie, clicked “Cancel” button). The remaining 2628 alerts did not lead to a cancelation response and were issued for the following reasons: 28.68% (1301) for “condition change” (the patient’s condition has changed—arrange an early follow-up), 15.87% (720) for “others” (none of the above reasons—arrange an early follow-up), 7.96% (361) for “self-pay” (duplicate prescribing, but the patient wishes to pay at his/her own expense for some reasonable reasons), 3.11% (141) for “lost prescription,” and 2.31% (105) for “physician on leave.” Notably, the most common reason for issuing duplicate drugs was “condition change.” The alerts enabled the physicians to review their prescriptions once again and, consequently, to prevent the duplication of medications.

In Table 4, the results show that 1843 prescriptions were confirmed as duplicate medication prescriptions from the 4227 potential duplicate medication prescriptions, or that 5.83% of prescriptions (1843/31,614) might result in duplicate medications.

Table 4. Prescriptions confirmed as duplicate medication prescriptions.

Department	Potential duplicate medication, n (%)	Canceled drug[a], n	CDS[b] engine prescription[c], n	Canceled drugs/CDS engine prescriptions, %
Medicine	2685 (63.52)	1025	17,189	5.96
Surgery	1019 (24.11)	576	9031	6.38
Gynecology-Pediatrics	215 (5.09)	86	2209	3.89
Others	308 (7.29)	156	3185	4.90
Total	4227 (100)	1843	31,614	5.83

[a] The prescription had at least one drug confirmed as a duplicate drug and the doctor canceled this drug.

[b] CDS: clinical decision support.

[c] The prescription checked with the CDS engine.

Discussion

Principal Results

In our study, we developed a CDS engine to access the PharmaCloud (a national medication repository) to retrieve the medication records of patients from the previous 3 months. As per Taiwan NHI policies, health care facilities are required to upload their patients' prescriptions within 24 hours after a visit. Thus, the CDS engine can access a fairly complete and up-to-date medication history. A previous study [22] showed that incomplete or delayed sharing of EHRs across health care facilities made it difficult for a CDSS to perform thorough checks of potential duplicate medications, resulting in a duplicate medication detection rate of 2.4%. Our approach increased the previously reported duplicate medication detection rate from 2.4% to 5.83% of total prescriptions. It shows that the more complete the medication history is, the better the protection from duplicate medication.

Nowadays, medication safety is a top priority for both patients and health care providers. However, it requires additional cost. Under our CDS engine framework, clinical encounter time was slightly (0.7 min) longer than when a CPOE system was used alone (from 3.6 min to 4.3 min) despite the ability to enhance medication safety. However, as we adopt more advanced and faster communication and computer technology to build CDS engines, the increased time can be mitigated. Thus, while implementing the CDS engine, to guarantee the desired level of medication safety, we should carefully evaluate the adopted solutions to meet the time requirements in clinical practice. Although the CDS engine takes an additional time of 0.7 min, it is still more efficient than the previous methods used in Taiwan [28]. Previously, CPOE invoked medication information stored in the health smart card to support medication decisions, but the prescription information was written by the physician for each patient, which required additional time (1.88 min) and resulted in more time utilized than when using the CDS engine with PharmaCloud.

Knowledge-based CDS generally offers two categories: "stand-alone CDS" and "CDS coupled with CPOE" [29]. The former is not directly integrated into the clinical workflow; a physician must enter patient information into both CPOE and CDSS, which can cause double data entry. The physician has to switch between the systems. Furthermore, it could not issue reminders when the physician prescribed medication using CPOE. Thus, this method is time-consuming and may compromise the clinical process, particularly during busy clinical practice. The latter one prevented double data entry and issued reminders when the physician prescribed medications. However, in the context of ever-improving and new medical knowledge, new clinical guidelines, regulations, policies, and EHRs (in general, due to limited budgets and resources, EHR systems are usually adopted incrementally by health care facilities [17,18]), CDS functions must be kept updated to prevent use of outdated knowledge [30]. Previous studies have shown that CDS rules are usually hard-coded or tight-bundled with CPOE or incorporated into CPOE [31,32]; thus, the CPOE program has to be updated once the rules are updated. This maintenance of rapidly changing knowledge and systems with complex rule sets can be expensive [29]. Therefore, we provided an innovative CDS engine and adopted the CDS engine to detect potential duplicate medication in this study. Firstly, the CDS engine provides a decoupled decision support module, a thematic decision rules design approach. Health care facilities can design or update these independent configurable CDS knowledge modules separately, such as potential duplicate medication checks, drug-drug interactions, and drug allergies. This would increase the scalability and extensibility of CDS with CPOE. Secondly, the CDS engine adapter, a service-oriented architecture (SOA)-based design, can provide a Web application programming interface for CPOE to invoke on demand the thematic CDS module with few configurations. It can also reduce the change in the original CPOE.

Moreover, as described in the Methods section ("Extension of Decision Support Function"), we can add EHR adapters to retrieve other EHR repositories not restricted in PharmaCloud. As increasing numbers of EHRs gradually become available, our CDS engine approach can rapidly meet the changing requirements of CDSS to provide more complete medical histories and ensure added safety. Other studies also indicated that the future adoption of CDS would ideally be modularized [33] with SOA design pattern [29] to account for the ever-changing medical knowledge. Thus, the innovative CDS engine framework can fulfill the trends in the field of medicine. The NHIA will continue to provide additional medical records, such as laboratory test results, surgeries, dental procedures, controlled drug management, and medicine allergies. With a more complete EHR, future iterations of the CDS engine would focus on the integration of more EHR repositories and designing of more CDS theme modules (ie, drug-drug interactions, dose calculations, pregnancy medication reminders, and allergy reminders) to extend CDS coverage to the health care facilities.

Electronic health records are generally recognized as key components of CDSSs [34]. Integrating EHRs with CDS functions is likely to become a widespread trend [32]. Many countries have taken steps to develop relevant infrastructures and regulations and provide incentive policies to facilitate the adoption and integration of EHR systems [35-37]. However, few studies have discussed that the CPOE framework automatically invokes the centralized national EHR for decision support when physicians prescribe medications. In Australia, a well-established nationalized EHR repository, My Health Record, primarily allows residents to maintain and share their health records with their general physicians or health care facilities. The Australian Digital Health Agency is currently investigating the secondary use of My Health Record [38], and our CDS engine application scenarios can be a reference for other countries who own centralized EHR repositories and are seeking secondary use applications.

Our study showed that 24.59% of patients signed the PharmaCloud consent form to allow their physicians to access their medication records. In our context, direct CDS engine users include authorized physicians, not patients. However, only after the patients sign the PharmaCloud consent form can the physicians access the patients' medication histories in the PharmaCloud from their CPOE system. The lack of patient participation in the PharmaCloud system would increase the difficulty associated with the efficient implementation of protections against duplicate medication prescription across health care facilities. In Australia, My Health Record was originally adopted with opt-in consent; however, because the uptake rate remained low, two trial areas adopted opt-out consent until 2016 to increase the uptake rate. As a result, the uptake rates were obviously higher in these areas than in the non-trial areas [39-41]. Based on the implementation of the My Health Record system, we suggest that NHIA should consider adopting an opt-out rather than opt-in approach to increase the PharmaCloud usage for the CDS engine to provide a more comprehensive CDS support in health care facilities. Although an opt-out system is likely to increase the use of PharmaCloud, there are many complex factors and different national conditions that could affect patients' participation in an innovative information system [42,43] and further research is needed to assess such factors.

Limitations

The system proposed in this study has certain limitations. Firstly, the PharmaCloud batch mode requires patient consent to access his or her medication history prior to their hospital visit. In our approach, the PharmaCloud system provided only a batch download mode for the CDS engine to retrieve patient medication history, and the CDS engine local repository was updated once daily, usually at midnight. Therefore, it was not possible to use the CDS engine for checking prescriptions of walk-in patients. A previous study indicated that walk-in patients represent approximately 44% of total patients [44]. Thus, for such patients, physicians can only use a Web browser to access the PharmaCloud system to manually check the patient's medication history. Secondly, due to our PharmaCloud consent rate of only 24.59%, the potential duplicate medication detection rate may have been underestimated. Finally, the CDS engine was adopted in a single teaching hospital; thus, the impact of adopting the CDS engine cannot be generalized to other hospitals at a national or international level.

Conclusion

In this study, we developed a modularized CDS engine to access a national medication repository, the PharmaCloud, for detection of duplicate medication across health care facilities in Taiwan. Because of the modularized design, the CDS engine could easily extend functions for detection of ADR events when more and more EHR systems are adopted. Moreover, the CDS engine can retrieve more updated and completed patients' medication histories in the PharmaCloud, so it can have better performance for detection of duplicate medication.

Although our CDS engine approach could enhance medication safety, it would make encounter time longer. Fortunately, this problem can be mitigated by careful evaluation of adopted solutions for implementation of the CDS engine.

Because the PharmaCloud system provided batch download mode only for the CDS engine to retrieve patients' medication history, the CDS engine local repository could not be updated in a timely manner. Thus, the CDS engine might not be able to provide walk-in patients with protection from duplicate medication. To tackle this problem, we suggest PharmaCloud should consider the opt-out consent policy to increase the usability of the CDS engine to provide more comprehensive CDS support in health care facilities.

Acknowledgments

This work was partially supported by the Industrial Technology Research Institute of Taiwan (ITRI) with Project Grant No F301AR9A10. The authors would like to express our appreciation to Alina, Kent Chen, and the colleagues of Information Management Department, Taipei Medical University Hospital, for their technical support.

Conflicts of Interest

None declared.

References

1. Abookire SA, Teich JM, Sandige H, Paterno MD, Martin MT, Kuperman GJ, et al. Improving allergy alerting in a computerized physician order entry system. Proc AMIA Symp 2000:2-6 [FREE Full text] [Medline: 11080034]
2. WHO Collaborating Centre for Drug Statistics Methodology. 2011. ATC Structure and principles URL: https://www.whocc.no/atc/structure_and_principles/ [accessed 2017-12-23] [WebCite Cache ID 6vuzV5Qwv]
3. Coloma PM, Schuemie MJ, Trifirò G, Gini R, Herings R, Hippisley-Cox J, EU-ADR Consortium. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol Drug Saf 2011 Jan;20(1):1-11. [doi: 10.1002/pds.2053] [Medline: 21182150]
4. Huang S, Wang P, Tseng W, Syu F, Lee M, Shih R, et al. NHI-PharmaCloud in Taiwan--a preliminary evaluation using the RE-AIM framework and lessons learned. Int J Med Inform 2015 Oct;84(10):817-825. [doi: 10.1016/j.ijmedinf.2015.06.001] [Medline: 26113462]
5. Aronson JK. Medication errors: definitions and classification. Br J Clin Pharmacol 2009 Jun;67(6):599-604 [FREE Full text] [doi: 10.1111/j.1365-2125.2009.03415.x] [Medline: 19594526]
6. Magid S, Forrer C, Shaha S. Duplicate orders: an unintended consequence of computerized provider/physician order entry (CPOE) implementation: analysis and mitigation strategies. Appl Clin Inform 2012;3(4):377-391 [FREE Full text] [doi: 10.4338/ACI-2012-01-RA-0002] [Medline: 23646085]
7. Pretorius RW, Gataric G, Swedlund SK, Miller JR. Reducing the risk of adverse drug events in older adults. Am Fam Physician 2013 Mar 01;87(5):331-336 [FREE Full text] [Medline: 23547549]
8. Pharmacy A, Murphy J, Lee M. PSAP 2015 Book 2 CNS/Pharmacy Practice: Pharmacotherapy Self-Assessment Program. Lenexa, KS: American College of Clinical Pharmacy; 2015.
9. Naples JG, Hanlon JT, Schmader KE, Semla TP. Recent literature on medication errors and adverse drug events in older adults. J Am Geriatr Soc 2016 Feb;64(2):401-408 [FREE Full text] [doi: 10.1111/jgs.13922] [Medline: 26804210]
10. Bertie H, Robson R. NHS England. 2015. Pharmaceutical waste reduction in the NHS URL: https://www.england.nhs.uk/wp-content/uploads/2015/06/pharmaceutical-waste-reduction.pdf [accessed 2017-12-23] [WebCite Cache ID 6vv047pIS]
11. Chien H, Ko J, Chen Y, Weng S, Yang W, Chang Y, et al. Study of Medication Waste in Taiwan. J Exp Clin Med 2013 Apr;5(2):69-72. [doi: 10.1016/j.jecm.2013.02.003]
12. Kinoshita H, Kobayashi Y, Fukuda T. Duplicative medications in patients who visit multiple medical institutions among the insured of a corporate health insurance society in Japan. Health Policy 2008 Jan;85(1):114-123. [doi: 10.1016/j.healthpol.2007.07.003] [Medline: 17728002]
13. Agrawal A. Medication errors: prevention using information technology systems. Br J Clin Pharmacol 2009 Jun;67(6):681-686 [FREE Full text] [doi: 10.1111/j.1365-2125.2009.03427.x] [Medline: 19594538]
14. Kuperman GJ, Bobb A, Payne TH, Avery AJ, Gandhi TK, Burns G, et al. Medication-related clinical decision support in computerized provider order entry systems: a review. J Am Med Inform Assoc 2007 Sep;14(1):29-40. [doi: 10.1197/jamia.M2170] [Medline: 17068355]
15. National Health Insurance Administration Ministry of Health and Welfare. 2016 Apr 14. Increased cloud-based health care efficiency enhances medication safety URL: https://www.nhi.gov.tw/English/News_Content.aspx?n=996D1B4B5DC48343&sms=F0EAFEB716DE7FFA&s=4562FE367994D3C8 [accessed 2018-01-08] [WebCite Cache ID 6vv1aAIRy]
16. Chen C, Kuo L, Cheng K, Shen W, Bai K, Wang C, et al. The effect of medication therapy management service combined with a national PharmaCloud system for polypharmacy patients. Comput Methods Programs Biomed 2016 Oct;134:109-119. [doi: 10.1016/j.cmpb.2016.07.008] [Medline: 27480736]
17. Sequist TD, Cullen T, Hays H, Taualii MM, Simon SR, Bates DW. Implementation and use of an electronic health record within the Indian Health Service. J Am Med Inform Assoc 2007;14(2):191-197 [FREE Full text] [doi: 10.1197/jamia.M2234] [Medline: 17213495]
18. Huerta T, Thompson M, Ford E, Ford W. Implementing electronic lab order entry management in hospitals: incremental strategies lead to better productivity outcomes. Int J Inf Manag 2013 Feb;33(1):40-47. [doi: 10.1016/j.ijinfomgt.2012.05.008]
19. HIMSS Analytics. 2017. Electronic medical record adoption model URL: http://www.himssanalytics.org/emram [accessed 2017-12-23] [WebCite Cache ID 6vv1rl9Hn]
20. PostgreSQL. PostgreSQL 2017 URL: https://www.postgresql.org/ [accessed 2017-12-23] [WebCite Cache ID 6vv21IgUS]
21. Long A, Chang P, Li Y, Chiu W. The use of a CPOE log for the analysis of physicians' behavior when responding to drug-duplication reminders. Int J Med Inform 2008 Aug;77(8):499-506. [doi: 10.1016/j.ijmedinf.2007.10.002] [Medline: 18182321]
22. Hsu M, Yeh Y, Chen C, Liu C, Liu C. Online detection of potential duplicate medications and changes of physician behavior for outpatients visiting multiple hospitals using national health insurance smart cards in Taiwan. Int J Med Inform 2011 Mar;80(3):181-189. [doi: 10.1016/j.ijmedinf.2010.11.003] [Medline: 21183402]
23. Liu C, Yang P, Yeh Y, Wang B. The impacts of smart cards on hospital information systems--an investigation of the first phase of the national health insurance smart card project in Taiwan. Int J Med Inform 2006 Feb;75(2):173-181. [doi: 10.1016/j.ijmedinf.2005.07.022] [Medline: 16125452]

24. Wetterneck TB, Walker JM, Blosky MA, Cartmill RS, Hoonakker P, Johnson MA, et al. Factors contributing to an increase in duplicate medication order errors after CPOE implementation. J Am Med Inform Assoc 2011;18(6):774-782 [FREE Full text] [doi: 10.1136/amiajnl-2011-000255] [Medline: 21803925]
25. Drools. 2017. URL: http://www.drools.org/ [accessed 2017-12-23] [WebCite Cache ID 6vv29O8v]
26. MDN Web Docs. 2017. Ajax URL: https://developer.mozilla.org/en-US/docs/Web/Guide/AJAX [accessed 2017-12-23] [WebCite Cache ID 6vv6YXJne]
27. The R Foundation. 2017. What is R? URL: https://www.r-project.org/about.html [accessed 2017-12-23] [WebCite Cache ID 6vv2Wt3Ng]
28. Yeh Y, Hsu M, Chen C, Lo Y, Liu C. Detection of potential drug-drug interactions for outpatients across hospitals. Int J Environ Res Public Health 2014 Jan 27;11(2):1369-1383 [FREE Full text] [doi: 10.3390/ijerph110201369] [Medline: 24473112]
29. Berner E. Clinical Decision Support Systems: Theory and Practice (Health Informatics) 3rd Edition. Cham, Switzerland: Springer; 2016.
30. Hicks JK, Dunnenberger HM, Gumpper KF, Haidar CE, Hoffman JM. Integrating pharmacogenomics into electronic health records with clinical decision support. Am J Health Syst Pharm 2016 Dec 01;73(23):1967-1976 [FREE Full text] [doi: 10.2146/ajhp160030] [Medline: 27864204]
31. Sim E, Tan D, Abdullah H. The use of computerized physician order entry with clinical decision support reduces practice variance in ordering preoperative investigations: A retrospective cohort study. Int J Med Inform 2017 Dec;108:29-35. [doi: 10.1016/j.ijmedinf.2017.09.015] [Medline: 29132628]
32. Melton BL. Systematic review of medical informatics-supported medication decision making. Biomed Inform Insights 2017;9:1178222617697975 [FREE Full text] [doi: 10.1177/1178222617697975] [Medline: 28469432]
33. Busis N. How can I choose the best electronic health record system for my practice? Neurology 2010 Nov 02;75(18 Suppl 1):S60-S64. [doi: 10.1212/WNL.0b013e3181fc9888] [Medline: 21041774]
34. HealthIT.gov. 2015. Meaningful use URL: https://www.healthit.gov/providers-professionals/meaningful-use-definition-objectives [accessed 2017-12-23] [WebCite Cache ID 6vv31Pan0]
35. Jha AK, Doolan D, Grandt D, Scott T, Bates DW. The use of health information technology in seven nations. Int J Med Inform 2008 Dec;77(12):848-854. [doi: 10.1016/j.ijmedinf.2008.06.007] [Medline: 18657471]
36. Adler-Milstein J, DesRoches C, Furukawa M, Worzala C, Charles D, Kralovec P, et al. More than half of US hospitals have at least a basic EHR, but stage 2 criteria remain challenging for most. Health Aff (Millwood) 2014 Sep;33(9):1664-1671. [doi: 10.1377/hlthaff.2014.0453] [Medline: 25104826]
37. Mack D, Zhang S, Douglas M, Sow C, Strothers H, Rust G. Disparities in primary care EHR adoption rates. J Health Care Poor Underserved 2016 Feb;27(1):327-338 [FREE Full text] [doi: 10.1353/hpu.2016.0016] [Medline: 27587942]
38. Australian Digital Health Agency. 2017. My Health Record URL: https://myhealthrecord.gov.au/internet/mhr/publishing.nsf/Content/home [accessed 2017-12-23] [WebCite Cache ID 6vv3qCn0w]
39. Torrens E, Walker S. Demographic characteristics of Australian health consumers who were early registrants for opt-in personally controlled electronic health records. Health Inf Manag 2017 Sep;46(3):127-133. [doi: 10.1177/1833358317699341] [Medline: 28537210]
40. Walsh L, Hill S, Allan M, Balandin S, Georgiou A, Higgins I, et al. A content analysis of the consumer-facing online information about My Health Record: implications for increasing knowledge and awareness to facilitate uptake and use. Health Inf Manag 2017 Jan 01;59(30):1833358317712200. [doi: 10.1177/1833358317712200] [Medline: 28589741]
41. Donna SS. Wentworth Healthcare provider of the Nepean Blue Mountains PHN. 2017. My Health Record-applying lessons learned from the Opt Out Trial NBMPHN URL: http://www.phcris.org.au/phplib/filedownload.php?file=/elib/lib/downloaded_files/conference/presentations/8658_conf_abstract_msdonnas.pdf [accessed 2017-12-23] [WebCite Cache ID 6vv3TXXDP]
42. Partel K. Deeble Institute Issues Brief 30. 2015 Oct 30. Toward better implementation: Australia's My Health Record URL: http://ahha.asn.au/publication/issue-briefs/deeble-institute-issues-brief-no-13-toward-better-implementation-australias [accessed 2018-01-16] [WebCite Cache ID 6wWWgo0fx]
43. Xu J, Quaddus M. Exploring the factors influencing end users' acceptance of knowledge management systems: development of a research model of adoption and continued use. J Organ End User Com 2007;19(4):57-79. [doi: 10.4018/joeuc.2007100104]
44. Yu X, Hooft P, Delooz H. Emergency department walk-in patients study. Eur J Emerg Med 1996 Sep;3(3):163-174. [Medline: 9023495]

Abbreviations

ADR: adverse drug reaction
ATC: Anatomical Therapeutic Chemical
CDS: clinical decision support
CDSS: clinical decision support system

CPOE: computerized physician order entry
EHR: electronic health record
NHI: National Health Insurance
NHIA: National Health Insurance Administration
SOA: service-oriented architecture
VPN: virtual private network

Edited by G Eysenbach; submitted 30.09.17; peer-reviewed by F Lai, P Chai; comments to author 15.10.17; revised version received 23.12.17; accepted 23.12.17; published 19.01.18.

Please cite as:
Yang CY, Lo YS, Chen RJ, Liu CT
A Clinical Decision Support Engine Based on a National Medication Repository for the Detection of Potential Duplicate Medications: Design and Evaluation
JMIR Med Inform 2018;6(1):e6
URL: http://medinform.jmir.org/2018/1/e6/
doi:10.2196/medinform.9064
PMID:

Review

Quality of Decision Support in Computerized Provider Order Entry: Systematic Literature Review

Delphine Carli[1,2], PharmD, PhD; Guillaume Fahrni[3], MSc; Pascal Bonnabry[1,2], PhD; Christian Lovis[3,4], MD, PhD

[1]Division of Pharmacy, University Hospitals of Geneva, Geneva, Switzerland

[2]School of Pharmaceutical Sciences, University of Geneva, University of Lausanne, Geneva, Switzerland

[3]Division of Medical Information Sciences, University Hospitals of Geneva, Geneva, Switzerland

[4]School of Medicine, University of Geneva, Geneva, Switzerland

Corresponding Author:
Delphine Carli, PharmD, PhD
Division of Pharmacy
University Hospitals of Geneva
Rue Gabrielle-Perret-Gentil 4
Geneva, 1211
Switzerland
Phone: 41 786532871
Fax: 41 223726255
Email: delphine.carli@chuv.ch

Abstract

Background: Computerized decision support systems have raised a lot of hopes and expectations in the field of order entry. Although there are numerous studies reporting positive impacts, concerns are increasingly high about alert fatigue and effective impacts of these systems. One of the root causes of fatigue alert reported is the low clinical relevance of these alerts.

Objective: The objective of this systematic review was to assess the reported positive predictive value (PPV), as a proxy to clinical relevance, of decision support systems in computerized provider order entry (CPOE).

Methods: A systematic search of the scientific literature published between February 2009 and March 2015 on CPOE, clinical decision support systems, and the predictive value associated with alert fatigue was conducted using PubMed database. Inclusion criteria were as follows: English language, full text available (free or pay for access), assessed medication, direct or indirect level of predictive value, sensitivity, or specificity. When possible with the information provided, PPV was calculated or evaluated.

Results: Additive queries on PubMed retrieved 928 candidate papers. Of these, 376 were eligible based on abstract. Finally, 26 studies qualified for a full-text review, and 17 provided enough information for the study objectives. An additional 4 papers were added from the references of the reviewed papers. The results demonstrate massive variations in PPVs ranging from 8% to 83% according to the object of the decision support, with most results between 20% and 40%. The best results were observed when patients' characteristics, such as comorbidity or laboratory test results, were taken into account. There was also an important variation in sensitivity, ranging from 38% to 91%.

Conclusions: There is increasing reporting of alerts override in CPOE decision support. Several causes are discussed in the literature, the most important one being the clinical relevance of alerts. In this paper, we tried to assess formally the clinical relevance of alerts, using a near-strong proxy, which is the PPV of alerts, or any way to express it such as the rate of true and false positive alerts. In doing this literature review, three inferences were drawn. First, very few papers report direct or enough indirect elements that support the use or the computation of PPV, which is a gold standard for all diagnostic tools in medicine and should be systematically reported for decision support. Second, the PPV varies a lot according to the typology of decision support, so that overall rates are not useful, but must be reported by the type of alert. Finally, in general, the PPVs are below or near 50%, which can be considered as very low.

(JMIR Med Inform 2018;6(1):e3) doi:10.2196/medinform.7170

KEYWORDS

decision support systems, clinical; medical order entry systems; system, medication alert; sensitivity; specificity; predictive value of tests

Introduction

Computerized patient records and computerized provider order entry (CPOE) systems are recognized as major tools in efforts to improve the safety and efficiency of care. Computerized patient records are the cornerstone of information sharing among care providers, and increasingly with patients; they contribute to improving the continuum of care and patient safety. The way CPOE improves processes rests on 3 pillars. The first pillar is formal structured order entry, which improves both completeness and readability. The second embeds CPOE into complete care processes such as medication loops or clinical pathways. The third pillar is the decision support capability during the ordering process, such as the provision of extensive information on the drugs being prescribed or the links made between the current order and other elements of the patient's record such as problems, laboratory results, and other drugs or diagnoses. Numerous studies have reported the positive effects of clinical decision support systems (CDSS) on patient outcomes such as fewer duplicate orders, dosage errors, drug interactions, and missed or delayed actions using reminders, to name a few [1-4]. The benefits of CPOE have already been demonstrated in the improved cost-efficiency of care, either directly, by lowering adverse events and duplicate orders, or indirectly, by reducing lengths of stay [5,6]. Nevertheless, the burden of alerts and reminders must not be too high or *alert fatigue* could cause clinicians to override both important and unimportant alerts, thus jeopardizing the improvements in safety that a CDSS should be expected to bring [7]. In other words, the CDSS's specificity (Sp) must be high. A few studies have reported on the unintended effects of CDSS in CPOE [8-10] and their occasional dramatic consequences on patient safety. These were related to delays in reporting adverse events, and thus therapy, leading to specific infectious or thrombotic complications in treatment [11] or to the cancellation of QT interval-alert generation after proposed measures to reduce alert overload [12]. This is not a marginal problem. For example, a 2013 study published by Yeh, analyzing more than 1 million prescriptions from outpatient settings in Taiwan, reported a 91.5% override rate on the approximately 11,000 drug-drug interaction alerts proposed [13]. Understanding the reasons why clinicians override CDSS in CPOE has since received a lot of attention [14,15]. In recent years, numerous studies have been published on the topic of alert improvements for CPOE. These addressed the theoretical background, such as models and frameworks [16], data representation [17] or behavioral theories [18], usability and interfaces [19,20], perceptions and expectations [21], simulation [22], effectiveness monitoring [23,24], and decision support Sp [25], among other issues.

This study focuses on the predictive value of CPOE alerts. One can consider the CDSS in CPOE to be akin to any other decision support instrument in medicine: a tool with positive predictive values (PPVs) and negative predictive values (NPVs). As mentioned above, some previous studies have focused on evaluating the predictive value of decision support in CPOE, and the PPVs reported were usually below 20% and as low as 5% [26,27]. A study by van der Sijs et al stated that 49% to 96% of alerts were overridden [28] and identified a range of human factors responsible:

- alert fatigue due to a poor signal-to-noise ratio as a result of a low PPV
- usability issues such as bad ergonomics, misinterpretation, or unnoticed alerts
- disagreements with guidelines
- physicians' belief in their own knowledge
- lack of time

Further understanding has been provided by questionnaires and focus groups that allowed physicians to evaluate the most important factors for useful, easy-to-use alerts [29,30]. These showed that drug-related alerts were rated more useful than alerts reminding the clinician of the state of the patient's health or disease. Shah et al suggested that an approach based on a careful selection of alerts so as to improve the relevancy, severity, likelihood, and strength of clinical evidence would improve the acceptance of alerts [31]. Bates et al put forward “Ten commandments for effective clinical decision support” such as speed of the information system, anticipation of clinician needs and provide information to clinicians at the time they need it, integration suggestions with practice, offer an alternative, change of direction rather than stop or management, and maintenance of knowledge-based systems [32].

As stated, most alerts are overridden. Although numerous authors speak about the number of alerts, or the pertinence of alerts, we have been interested in trying to assess clearly the PPV of alerts, and thus the rate of true and false positive alerts. In doing this review, three inferences were drawn. First, very few papers report direct or enough indirect elements that support the use or the computation of PPV, which is a gold standard for all diagnostic tools in medicine, which is why it should be systematically reported for decision support. Second, the PPV varies a lot according to the typology of decision support and would have to be reported by the type of alert. Third is that, in general, the PPVs can be considered as very low—below 50% or near 50%.

Due to the high expectations health care professionals have for CDSS in CPOE, as well as the related costs and potential unintended consequences, we decided to carry out a systematic review of the literature on CPOE, CDSS, and predictive value, and their associations with alert fatigue. We start from the assumption that a low PPV would explain why majority of alerts are overridden. We framed this systematic review to determine the real PPV of CPOE alerts.

Methods

Selection Criteria

We targeted publications evaluating clinically relevant alert in computerized patient records implementing CPOE.

Search Strategy

A search of the literature was made using PubMed for work published between February 2009 and March 2015, using the following queries: (CPOE[all fields] OR “Medical Order Entry Systems”[all fields] OR “Alert Systems”[all fields] OR “Order

Entry”[all fields] OR “Decision support Systems”[all fields]) AND (sensitivity[All Fields] OR sensibility[All Fields] OR predictive[All Fields]) OR (fatigue[All Fields] OR overload[All Fields] OR overcharge[All Fields] OR burden[All Fields] OR override[All Fields] OR overalerting[All Fields] OR ignore[All Fields]). The following meanings were searched for *decision support*: CPOE, medical order entry systems, alert systems, order entry, and decision support systems. The following meanings were searched for *relevance*: sensitivity, sensibility, predictive, fatigue, overload, overcharge, burden, override, over alerting, and ignore.

The following limits were applied to all queries: English language, only papers available in full text, assessing medication, and numerical data available.

We excluded qualitative studies, user-satisfaction or opinion surveys, physician adherence studies, and analyses of the impact of human factors.

Selection of Relevant Publications

First, the 3 reviewers (DC, GF, and CL) selected references independently based on their titles and according to the review study’s inclusion and exclusion criteria. When results were discordant, the final choice was made by consensus. Next, they independently read and assessed the abstracts of all the papers identified. When no abstract was available, full-text papers were retrieved and reviewed so that only relevant papers were retained. Again, the 3 reviewers solved any disagreements by consensus. In the absence of an agreement, the abstract was provisionally included for consideration subject to reading the full text.

Abstracts that were rated as relevant to the research question were kept, and all full-text papers were retrieved. Then, each retrieved paper’s reference section was searched for additional relevant literature that might be included.

Of the reviewers, 2 (DC and GF) assessed the quality of the papers selected by using a standardized evaluation process based on the exclusion and inclusion criteria. For papers to be selected for the final review, the levels of predictive value, sensitivity (Se; ability to generate alerts in potentially dangerous situations), or Sp (inability to prevent irrelevant alerts) were retrieved or calculated if possible. Se was defined as the number of patients with an adverse drug event (ADE) detected by an alert, out of the total number of patients with a positive ADE. Sp was defined as the number of patients without an ADE and with no warning alert, out of the total number of patients without an ADE. The PPV was defined as the number of relevant medication alerts (true positives) out of the total number of alerts (sum of true and false positives). Evaluation disagreements between the 2 reviewers were resolved by the third reviewer (CL).

Results

Selection of Studies

The database search retrieved 928 matching references. A first evaluation based on MEDLINE summary allowed identifying 402 potentially interesting papers. Then, a second deeper analysis based on abstracts and applying the inclusion and exclusion criteria resulted in the exclusion of 311 articles, thus reducing the initial set to 91 reports. Out of these, 26 full-text papers were retrieved, reviewed, and included in the next phase of the review. The additional search through the selected studies’ reference sections resulted in 20 additional potentially relevant papers. Of these, 4 were included in our analysis. The review selection process is summarized in Figure 1.

Description of Studies

Including the additional search references, the final sample of 17 studies that met our eligibility criteria, as listed in Table 1, were published between 1998 and 2015. The papers predominantly analyzed interruptive alerts (n=7/8 notified). Various alert targets were used and are described in Table 2. The main ones described were drug-lab interactions (n=11), drug-dosage interactions (n=8), drug-drug interactions (n=6), duplicate orders (n=3), and drug–allergy interactions (n=3).

These papers report the predictive value or Se and Sp of the alerts studied. As shown in Table 3, four papers did not report any PPV, although this study’s authors were able to calculate it for two of those papers. The PPV found in the papers were usually low and heterogeneous, mostly between 20% and 40%. Despite the diversity of target alerts, alert notifications, study designs, and study periods of the papers included in this review, it seems that PPVs were higher for drug-lab interactions (2.3%-83%) than they were for drug-dosage interactions (8%-13.8%), or drug-drug interactions (1.6%-48%). Furthermore, advanced CDSS [49] showed higher PPV than the more basic ones (17%-97%).

The Types of Alert Influencing PPV

In general, PPV increased when the risk increased. For example, PPV was higher for drug-dosage interactions than for drug-lab interactions. This is probably because of the higher risk of experiencing an ADE [48]. Furthermore, the PPV was lower in prevention (the opportunity to prevent ADEs) than in detection (evaluate or treat possible existing ADEs): 24% versus 97% [46]. Indeed PPV is related to the prevalence (Prev) unlike Se and Sp, which are only related to the test as defined as defined as follows: $PPV=(Se\times Prev)\div(Se\times Prev+(1-Sp)\times(1-Prev))$. Therefore, in prevention settings, the prevalence of disease is likely to be very low, so the PPV will also therefore be low. Additionally, it was shown that the PPV of alerts targeting drug-lab interactions varied with the choice of the alarm signal. Indeed, for a laboratory value lower than the maximum defined value, the PPV of the alert was 36% (95% CI 29-43). If an alert was activated after at least a 50% decrease in the value between the last two laboratory results, the PPV increased to 83% (95% CI 62-104). For two consecutive decreases, with at least a 25% difference between the third most recent and the most recent platelet count, the PPV was 40% (95% CI 32-48) [50].

Furthermore, it has been shown that the PPV of safety alerts aimed at high-risk patients was higher (PPV=14%) than when dealing with initiation of a drug (PPV=6%), ongoing use of a drug (PPV=6%), advice (PPV=7%), and medication used to treat an ADE (PPV=0%) [28]. In summary, the PPV of alerts is usually very low. However, several factors seem to improve PPV.

Contextual Information Improves PPV

The PPV of advanced alerts is higher than for basic alerts because they are more specific. Advanced CDSS, such as using patients' characteristics and laboratory test results, have a higher PPV than basic ones. For example, Eppenga et al showed that using information from the laboratory and a few other specific pieces of information increased the PPV from 12.2% to 23.3% (*P*<.05) and that PPV was higher in advanced systems than in basic ones (17% vs 5.8%, *P*<.05) [37]. Numerous factors can influence the PPV, mostly because they will have influence of the population considered for the alert. For example, not specifying the administration route can sometimes decrease the PPV, for example in some topical treatments. This is because the risk of developing an ADE can vary according to the administration route [50]. Further advances in dosing alert systems should aim to improve the Se of alerts. The Se of the system for identifying dosing errors increased from 54.1% (95% CI 47.8-60.3) to 60.3% (95% CI 54.0-66.3) in a customized dose range system (*P*=.02). The system's Se for underdosage was 49.6% without customization, and this increased to 60.3% with customization (*P*=.01) [47]. Furthermore, it has been highlighted that PPV differs according to patients' characteristics and comorbidity: for alerts on the risk of developing hypoglycemia, the PPV was higher for patients with sulfonylureas in their drug regimens (45.7% vs 28.4%, *P*=.04) and for patients with three or more chronic medical conditions (35.7% vs 22.7%, *P*=.049). The PPV of an alert warning of the risk of developing hyperkalemia was higher for patients with serum creatinine >2.0 mg/dL (50.0% vs 16.0%, *P*=.01) [38].

The PPV can vary according to the types of alerts. Among the 5 types of alerts with the best PPV (34.1%-73.3%), 3 were drug-lab interactions, which are advanced alerts. In parallel, of the 10 alerts described as being the least relevant (PPV between 0% and 4.5%), 8 were drug-drug interactions [37].

Finally, the PPV varies according to the specific goal. A study of alerts aimed at identifying 4 types of ADE showed that some of them could have a lower PPV: the PPV was only 4.0% (95% CI 1.3-9.1) for hypokalemia versus 31.2% (95% CI 18.2-46.6) for hypoglycemia, 31.1% (95% CI 25.1-37.8) for hyperkalemia, and 20.6% (95% CI 11.7-32.1) for thrombocytopenia. Furthermore, the effect of an alert can differ according to the medical specialty. In a study by Riggio et al, a surgery department ordered laboratory tests earlier than general medicine department when alerts were activated, probably because surgeons were more aware of the importance of the platelet counts that were being observed in the study [33]. The PPV can also vary according to the alert's pharmacological target. For example, anti-infective drugs are excluded from alerts concerning drug dosage interactions to limit the number of false positives because these drugs could present patients specific dosing adjustment and multiple indications [44].

Figure 1. Flowchart describing the systematic literature review process.

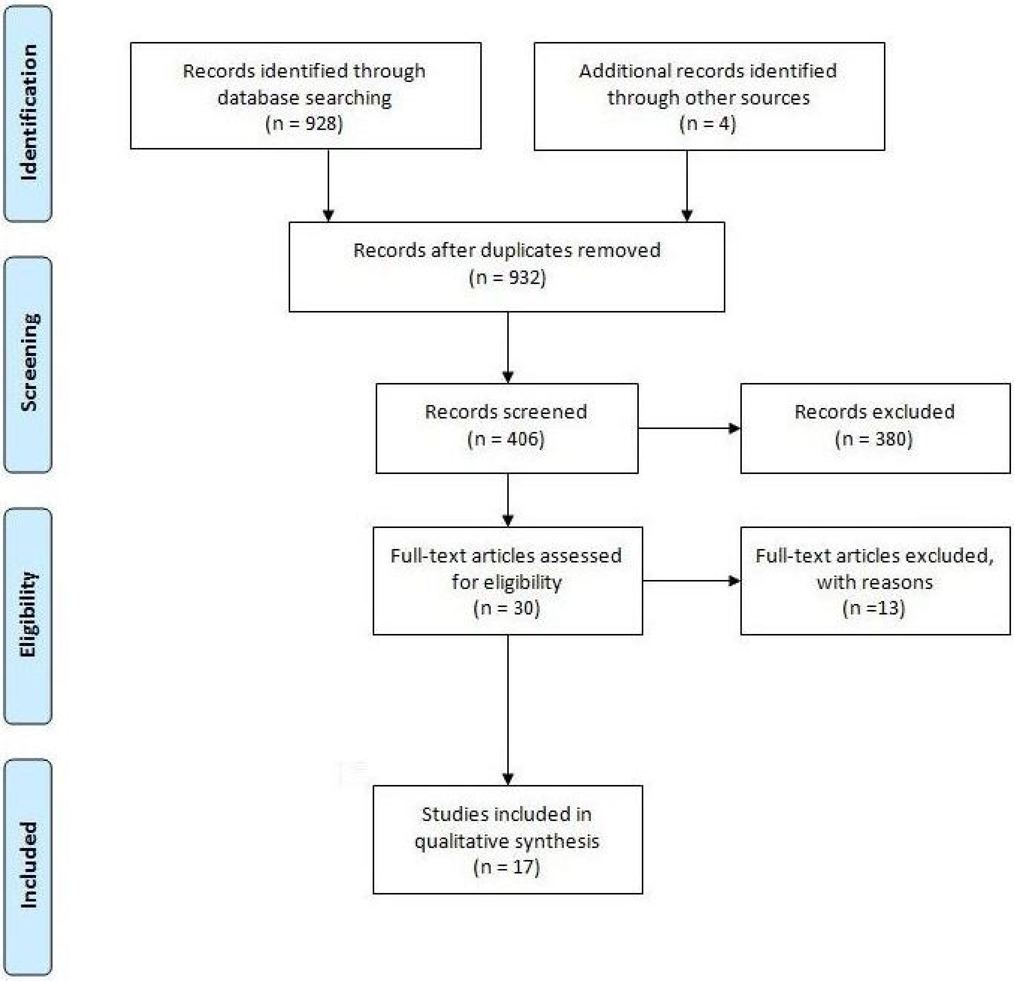

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 928)
Additional records identified through other sources (n = 4)
Records after duplicates removed (n = 932)
Records screened (n = 406)
Records excluded (n = 380)
Full-text articles assessed for eligibility (n = 30)
Full-text articles excluded, with reasons (n =13)
Studies included in qualitative synthesis (n = 17)

Table 1. Characteristics of the studies included in the paper.

Source	Study design	Study period	Study site	Specialty	Patient care	Number of patients with an alert
Riggio et al, 2008 [33]	Control and intervention	3 weeks	728-bed hospital	Medicine, surgery, pediatric	Inpatient	Control group: 47; intervention group: 53
Cash, 2009 [34]	Retrospective analysis	N/A[a]	Hospital	Pediatric	Inpatient	N/A
Van der Sijs et al, 2010 [35]	Control/intervention	1 month	807-bed hospital	N/A	Inpatient	N/A
FitzHenry et al, 2011 [36]	Retrospective analysis	7 months	807-bed hospital	N/A	Inpatient	2404
Eppenga et al, 2014 [37]	Cross-sectional	5 days	800-bed hospital	N/A	Inpatient	619
Moore et al, 2009 [38]	Prospective observational	5 months	684-bed hospital	N/A	N/A	456
Fritz et al, 2012 [39]	Prospective observational	N/A	850-bed hospital	Internal medicine	Inpatient	100
Harinstein et al, 2012 [40]	Prospective observational	8 weeks	Medical center	Medical and cardiac intensive care	Inpatient	64
Zorina et al, 2012 [41]	Cross-sectional	1 year	850-bed hospital	Neurological	Inpatient	484
Beeler et al, 2013 [42]	Retrospective analysis	90 weeks	850-bed hospital	N/A	Inpatient	922
Rommers et al, 2013 [27]	Prospective observational	5 months	Hospital	Internal medicine, cardiology, lung, gastrointestinal, hematology	Inpatient	931
Stultz et al, 2013 [43]	Retrospective analysis	1 month	350-bed hospital	Pediatric	Inpatient	573
Stultz et al, 2014 [44]	Retrospective analysis	1 month	350-bed hospital	Pediatric	Inpatient	189
Dormann et al, 2004 [45]	Prospective study/retrospective analysis	6 months	Hospital	Gastroenterological	N/A	377
Raschke et al, 1998 [46]	Prospective case	6 months	650-bed hospital	Nonobstetrics	N/A	9306
Silverman et al, 2004 [47]	Retrospective analysis	3 one-year periods	726-bed teaching institution	Tertiary care	N/A	N/A
Handler et al, 2007 [48]	Systematic review	12 studies	Hospital	N/A	N/A	N/A

[a] N/A: not applicable.

Table 2. Characteristics of alerts included in the paper.

Source	Alert notification	Alert origin	Alert target
Riggio et al, 2008	Interruptive alert	—	• Drug-lab interaction: heparin-induced thrombocytopenia
Cash, 2009	Interruptive alert	—	• Drug-drug interaction • Drug-lab interaction • Duplicate order • Drug-dosage interaction • Drug-allergy interaction
Van der Sijs et al, 2010	Interruptive alert	Commercial system	• Drug-dosage interaction: overdosage • Drug-drug interaction • Drug–dosage interaction • Drug-allergy interaction • Drug-pregnancy interaction: contraindication • Duplicate order • Drug-lab interaction: bad renal function • Drug-pharmacogenetic interaction: poor metabolizer
FitzHenry et al, 2011	Interruptive alert	—	• Drug-dosage interaction: warfarin
Eppenga et al, 2014	Interruptive alert	—	Basic • Drug-drug interaction • Duplicate order Advanced • Drug-drug interaction • Drug-dosage interaction • Drug-lab interaction • Drug-lab interaction: missing laboratory value • Drug-disease interaction • Drug-age interaction
Moore et al, 2009	—	—	• Drug-lab interaction developing adverse drug event (ADE): hypoglycemia, hypokalemia, hyperkalemia, and thrombocytopenia
Fritz et al, 2012	—	Commercial system	• Drug-drug interaction
Harinstein et al, 2012	—	Commercial system	• Drug-lab interaction: drug-induced thrombocytopenia
Zorina et al, 2012	—	Commercial system	• Drug-drug interaction
Beeler et al, 2013	Noninterruptive alert	—	• Drug-drug interaction
Rommers et al, 2013	—	—	• Drug-lab interaction: ADE system
Stultz et al, 2013	Interruptive alert	—	• Drug-dosage interaction
Stultz et al, 2014	Interruptive alert	—	• Drug-dosage interaction
Dormann et al, 2003	—	—	• Drug-lab interaction: predicted ADE
Raschke et al, 1998	—	—	• Drug-monitoring interaction: predicted ADE • Drug-age interaction: predicted ADE • Drug-lab interaction: predicted ADE
Silverman et al, 2004	—	—	• ADE detection system • Drug-allergy interaction • Drug-drug interaction • Therapeutic duplication • Drug-dosage interaction • Drug-lab interaction

Source	Alert notification	Alert origin	Alert target
Handler et al, 2007	—	—	• Antidote • Drug-lab interaction • Drug-dosage interaction: subtherapeutic medication levels

Table 3. Positive predictive value (PPV), sensitivity or specificty for studies included in the review.

Source	Number of alerts	Positive predictive value (%)	Sensitivity (%)	Specificity (%)	False positive (%)
Riggio et al, 2008	41,922	2.3	87	87	N/A[a]
Cash, 2009	—	1.4	N/A	N/A	N/A
Van der Sijs et al, 2010[a]	—	—	38-79 (n=29)	11-89 (n=19)	N/A
FitzHenry et al, 2011[b]	2308	—	N/A	N/A	46-85
Eppenga et al, 2014	Basic 2607/advanced 2256	Basic: 5.8 (n=150/2607)/advanced: 17 (*P*<.05)	N/A	N/A	N/A
Moore et al, 2009	611	4.0 (n=125)-31.2 (n=218)	N/A	N/A	N/A
Fritz et al, 2012	743	5.7 (n=3/53)-8 (n=29/362)	9.1(n=3/53) -87.9 (n=29/362)	N/A	N/A
Harinstein et al, 2012	350 (204/12/134)	36 (n=73/204)-83(n=10/12)	N/A	N/A	N/A
Zorina et al, 2012	1759/1082[c]	24/48[c]	70.6/72.4[c]	N/A	N/A
Beeler et al, 2013[d]	7902	1.6 (n=47/2866) (*P*=.002)	N/A	N/A	N/A
Rommers et al, 2013	2650 (963/722/437)	8 (n=204/2650)	N/A	N/A	N/A
Stultz et al, 2013[e]	3774	13.8	N/A	N/A	N/A
Stultz et al, 2014	257	Odds ratio (OR) 8 (95% CI 6.8-9.3)	OR 60.3 (95% CI 54.0-66.3) *P*=.02	OR 96.2 (95% CI: 96.0-96.3)	N/A
Dormann et al, 2003	2328 (1748/580)	Prospective study 25(n=574/2328)(13-40)/retrospective analysis 32(18-67)	91/40	23/76	N/A
Raschke et al, 1998	1116 (803/313)	24 (n=5/21)-97(n=190/196)	N/A	N/A	N/A
Silverman et al, 2004	3117/7390/6136	0-60	N/A	N/A	N/A
Handler et al, 2007	—	Antidotes: 9-11	N/A	N/A	N/A
	—	Laboratory test result: 3-27	N/A	N/A	N/A
		Supratherapeutic medication levels: 3-50	N/A	N/A	N/A

[a] N/A: not applicable.

[c] No PPV available.

[c] Values for two different programs of clinical decision support systems.

[d] Positive predictive value calculated for the review: PPV was defined as the quotient of the number of advice/interventions to prevent a possible adverse drug event and the total number of alerts generated.

[e] PPV calculated for the review: PPV was defined as number of correct alerts in comparison with Lexicomp.

Discussion

Principal Findings

The PPV found in the papers were rather low: 20% to 40%. Despite the heterogeneity of papers, it seems that several factors influence PPV. First, the PPV can vary with the types of alert such as the risk patients trying to be prevented. Furthermore, several factors seem to improve PPV such as contextual information. Indeed alerts that are more specific have a higher PPV than basic alerts specifying the administration route or patients' characteristics for example. Moreover, PPV can differ according to alert's pharmacological target or medical specialty.

Even the most basic systems usually show good Se. They thereby allow medical professionals to detect drug-related problems more comprehensively: a pharmacy department increased the number of its clinical interventions by 15% after the introduction of a CDSS [47]. However, the impact of a true positive alert can be paradoxical. For example, patients presented no reduction in ADEs, time to therapeutic intervention, or time to laboratory testing in an alert group, and physicians waited 1.6 days longer before stopping a treatment inducing ADE in that alert group (*P*=.049) [49]. This result could be because of alert fatigue induced by a low PPV.

This study has several limitations. First of all, we conducted our research using only PubMed, and carried no queries using EMBASE, Web of Science, or conference proceedings. The results are based on few reports, as only few studies reported all characteristics required to assess properly the contexts of decision support and their associated predictive values. There was a wide heterogeneity in how results were reported, completeness, and evaluation methodologies, thus limiting the reliability of pooling the PPV of alerts across publications. Because PPV varies with prevalence, the patient context, including population, hospital settings, and the like, has influence, and could not be considered. Thus, these results introduce some types of biases into the overall assessment.

Studies about interruptive alerts had some homogeneity in their methodology, and studies on decision support were mostly about 3 types interactions: drug-lab, drug-drug, and drug-dosage. These 3 types of interactions were the easiest to implement, and there are several large databases available for each of them. In general, systems that do not take patients' specific clinical information into account and use only external databases demonstrate the lowest PPV; systems that have a specific source of knowledge and use the greatest number of patients' individual characteristics have the highest PPV.

Conclusions

The PPV of clinical decision support systems for CPOE, as reported in the literature, varies massively, from 5.8% to 83%, with the majority of results between 20% and 40%. Drug-drug interaction alerts have the lowest PPV, and drug-lab alerts have the highest.

Our literature review leads us to suggest that the best strategy to use with a CPOE is to adapt and carefully optimize the database driving the knowledge for activating alerts. Furthermore, the CDSS should take into account as many of the patient's characteristics as possible. The efficiency of the alerts, and thus their PPV, is more important than a very large database of knowledge that may generate lots of false positives, which reduce PPV and generate alert fatigue.

Advanced alert systems should aim to improve PPV of alerts, while keeping a good Se. To reduce the number of false positive alerts, contextual data from different sources, such as the pharmacy, demographic data, or laboratory tests, should be integrated into the system.

The US Institute of Medicine has suggested that systems should be designed so as to make it "hard for people to do the wrong thing and easy for people to do the right thing" [51]. However, with PPVs as low as those seen in the literature, it seems, unfortunately, that many computerized patient records tend to make it hard for people to do the right thing and easy for people to do the wrong thing.

Conflicts of Interest

None declared.

References

1. Abramson EL, Kaushal R. Computerized provider order entry and patient safety. Pediatr Clin North Am 2012 Dec;59(6):1247-1255. [doi: 10.1016/j.pcl.2012.08.001] [Medline: 23116522]
2. Bates DW, Gawande AA. Improving safety with information technology. N Engl J Med 2003 Jun 19;348(25):2526-2534. [doi: 10.1056/NEJMsa020847] [Medline: 12815139]
3. Radley DC, Wasserman MR, Olsho LE, Shoemaker SJ, Spranca MD, Bradshaw B. Reduction in medication errors in hospitals due to adoption of computerized provider order entry systems. J Am Med Inform Assoc 2013 May 01;20(3):470-476 [FREE Full text] [doi: 10.1136/amiajnl-2012-001241] [Medline: 23425440]
4. van Rosse F, Maat B, Rademaker CM, van Vught AJ, Egberts AC, Bollen CW. The effect of computerized physician order entry on medication prescription errors and clinical outcome in pediatric and intensive care: a systematic review. Pediatrics 2009 Apr;123(4):1184-1190. [doi: 10.1542/peds.2008-1494] [Medline: 19336379]
5. Forrester SH, Hepp Z, Roth JA, Wirtz HS, Devine EB. Cost-effectiveness of a computerized provider order entry system in improving medication safety ambulatory care. Value Health 2014 Jun;17(4):340-349 [FREE Full text] [doi: 10.1016/j.jval.2014.01.009] [Medline: 24968993]
6. Vermeulen KM, van Doormaal JE, Zaal RJ, Mol PG, Lenderink AW, Haaijer-Ruskamp FM, et al. Cost-effectiveness of an electronic medication ordering system (CPOE/CDSS) in hospitalized patients. Int J Med Inform 2014 Aug;83(8):572-580. [doi: 10.1016/j.ijmedinf.2014.05.003] [Medline: 24929633]
7. Slight SP, Seger DL, Nanji KC, Cho I, Maniam N, Dykes PC, et al. Are we heeding the warning signs? Examining providers' overrides of computerized drug-drug interaction alerts in primary care. PLoS One 2013;8(12):e85071 [FREE Full text] [doi: 10.1371/journal.pone.0085071] [Medline: 24386447]
8. Magid S, Forrer C, Shaha S. Duplicate orders: an unintended consequence of computerized provider/physician order entry (CPOE) implementation: analysis and mitigation strategies. Appl Clin Inform 2012;3(4):377-391 [FREE Full text] [doi: 10.4338/ACI-2012-01-RA-0002] [Medline: 23646085]
9. Ash JS, Sittig DF, Campbell EM, Guappone KP, Dykstra RH. Some unintended consequences of clinical decision support systems. AMIA Annu Symp Proc 2007 Oct 11:26-30 [FREE Full text] [Medline: 18693791]

10. Cowan L. Literature review and risk mitigation strategy for unintended consequences of computerized physician order entry. Nurs Econ 2013;31(1):27-31, 11. [Medline: 23505740]
11. Strom BL, Schinnar R, Aberra F, Bilker W, Hennessy S, Leonard CE, et al. Unintended effects of a computerized physician order entry nearly hard-stop alert to prevent a drug interaction: a randomized controlled trial. Arch Intern Med 2010 Sep 27;170(17):1578-1583. [doi: 10.1001/archinternmed.2010.324] [Medline: 20876410]
12. van der Sijs H, Kowlesar R, Aarts J, Berg M, Vulto A, van Gelder T. Unintended consequences of reducing QT-alert overload in a computerized physician order entry system. Eur J Clin Pharmacol 2009 Sep;65(9):919-925 [FREE Full text] [doi: 10.1007/s00228-009-0654-3] [Medline: 19415251]
13. Yeh ML, Chang YJ, Wang PY, Li YC, Hsu CY. Physicians' responses to computerized drug-drug interaction alerts for outpatients. Comput Methods Programs Biomed 2013 Jul;111(1):17-25. [doi: 10.1016/j.cmpb.2013.02.006] [Medline: 23608682]
14. Perna G. Clinical alerts that cried wolf. As clinical alerts pose physician workflow problems, healthcare IT leaders look for answers. Healthc Inform 2012 Apr;29(4):18, 20. [Medline: 22574398]
15. Coleman JJ, van der Sijs H, Haefeli WE, Slight SP, McDowell SE, Seidling HM, et al. On the alert: future priorities for alerts in clinical decision support for computerized physician order entry identified from a European workshop. BMC Med Inform Decis Mak 2013;13:111 [FREE Full text] [doi: 10.1186/1472-6947-13-111] [Medline: 24083548]
16. Wipfli R, Lovis C. Alerts in clinical information systems: building frameworks and prototypes. Stud Health Technol Inform 2010;155:163-169. [Medline: 20543324]
17. Luna D, Otero V, Canosa D, Montenegro S, Otero P, de Quirós FG. Analysis and redesign of a knowledge database for a drug-drug interactions alert system. Stud Health Technol Inform 2007;129(Pt 2):885-889. [Medline: 17911843]
18. Schedlbauer A, Prasad V, Mulvaney C, Phansalkar S, Stanton W, Bates DW, et al. What evidence supports the use of computerized alerts and prompts to improve clinicians' prescribing behavior? J Am Med Inform Assoc 2009;16(4):531-538 [FREE Full text] [doi: 10.1197/jamia.M2910] [Medline: 19390110]
19. Khajouei R, Jaspers MW. CPOE system design aspects and their qualitative effect on usability. Stud Health Technol Inform 2008;136:309-314. [Medline: 18487749]
20. Tsopra R, Jais JP, Venot A, Duclos C. Comparison of two kinds of interface, based on guided navigation or usability principles, for improving the adoption of computerized decision support systems: application to the prescription of antibiotics. J Am Med Inform Assoc 2014 Feb;21(e1):e107-e116 [FREE Full text] [doi: 10.1136/amiajnl-2013-002042] [Medline: 24008427]
21. Jung M, Hoerbst A, Hackl WO, Kirrane F, Borbolla D, Jaspers MW, et al. Attitude of physicians towards automatic alerting in computerized physician order entry systems. A comparative international survey. Methods Inf Med 2013;52(2):99-108. [doi: 10.3414/ME12-02-0007] [Medline: 23187311]
22. van der Sijs H, van Gelder T, Vulto A, Berg M, Aarts J. Understanding handling of drug safety alerts: a simulation study. Int J Med Inform 2010 May;79(5):361-369. [doi: 10.1016/j.ijmedinf.2010.01.008] [Medline: 20171929]
23. Reynolds G, Boyer D, Mackey K, Povondra L, Cummings A. Alerting strategies in computerized physician order entry: a novel use of a dashboard-style analytics tool in a children's hospital. AMIA Annu Symp Proc 2008 Nov 06:1108. [Medline: 18999063]
24. Zimmerman CR, Jackson A, Chaffee B, O'Reilly M. A dashboard model for monitoring alert effectiveness and bandwidth. AMIA Annu Symp Proc 2007 Oct 11:1176. [Medline: 18694272]
25. Smithburger PL, Buckley MS, Bejian S, Burenheide K, Kane-Gill SL. A critical evaluation of clinical decision support for the detection of drug-drug interactions. Expert Opin Drug Saf 2011 Nov;10(6):871-882. [doi: 10.1517/14740338.2011.583916] [Medline: 21542665]
26. McCoy AB, Thomas EJ, Krousel-Wood M, Sittig DF. Clinical decision support alert appropriateness: a review and proposal for improvement. Ochsner J 2014;14(2):195-202 [FREE Full text] [Medline: 24940129]
27. Rommers MK, Zwaveling J, Guchelaar H, Teepe-Twiss IM. Evaluation of rule effectiveness and positive predictive value of clinical rules in a Dutch clinical decision support system in daily hospital pharmacy practice. Artif Intell Med 2013 Sep;59(1):15-21. [doi: 10.1016/j.artmed.2013.04.001] [Medline: 23664455]
28. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc 2006;13(2):138-147 [FREE Full text] [doi: 10.1197/jamia.M1809] [Medline: 16357358]
29. Krall MA, Sittig DF. Subjective assessment of usefulness and appropriate presentation mode of alerts and reminders in the outpatient setting. Proc AMIA Symp 2001:334-338 [FREE Full text] [Medline: 11825206]
30. Krall MA, Sittig DF. Clinician's assessments of outpatient electronic medical record alert and reminder usability and usefulness requirements. Proc AMIA Symp 2002:400-404 [FREE Full text] [Medline: 12463855]
31. Shah NR, Seger AC, Seger DL, Fiskio JM, Kuperman GJ, Blumenfeld B, et al. Improving acceptance of computerized prescribing alerts in ambulatory care. J Am Med Inform Assoc 2006;13(1):5-11 [FREE Full text] [doi: 10.1197/jamia.M1868] [Medline: 16221941]
32. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc 2003;10(6):523-530 [FREE Full text] [doi: 10.1197/jamia.M1370] [Medline: 12925543]

33. Riggio JM, Cooper MK, Leiby BE, Walenga JM, Merli GJ, Gottlieb JE. Effectiveness of a clinical decision support system to identify heparin induced thrombocytopenia. J Thromb Thrombolysis 2009 Aug;28(2):124-131. [doi: 10.1007/s11239-008-0279-x] [Medline: 18839278]
34. Cash JJ. Alert fatigue. Am J Health Syst Pharm 2009 Dec 01;66(23):2098-2101. [doi: 10.2146/ajhp090181] [Medline: 19923309]
35. van der Sijs H, Bouamar R, van Gelder T, Aarts J, Berg M, Vulto A. Functionality test for drug safety alerting in computerized physician order entry systems. Int J Med Inform 2010 Apr;79(4):243-251. [doi: 10.1016/j.ijmedinf.2010.01.005] [Medline: 20149722]
36. FitzHenry F, Doran J, Lobo B, Sullivan TM, Potts A, Feldott CC, et al. Medication-error alerts for warfarin orders detected by a bar-code-assisted medication administration system. Am J Health Syst Pharm 2011 Mar 01;68(5):434-441. [doi: 10.2146/ajhp090666] [Medline: 21330686]
37. Eppenga WL, Derijks HJ, Conemans JM, Hermens WA, Wensing M, De Smet PA. Comparison of a basic and an advanced pharmacotherapy-related clinical decision support system in a hospital care setting in the Netherlands. J Am Med Inform Assoc 2012;19(1):66-71 [FREE Full text] [doi: 10.1136/amiajnl-2011-000360] [Medline: 21890873]
38. Moore C, Li J, Hung C, Downs J, Nebeker JR. Predictive value of alert triggers for identification of developing adverse drug events. J Patient Saf 2009 Dec;5(4):223-228. [doi: 10.1097/PTS.0b013e3181bc05e5] [Medline: 22130215]
39. Fritz D, Ceschi A, Curkovic I, Huber M, Egbring M, Kullak-Ublick GA, et al. Comparative evaluation of three clinical decision support systems: prospective screening for medication errors in 100 medical inpatients. Eur J Clin Pharmacol 2012 Aug;68(8):1209-1219. [doi: 10.1007/s00228-012-1241-6] [Medline: 22374346]
40. Harinstein LM, Kane-Gill SL, Smithburger PL, Culley CM, Reddy VK, Seybert AL. Use of an abnormal laboratory value-drug combination alert to detect drug-induced thrombocytopenia in critically Ill patients. J Crit Care 2012 Jun;27(3):242-249. [doi: 10.1016/j.jcrc.2012.02.014] [Medline: 22520497]
41. Zorina OI, Haueis P, Semmler A, Marti I, Gonzenbach RR, Guzek M, et al. Comparative evaluation of the drug interaction screening programs MediQ and ID PHARMA CHECK in neurological inpatients. Pharmacoepidemiol Drug Saf 2012 Aug;21(8):872-880. [doi: 10.1002/pds.3279] [Medline: 22517594]
42. Beeler PE, Eschmann E, Rosen C, Blaser J. Use of an on-demand drug-drug interaction checker by prescribers and consultants: a retrospective analysis in a Swiss teaching hospital. Drug Saf 2013 Jun;36(6):427-434. [doi: 10.1007/s40264-013-0022-1] [Medline: 23516005]
43. Stultz JS, Nahata MC. Appropriateness of commercially available and partially customized medication dosing alerts among pediatric patients. J Am Med Inform Assoc 2014 Feb;21(e1):e35-e42 [FREE Full text] [doi: 10.1136/amiajnl-2013-001725] [Medline: 23813540]
44. Stultz JS, Porter K, Nahata MC. Sensitivity and specificity of dosing alerts for dosing errors among hospitalized pediatric patients. J Am Med Inform Assoc 2014 Oct;21(e2):e219-e225 [FREE Full text] [doi: 10.1136/amiajnl-2013-002161] [Medline: 24496386]
45. Dormann H, Criegee-Rieck M, Neubert A, Egger T, Levy M, Hahn EG, et al. Implementation of a computer-assisted monitoring system for the detection of adverse drug reactions in gastroenterology. Aliment Pharmacol Ther 2004 Feb 01;19(3):303-309. [Medline: 14984377]
46. Raschke RA, Gollihare B, Wunderlich TA, Guidry JR, Leibowitz AI, Peirce JC, et al. A computer alert system to prevent injury from adverse drug events: development and evaluation in a community teaching hospital. J Am Med Assoc 1998 Oct 21;280(15):1317-1320. [Medline: 9794309]
47. Silverman JB, Stapinski CD, Huber C, Ghandi TK, Churchill WW. Computer-based system for preventing adverse drug events. Am J Health Syst Pharm 2004 Aug 01;61(15):1599-1603. [Medline: 15372836]
48. Handler SM, Altman RL, Perera S, Hanlon JT, Studenski SA, Bost JE, et al. A systematic review of the performance characteristics of clinical event monitor signals used to detect adverse drug events in the hospital setting. J Am Med Inform Assoc 2007;14(4):451-458 [FREE Full text] [doi: 10.1197/jamia.M2369] [Medline: 17460130]
49. Kuperman GJ, Bobb A, Payne TH, Avery AJ, Gandhi TK, Burns G, et al. Medication-related clinical decision support in computerized provider order entry systems: a review. J Am Med Inform Assoc 2007;14(1):29-40 [FREE Full text] [doi: 10.1197/jamia.M2170] [Medline: 17068355]
50. Harinstein LM, Kane-Gill SL, Smithburger PL, Culley CM, Reddy VK, Seybert AL. Use of an abnormal laboratory value-drug combination alert to detect drug-induced thrombocytopenia in critically Ill patients. J Crit Care 2012 Jun;27(3):242-249. [doi: 10.1016/j.jcrc.2012.02.014] [Medline: 22520497]
51. Institute of Medicine (US) Committee on Quality of Health Care in America. In: Kohn LT, Corrigan JM, Donaldson MS, editors. To Err Is Human: Building a Safer Health System. Washington (DC): National Academies Press (US); 2000.

Abbreviations

ADE: adverse drug event
CDSS: clinical decision support systems
CPOE: computerized provider order entry

PPV: positive predictive value
Se: Sensitivity
Sp: Specificity

Edited by G Eysenbach; submitted 04.04.17; peer-reviewed by A Elmessiry, K Singh; comments to author 05.06.17; revised version received 25.08.17; accepted 16.09.17; published 24.01.18.

Please cite as:
Carli D, Fahrni G, Bonnabry P, Lovis C
Quality of Decision Support in Computerized Provider Order Entry: Systematic Literature Review
JMIR Med Inform 2018;6(1):e3
URL: http://medinform.jmir.org/2018/1/e3/
doi:10.2196/medinform.7170
PMID:29367187

Original Paper

Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis

Brett K Beaulieu-Jones[1,2], PhD; Daniel R Lavage[3], BS; John W Snyder[3], NDTR, RDN; Jason H Moore[2], PhD; Sarah A Pendergrass[3], PhD; Christopher R Bauer[3], BA, PhD

[1]Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[2]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, United States
[3]Biomedical and Translational Informatics Institute, Geisinger, Danville, PA, United States

Corresponding Author:
Christopher R Bauer, BA, PhD
Biomedical and Translational Informatics Institute
Geisinger
100 N Academy Ave
Danville, PA,
United States
Phone: 1 603 369 7709
Email: cbauer@geisinger.edu

Abstract

Background: Missing data is a challenge for all studies; however, this is especially true for electronic health record (EHR)-based analyses. Failure to appropriately consider missing data can lead to biased results. While there has been extensive theoretical work on imputation, and many sophisticated methods are now available, it remains quite challenging for researchers to implement these methods appropriately. Here, we provide detailed procedures for when and how to conduct imputation of EHR laboratory results.

Objective: The objective of this study was to demonstrate how the mechanism of missingness can be assessed, evaluate the performance of a variety of imputation methods, and describe some of the most frequent problems that can be encountered.

Methods: We analyzed clinical laboratory measures from 602,366 patients in the EHR of Geisinger Health System in Pennsylvania, USA. Using these data, we constructed a representative set of complete cases and assessed the performance of 12 different imputation methods for missing data that was simulated based on 4 mechanisms of missingness (missing completely at random, missing not at random, missing at random, and real data modelling).

Results: Our results showed that several methods, including variations of Multivariate Imputation by Chained Equations (MICE) and softImpute, consistently imputed missing values with low error; however, only a subset of the MICE methods was suitable for multiple imputation.

Conclusions: The analyses we describe provide an outline of considerations for dealing with missing EHR data, steps that researchers can perform to characterize missingness within their own data, and an evaluation of methods that can be applied to impute clinical data. While the performance of methods may vary between datasets, the process we describe can be generalized to the majority of structured data types that exist in EHRs, and all of our methods and code are publicly available.

(JMIR Med Inform 2018;6(1):e11) doi:10.2196/medinform.8960

KEYWORDS

imputation; missing data; clinical laboratory test results; electronic health records

Introduction

Justification

Missing data present a challenge to researchers in many fields, and this challenge is growing as datasets increase in size and scope. This is especially problematic for electronic health records (EHRs), where missing values frequently outnumber observed values. EHRs were designed to record and improve patient care and streamline billing, and not as resources for research [1]; thus, there are significant challenges to using these data to gain a better understanding of human health. As EHR data become increasingly used as a source of phenotypic information for biomedical research [2], it is crucial to develop strategies for coping with missing data.

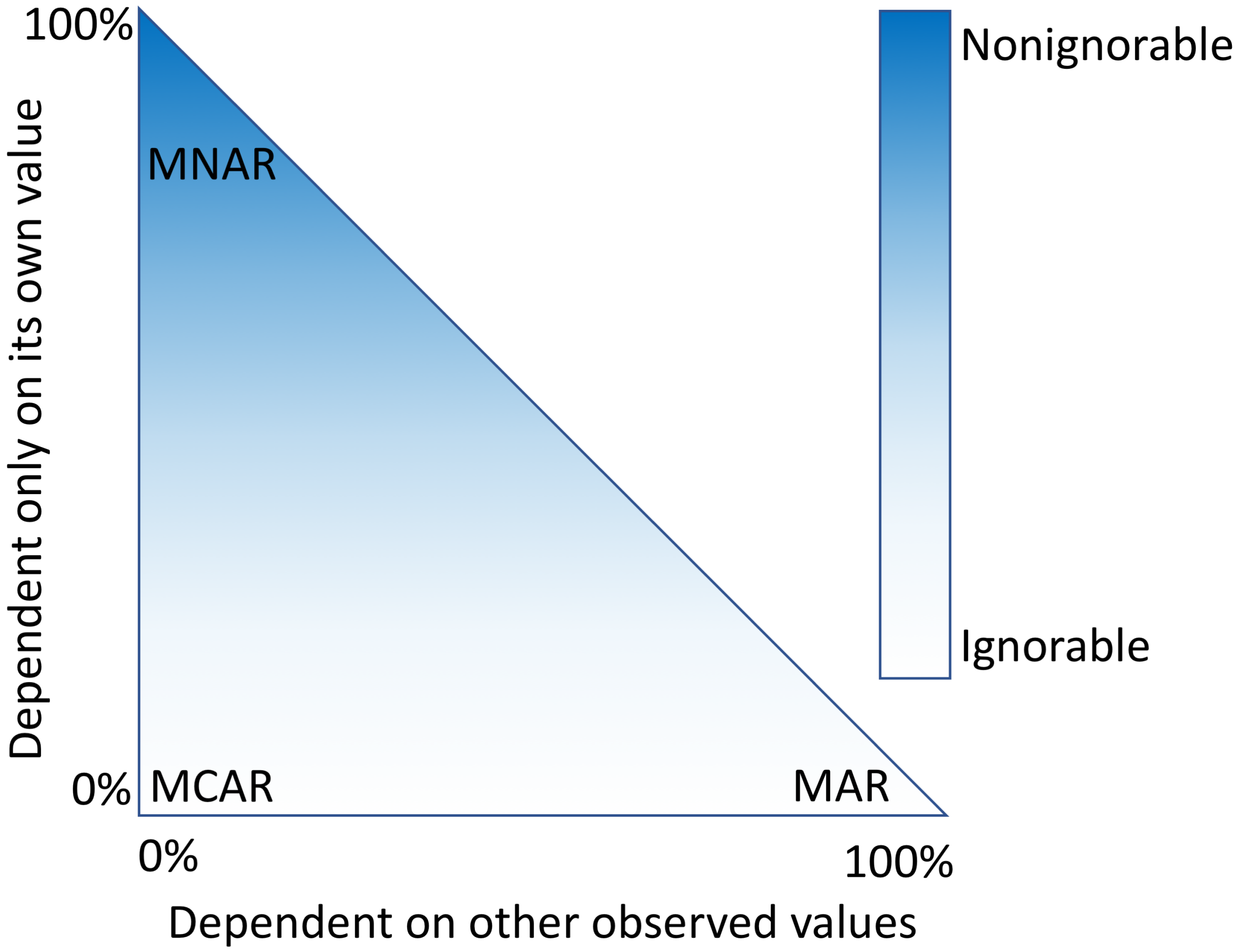
Clinical laboratory assay results are a particularly rich data source within the EHR, but they also tend to have large amounts of missing data. These data may be missing for many different reasons. Some tests are used for routine screening, but screening may be biased. Other tests are only conducted if they are clinically relevant to very specific ailments. Patients may also receive care at multiple health care systems, resulting in information gaps at each institution. Age, sex, socioeconomic status, access to care, and medical conditions can all affect how comprehensive the data are for a given patient. Accounting for the mechanisms that cause data to be missing is critical, since failure to do so can lead to biased conclusions.

Background

Aside from the uncertainty associated with a variable that is not observed, many analytical methods, such as regression or principal components analysis, are designed to operate only on a complete dataset. The easiest way to implement these procedures is to remove variables with missing values or remove individuals with missing values. Eliminating variables is justifiable in many situations, especially if a given variable has a large proportion of missing values, but doing so may restrict the scope and power of a study. Removing individuals with missing data is another option known as complete-case analysis. This is generally not recommended unless the fraction of individuals that will be removed is small enough to be considered trivial, or there is good reason to believe that the absence of a value is due to random chance. If there are systematic differences between individuals with and without observations, complete-case analysis will be biased.

An alternative approach is to fill in the fields that are missing data with estimates. This process, called imputation, requires a model that makes assumptions about why only some values were observed. Missingness mechanisms fall somewhere in a spectrum between 3 scenarios (Figure 1).

Figure 1. Two general paradigms are commonly used to describe missing data. Missing data are considered ignorable if the probability of observing a variable has no relation to the value of the observed variable and are considered nonignorable otherwise. The second paradigm divides missingness into 3 categories: missing completely at random (MCAR: the probability of observing a variable is not dependent on its value or other observed values), missing at random (MAR: the probability of observing a variable is not dependent on its own value after conditioning on other observed variables), and missing not at random (MNAR: the probability of observing a variable is dependent on its value, even after conditioning on other observed variables). The x-axis indicates the extent to which a given value being observed depends on other values of other observed variables. The y-axis indicates the extent to which a given value being observed depends on its own value.


100%
MNAR
Dependent only on its own value
0%
MCAR
MAR
0%
100%
Dependent on other observed values
Nonignorable
Ignorable

When data are missing in a manner completely unrelated to both the observed and unobserved values, they are considered to be missing completely at random (MCAR) [3,4]. When data are MCAR, the observed data represent a random sample of the population, but this is rarely encountered in practice. Conversely, data missing not at random (MNAR) refers to a situation where the probability of observing a data point depends on the value of that data point [5]. In this case, the mechanism responsible for the missing data is biased and should not be considered ignorable [6]. For example, rheumatoid factor is an antibody detectable in blood, and the concentration of this antibody is correlated with the presence and severity of rheumatoid arthritis. This test is typically performed only for patients with some indication of rheumatoid arthritis. Thus, patients with high rheumatoid factor levels are more likely to have rheumatoid factor measures.

A more complicated scenario can arise when multiple variables are available. If the probability of observing a data point does not depend on the value of that data point, after conditioning on 1 or more additional variables, then that data point is said to be missing at random (MAR) [5]. For example, a variable, *X*, may be MNAR if considered in isolation. However, if we observe another variable, *Y*, that explains some of the variation in *X* such that, after conditioning on *Y*, the probability of observing *X* is no longer related to its own value, then *X* is said to be MAR. In this way, *Y* can transform *X* from MNAR to MAR (Figure 1). We cannot prove that *X* is randomly sampled unless we measure some of the unobserved values, but strong correlations, the ability to explain missingness, and domain knowledge may provide evidence that the data are MAR.

Imputation methods assume specific mechanisms of missingness, and assumption violations can lead to bias in the results of downstream analyses that can be difficult to predict [7,8]. Variances of imputed values are often underestimated, causing artificially low *P* values [9]. Additionally, for data MNAR, the observed values have a different distribution from the missing values. To cope with this, a model can be specified to represent the missing data mechanism, but such models can be difficult to evaluate and may have a large impact on results. Great caution should be taken when handling missing data, particularly data that are MNAR. Most imputation methods assume that data are MAR or MCAR, but it is worth reiterating that these are all idealized states, and real data invariably fall somewhere in between (Figure 1).

Objective

We aimed to provide a framework for characterizing and understanding the types of missing data present in the EHR. We also developed an open source framework that other researchers can follow when dealing with missing data.

Methods

Source Code

We provide the source code to reproduce this work in our repository on GitHub (GitHub, Inc) [10] under a permissive open source license. In addition, we used continuous analysis [11] to generate Docker Hub (Docker Inc) images matching the environment of the original analysis and to create intermediate results and logs. These artifacts are freely available [12].

Electronic Health Record Data Processing

All laboratory assays were mapped to Logical Observation Identifiers Names and Codes (LOINC). We restricted our analysis to outpatient laboratory results to minimize the effects of extreme results from inpatient and emergency department data. We used all laboratory results dated between August 8, 1996 and March 3, 2016, excluding codes for which less than 0.5% of patients had a result. The resulting dataset consisted of 669,212 individuals and 143 laboratory assays.

We removed any laboratory results that were obtained prior to the patient's 18th birthday or after their 90th. In cases where a date of death was present, we also removed laboratory results that were obtained within 1 year of death, as we found that the frequency of observations often spiked during this period and the values for certain laboratory tests were altered for patients near death. For each patient, a median date of observation was calculated based on their remaining laboratory results. We defined a temporal window of observation by removing any laboratory results recorded more than 5 years from the median date. We then calculated the median result of the remaining laboratory tests for each patient. As each variable had a different scale and many deviated from normality, we applied Box-Cox and Z-transformations to all variables. The final dataset used for all downstream analyses contained 602,366 patients and 146 variables (age, sex, body mass index [BMI], and 143 laboratory measures).

Variable Selection

We first ranked the laboratory measures by total amount of missingness, lowest to highest. At each rank, we calculated the percentage of complete cases for the set, including all lower-ranked measures. We also built a random forest classifier to predict the presence or absence of each variable. Based on these results and domain knowledge, we selected 28 variables that provided a reasonable trade-off between quantity and completeness and that we deemed to be largely MAR.

Predicting the Presence of Data

For each clinical laboratory measure, we used the scikit-learn [13] random forest classifier, to predict whether each value would be present. Each laboratory measure was converted to a binary label vector based on whether the measure was recorded. The values of all other laboratory measures, excluding comembers of a panel, were used as the training matrix input to the random forest. This process was repeated for each laboratory test using 10-fold cross-validation. We assessed prediction accuracy by the area under the receiver operating characteristic curve (AUROC) using the trapezoidal rule.

Sampling of Complete Cases

To generate a set of complete cases that resembled the whole population, we randomly sampled 100,000 patients without replacement. We then matched each of these individuals to the most similar patient who had a value for each of the 28 most common laboratory tests by matching sex and finding the minimal euclidean distance of age and BMI.

Simulation of Missing Data

Within the sampled complete cases, we selected the data for removal by 4 mechanisms

Simulation 1: Missing Completely at Random

We replaced values with NaN (indicator of missing data) at random. We repeated this procedure 10 times each for 10%, 20%, 30%, 40%, and 50% missingness, yielding 50 simulated datasets.

Simulation 2: Missing at Random

We selected 2 columns (*A* and *B*) and a quartile. For the values from column *A* within the quartile, we randomly replaced 50% of the values from column *B* with NaN. We repeated the procedure for each quartile and each laboratory test combination, yielding 3024 simulated datasets.

Simulation 3: Missing Not at Random

We selected a column and a quartile. When the column's value was in the quartile, we replaced it with NaN 50% of the time. We repeated this procedure for each of the 4 quartiles of each of the 28 laboratory values, generating a total 112 total simulated datasets.

Simulation 4: Missingness Based on Real Data Observations

From our complete-cases dataset, we matched each patient to the nearest neighbor, excluding self-matches, in the entire population based on their sex, age, and BMI. We then replaced any laboratory value in the complete cases with NaN if it was absent in the matched patient.

Imputation of Missing Data

Using our simulated datasets (simulations 1-4), we compared 18 common imputation methods (12 representative methods are shown in the figures below) from the fancyimpute [14] and the Multivariate Imputation by Chained Equations (MICE v2.30) [15] packages. Multimedia Appendix 1 (table) shows a full list of imputation methods and the parameters used for each.

Results

Our first step was to select a subset of the 143 laboratory measures for which imputation would be a reasonable approach. We began by ranking the clinical laboratory measures in descending order by the number of patients who had an observed value for that test. For each ranked laboratory test, we plotted the percentage of individuals missing a value, as well as the percentage of complete cases when that given test was joined with all the tests with lower ranks (ie, less missingness). These plots showed that the best trade-off between quantity of data and completeness was between 20 and 30 variables (Figure 2, part A). Beyond the 30 most common laboratory tests, the number of complete cases rapidly approached zero.

As age, sex, and BMI have a considerable impact on what clinical laboratory measures are collected, we evaluated the relationship between missingness and these covariates (Figure 2, parts B-D). We also used a random forest approach to predict the presence or absence of each measure based on the values of the other observed measures. MCAR data are not predictable, resulting in AUROCs near 0.5. Only 38 of the 143 laboratory tests had AUROCs less than 0.55 (Figure 2, part E). Very high AUROCs are most consistent with data that are MAR. For the top 30 candidate clinical laboratory measures based on the number of complete cases, the mean AUROC was 0.82. This suggested that the observed data could explain much of the mechanism responsible for the missing data within this set. We ultimately decided not to include the 29th-ranked laboratory test, specific gravity of urine (2965-2), since it had an AUROC of only 0.69 and is typically used for screening only within urology or nephrology departments (RV Levy, MD, personal conversation, June 2017). We included the lipid measures (ranks 25-28), since they had AUROC values near 0.82 and they are recommended for screening of patients depending on age, sex, and BMI [16]. Our data confirmed that age, sex, and BMI all predicted the presence of lipid measures (Multimedia Appendix 1, fig 1A-B).

To assess the accuracy of imputation methods, we required known values to compare with imputed values. Thus, we restricted our analysis to a subset of patients who were complete cases for the 28 selected variables (Table 1) [17]. Since the characteristics of this subset differed from those of the broader population (Figure 2, parts B-D), we used sampling and k-nearest neighbors (KNN) matching to generate a subset of the complete cases that better resembled the overall population. We then simulated missing data within this set by 4 mechanisms: MCAR, MAR, MNAR, and realistic patterns based on the original data.

We next evaluated our ability to predict the presence of each value in the simulated datasets. These simulations confirmed that our MCAR simulation had a low AUROC (Figure 3, part A). The MAR data (Figure 3, part B) and MNAR data (Figure 3, part C) were often well predicted, particularly for the MAR data and when data were missing from the tails of distributions. The AUROCs rarely exceeded 0.75 in the MNAR simulations, while values above 0.75 were typical in the MAR simulations. This provided additional support for our decision to restrict our focus to the top 28 laboratory measures, since they all had AUROCs between 0.9 and 0.75, which was outside the range of MNAR simulations (Figure 2, part F and Figure 3, part C).

We chose to test the accuracy of imputation for several methods from 2 popular and freely available libraries: the MICE package for R and the fancyimpute library for Python. We first applied each of these methods across simulations 1 to 3. For each combination, Figure 4 depicts the overall root mean square errors. Multimedia Appendix 1 (Supplemental Table and Figures 3-21) shows a breakdown of all the methods and parameters.

Figure 2. Summary of missing data across 143 clinical laboratory measures. (A) After ranking the clinical laboratory measures by the number of total results, the percentage of patients missing a result for each test was plotted (red points). At each rank, the percentage of complete cases for all tests of equal or lower rank were also plotted (blue points). Only variables with a rank ≤75 are shown. The vertical bar indicates the 28 tests that were selected for further analysis. (B) The full distribution of patient median ages is shown in blue, and the fraction of individuals in each age group that had a complete set of observations for tests 1-28 are shown in red. (C) Within the 28 laboratory tests that were selected for imputation analyses, the mean number of missing tests is depicted as a function of age. (D) Within the 28 laboratory tests that were selected for imputation, the mean number of missing tests is depicted as a function of body mass index (BMI). (E) Accuracy of a random forest predicting the presence or absence of all 143 laboratory tests. AUROC: area under the receiver operating characteristic curve. (F) Accuracy of a random forest predicting the presence or absence of the top 28 laboratory tests, by Logical Observation Identifiers Names and Codes (LOINC).

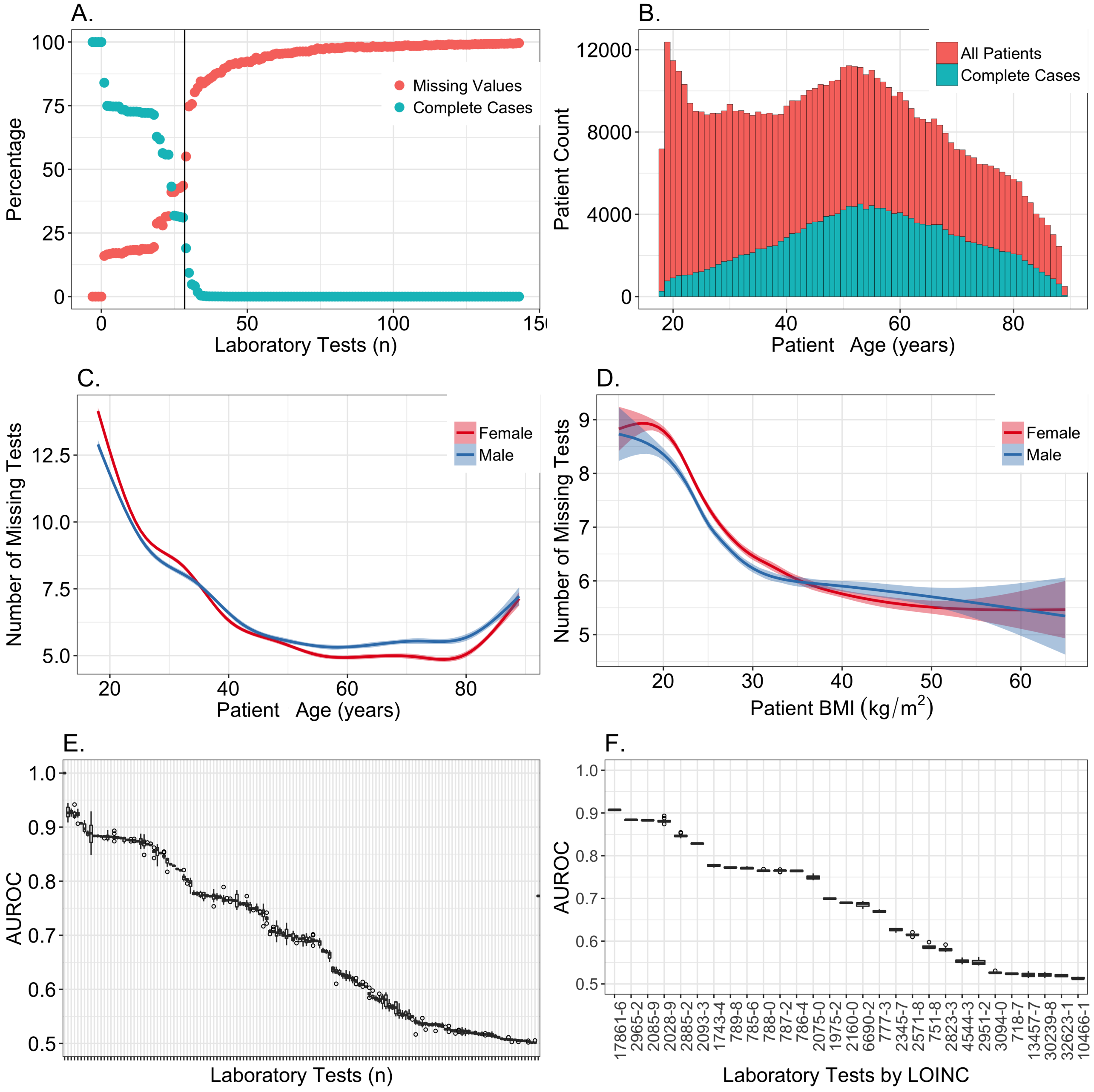


Table 1. Logical Observation Identifiers Names and Codes (LOINC) and descriptions of the most frequently ordered clinical laboratory measurements. The assays are ranked from the most common to the least.

LOINC	Description
718-7	Hemoglobin [Mass/volume] in Blood
4544-3	Hematocrit [Volume Fraction] of Blood by Automated count
787-2	Erythrocyte mean corpuscular volume [Entitic volume] by Automated count
786-4	Erythrocyte mean corpuscular hemoglobin concentration [Mass/volume] by Automated count
785-6	Erythrocyte mean corpuscular hemoglobin [Entitic mass] by Automated count
6690-2	Leukocytes [#/volume] in Blood by Automated count
789-8	Erythrocytes [#/volume] in Blood by Automated count
788-0	Erythrocyte distribution width [Ratio] by Automated count
32623-1	Platelet mean volume [Entitic volume] in Blood by Automated count
777-3	Platelets [#/volume] in Blood by Automated count
2345-7	Glucose [Mass/volume] in Serum or Plasma
2160-0	Creatinine [Mass/volume] in Serum or Plasma
2823-3	Potassium [Moles/volume] in Serum or Plasma
3094-0	Urea nitrogen [Mass/volume] in Serum or Plasma
2951-2	Sodium [Moles/volume] in Serum or Plasma
2075-0	Chloride [Moles/volume] in Serum or Plasma
2028-9	Carbon dioxide, total [Moles/volume] in Serum or Plasma
17861-6	Calcium [Mass/volume] in Serum or Plasma
1743-4	Alanine aminotransferase [Enzymatic activity/volume] in Serum or Plasma by With P-5'-P
30239-8	Aspartate aminotransferase [Enzymatic activity/volume] in Serum or Plasma by With P-5'-P
1975-2	Bilirubin.total [Mass/volume] in Serum or Plasma
2885-2	Protein [Mass/volume] in Serum or Plasma
10466-1	Anion gap 3 in Serum or Plasma
751-8	Neutrophils [#/volume] in Blood by Automated count
2093-3	Cholesterol [Mass/volume] in Serum or Plasma
2571-8	Triglyceride [Mass/volume] in Serum or Plasma
2085-9	Cholesterol in HDL[a] [Mass/volume] in Serum or Plasma
13457-7	Cholesterol in LDL[b] [Mass/volume] in Serum or Plasma by calculation

[a] HDL: high-density lipoprotein.

[b] LDL: low-density lipoprotein.

Figure 3. Area under the receiver operating characteristic curve (AUROC) of a random forest predicting whether data will be present or missing. (A) Missing completely at random simulation. (B) Missing at random simulation. (C) Missing not at random simulation.

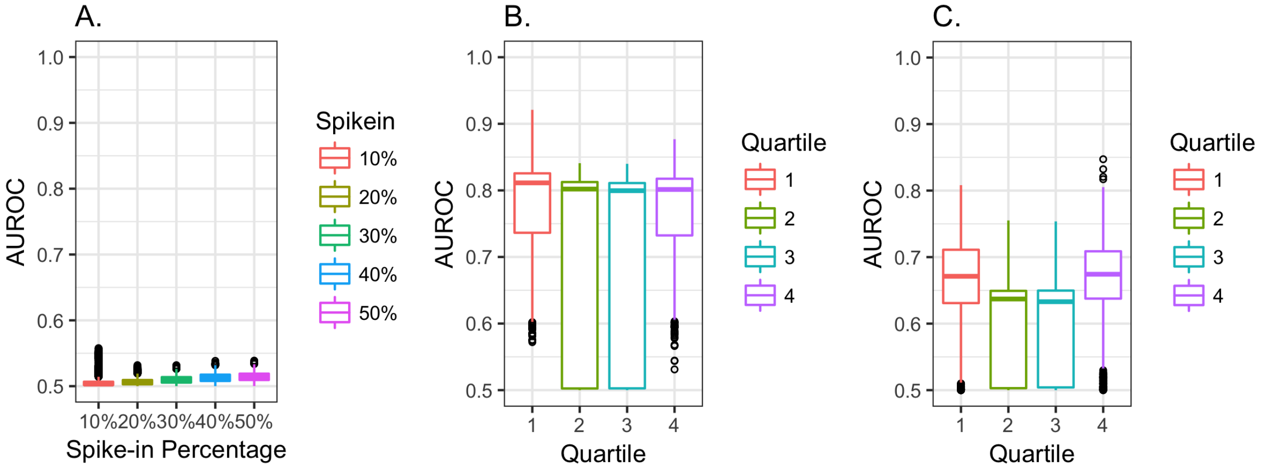


Figure 4. Imputation accuracy measured by root mean square error (RMSE) across simulations 1-3. (A) Missing completely at random (MCAR). (B) Missing at random (MAR). (C) Missing not at random (MNAR). FI: fancyimpute; KNN: k-nearest neighbors; MICE: Multivariate Imputation by Chained Equations; pmm: predictive mean matching; RF: random forest; SVD: singular value decomposition.

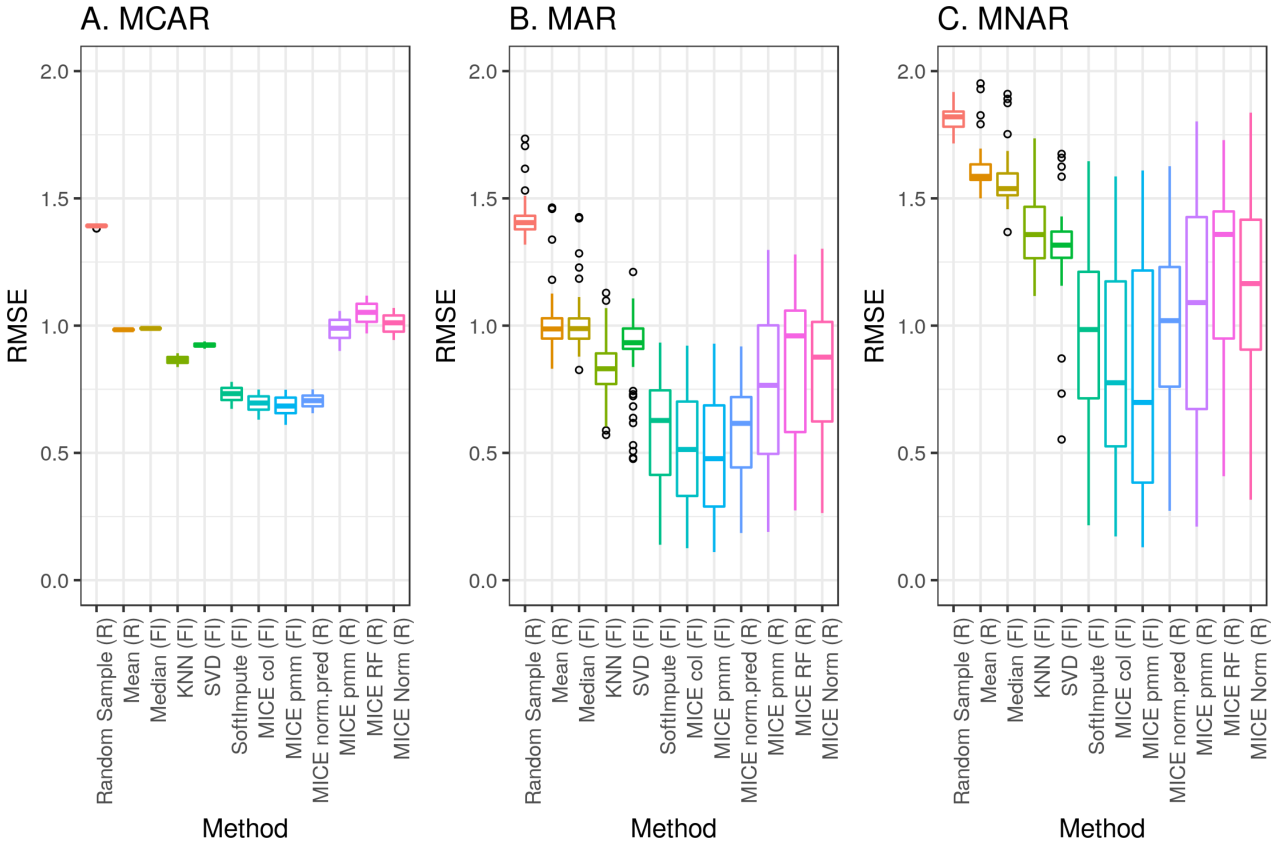
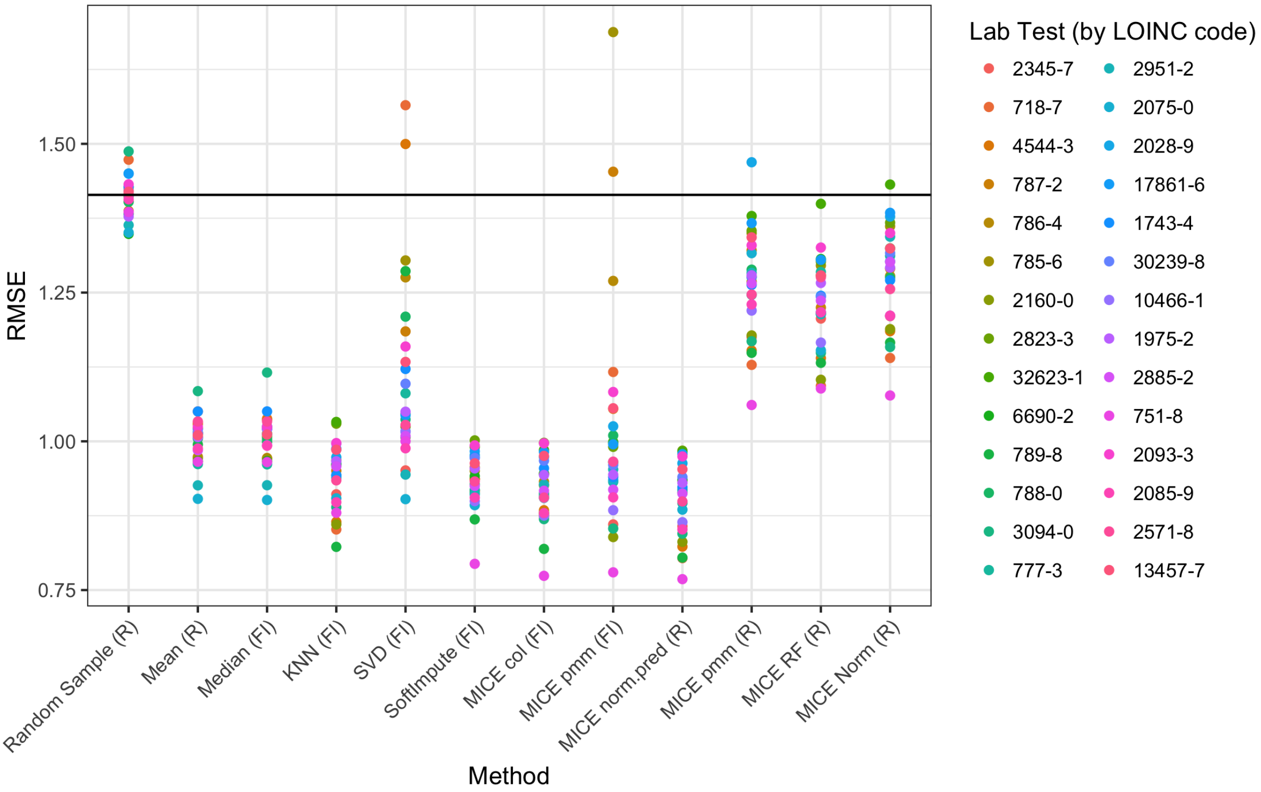


Figure 5. Imputation root mean square error (RMSE) for a subset of 10,000 patients from simulation 4. A total of 12 imputation methods were tested (x-axis), and each color corresponds to a Logical Observation Identifiers Names and Codes (LOINC) code. The black line shows the theoretical error from random sampling. FI: fancyimpute; KNN: k-nearest neighbors; MICE: Multivariate Imputation by Chained Equations; pmm: predictive mean matching; RF: random forest; SVD: singular value decomposition.

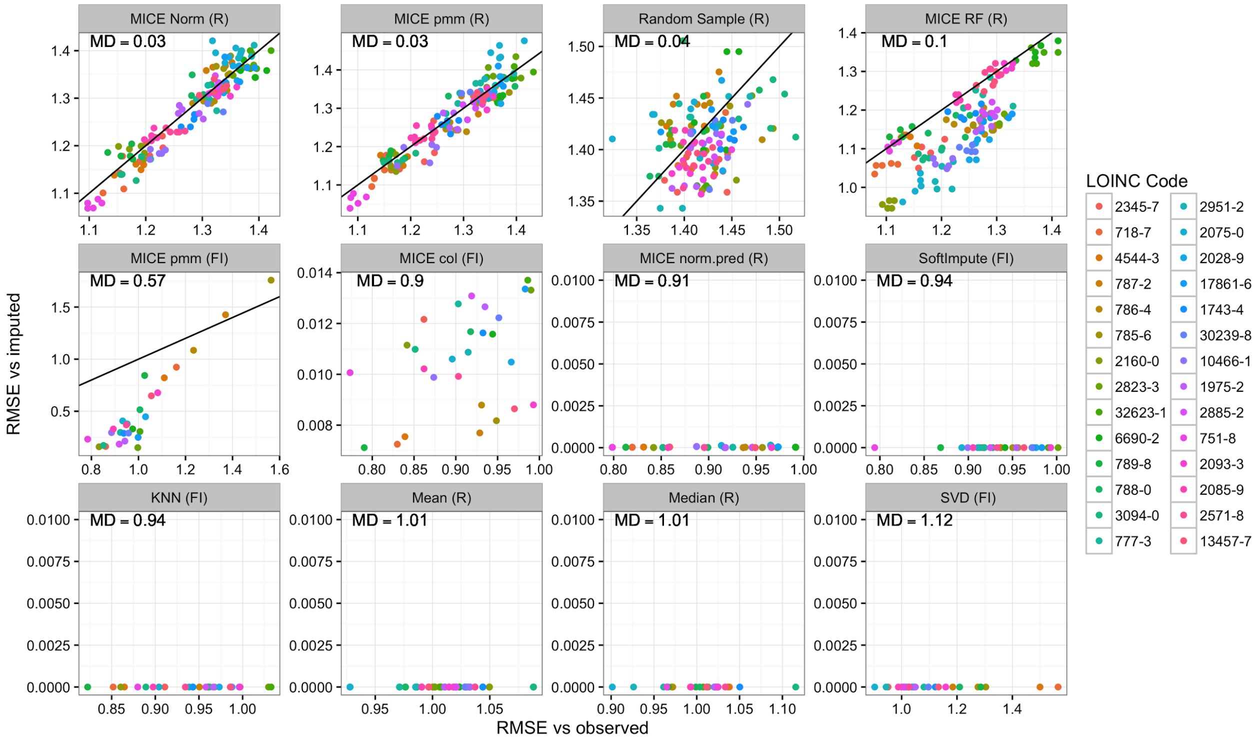


We next measured imputation accuracy based on the patterns of missingness that we observed in the real data (Figure 5). The main difference compared with simulations 1 to 3 was lower error for some of the deterministic methods (mean, median, and KNN). It is worth mentioning that the error was highly dependent on the variable that was being imputed. Specifically, for the fancyimpute MICE predictive mean matching (pmm) method, multicollinearity within some of the variables caused convergence failures that led to extremely large errors (Figure 5, method MICE pmm [FI]). These factors were relatively easy to address in the R package MICE pmm method by adjusting the predictor matrix [15].

In addition to evaluating the accuracy of imputation, it is also important to estimate the uncertainty associated with imputation. One approach to address this is multiple imputation, where each data point is imputed multiple times using a nondeterministic method. To determine whether each method properly captured the true uncertainty of the data, we compared the error between an imputed dataset and the observed data versus the error between 2 sets of imputed values for each method (Figure 6). If these errors are equal, then multiple imputation is likely producing good estimates of uncertainty. If, however, the error between 2 imputed datasets is less than that between each imputed dataset and the known values, then the imputation method is likely underestimating the variance.

Our results (Figure 6) demonstrate that many of the imputation methods are not suitable for multiple imputation. Of the methods that had the lowest error in the MCAR, MAR, and MNAR simulations we found 3 (softImpute, MICE col (fancyimpute), MICE norm.pred (R)) to have minimal variation between imputations. This was also true of KNN, singular value decomposition (SVD), mean, and median imputation. Only 3 methods (random sampling, MICE norm (R), and MICE pmm (R)) seemed to have similar error between the multiple imputations and the observed data and thus appear to be unbiased. The latter 2 had very similar performance and are the best candidates for multiple imputation. Two methods had intermediate performance. MICE random forest (R) was similar to several other MICE methods in terms of error relative to the observed data, but it produced slightly less variation between each imputed dataset. This seemed to affect some variables more than others but there was no obvious pattern. The MICE pmm (fancyimpute) was not deterministic but it did seem to achieve low error at the expense of increased bias. In this case, the variables that could be imputed with the lowest error also seemed to have the most bias. Since this method claims to be a reimplementation of the MICE pmm (R) method, this may be due to multicollinearity among the variables that could not easily be accounted for, as there was no simple way to alter the predictor matrix.

Figure 6. Assessment of multiple imputation for each method. Using simulation 4, missing values were imputed multiple times with each method. The x-axes show the root mean square error (RMSE) between the imputed data and the observed values. The y-axes show the RMSE between multiple imputations of the same data. The axis scales vary between panels to better show the range of variation. The laboratory tests are indicated by the color of the points. The black diagonal line represents unity (y=x). Panels are ordered by each method’s mean deviation (MD) from unity, indicated in the top left corner of each panel. In the last 7 panels, the unity line is not visible because the variation between multiple imputations was close to zero. FI: fancyimpute; KNN: k-nearest neighbors; MICE: Multivariate Imputation by Chained Equations; pmm: predictive mean matching; RF: random forest; SVD: singular value decomposition.



Discussion

Principal Results

It is not possible, or even desirable, to choose “the best” imputation method. There are many considerations that may not be generalizable between different sets of data; however, we can draw some general conclusions about how different methods compare in terms of error, bias, complexity, and difficulty of implementation. Based on our results, there seem to be 3 broad categories of methods.

The first category is the simple deterministic methods. These include mean or median imputation and KNN. While easy to implement, mean or median imputation may lead to severe bias and large errors if the unobserved data are more likely to come from the tails of the observed distribution (Figure 4, parts A-C, methods mean, median, and KNN). This will also cause the variance of the distribution to be underestimated if more than a small fraction of the data is missing. Since these methods are deterministic, they are also not suitable for multiple imputation (Figure 6, bottom row).

KNN is a popular choice for imputation that has been shown to perform very well for some types of data [18,19], but it was not particularly well suited for our data, regardless of the choice of k. This may be due to issues of data dimensionality [20] or to individuals not falling into well-separated groups based on their clinical laboratory results. This method is also not suitable for large datasets, since a distance matrix for all pairs of individuals is stored in memory during computation, and the size of the distance matrix scales with n^2.

The second category of algorithms could be called the sophisticated deterministic methods. These include SVD, softImpute, MICE col, and MICE norm.pred. SVD performed poorly compared with its counterparts and sometimes produced errors greater than simple random sampling (Figure 5, method SVD). The reasons for this are not clear, but we cannot recommend this method. SoftImpute, MICE col, and MICE norm.pred were among the lowest-error methods in all of our simulations (Figure 5, methods MICE col and norm.pred). The main limitation of these methods is that they cannot be used for multiple imputation (Figure 6, middle row).

The third broad category of algorithms comprises the stochastic methods, which included random sampling and most of the remaining methods in the MICE library. Random sampling almost always produced the highest error (Figure 4 and Figure 5, method random sample), but it has the advantage of being easy to implement and it requires no parameter selection. The MICE methods based on pmm, random forests, and Bayesian linear regression tended to perform similarly in terms of error in most of our simulations (Figure 4 and Figure 5, methods MICE pmm, RF, and norm).

Imputation methods that involve stochasticity allow for a fundamentally different type of analysis called multiple imputation. In this paradigm, multiple imputed datasets (a minimum of 3 and often 10-20 depending on the percentage of missing data) [21-23] are generated, and each is analyzed in the same way. At the end of all downstream analyses, the results are then compared. Typically, the ultimate result of interest is supported by a *P* value, a regression coefficient, an odds ratio, etc. In the case of a multiply imputed dataset, the researcher will have several output statistics that can be used to estimate a confidence interval for the result.

Multiple imputation has been gaining traction recently, and the MICE package has become one of the most popular choices for implementing this procedure. This package is powerful and very well documented [15] but, like all methods for imputation, caution must be exercised. In MICE, each variable is imputed one by one. This entire process is then repeated for a number of iterations such that the values imputed in 1 iteration can update the estimates for the next iteration. The result is a chain of imputed datasets, and this entire process is typically performed in parallel so that multiple chains are generated.

In MICE, several choices must be made. The first obvious choice is the imputation method (ie, equation). Many methods are available in the base package, additional methods can be added from other packages [24], and users can even define their own. We thoroughly evaluated 3 methods in the context of our dataset: pmm, Bayesian linear regression (norm), and random forest.

The pmm is the default choice, and it can be used on a mixture of numeric and categorical variables. We found pmm to have a good trade-off between error and bias, but for our dataset it was critical to remove several variables from the predictor matrix due to strong correlations (*R*>.85) and multicollinearity. Bayesian regression performed similarly but was less sensitive to these issues. If a dataset contains only numeric values, Bayesian regression may be a safer option. Random forest tended to produce results that were slightly biased for a subset of the variables without an appreciable reduction in error. Aside from random sampling, none of the other methods we evaluated were suitable for multiple imputation (Figure 6).

Conclusions

Many factors must be considered when analyzing a dataset with missing values. This starts by determining whether each variable should be considered at all. Two good reasons to reject a variable are if it has too many missing values or if it is likely to be MNAR. If a variable is deemed to be MNAR, it may still be possible to impute, but the mechanism of missingness should be explicitly modeled, and a sensitivity analysis is recommended to assess how much impact this could have on the final results [25,26]. While a statistical model of the mechanism of missingness is useful, there is no substitute for a deep familiarity with the data at hand and how they were generated.

Having selected the data, one must select an imputation method. Ideally, several methods should be tested in a realistic setting. Great care should be taken to construct a set of complete data that closely resemble all of the relevant characteristics of the data that one wishes to impute. Similar care should then be taken to remove some of these data in ways that closely resemble the observed patterns of missingness. If this is not feasible, one may also simulate a variety of datasets representing a range of possible data structures and missingness mechanisms. Any available imputation methods can then be applied to the simulated data, and error between the imputed data and their known values provide a metric of performance.

While the minimization of error is an important goal, a singular focus on this objective is likely to lead to bias. For each missing value, it is also important to estimate the uncertainty associated with it. This can be achieved by multiple imputation using an algorithm that incorporates stochastic processes. Multiple imputation has become the field standard because it provides confidence intervals for the results of downstream analyses. One should not naively assume that any stochastic process is free of bias. It is important to check that multiple imputation is providing variability that corresponds to the actual uncertainty of the imputed values using a set of simulated data.

Acknowledgments

We thank Dr Casey S Greene (University of Pennsylvania) for his helpful discussions. We also thank Dr Rebecca V Levy (Geisinger) for providing expert clinical domain knowledge.

This work was supported by the Commonwealth Universal Research Enhancement Program grant from the Pennsylvania Department of Health. BBJ and JM were also supported by US National Institutes of Health grants AI116794 and LM010098 to JM.

Authors' Contributions

BBJ, JM, SAP, and CRB conceived of the study. DRL and JWS performed data processing. BBJ and CRB performed analyses. BBJ, SAP, and CRB wrote the manuscript, and all authors revised and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplemental table and figures.

[PDF File (Adobe PDF File), 4MB - medinform_v6i1e11_app1.pdf]

References

1. Steinbrook R. Health care and the American Recovery and Reinvestment Act. N Engl J Med 2009 Mar 12;360(11):1057-1060. [doi: 10.1056/NEJMp0900665] [Medline: 19224738]
2. Flintoft L. Disease genetics: phenome-wide association studies go large. Nat Rev Genet 2014 Jan;15(1):2. [doi: 10.1038/nrg3637] [Medline: 24322724]
3. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Wash DC) 2013;1(3):1035 [FREE Full text] [doi: 10.13063/2327-9214.1035] [Medline: 25848578]
4. Bounthavong M, Watanabe JH, Sullivan KM. Approach to addressing missing data for electronic medical records and pharmacy claims data research. Pharmacotherapy 2015 Apr;35(4):380-387. [doi: 10.1002/phar.1569] [Medline: 25884526]
5. Bhaskaran K, Smeeth L. What is the difference between missing completely at random and missing at random? Int J Epidemiol 2014 Aug;43(4):1336-1339 [FREE Full text] [doi: 10.1093/ije/dyu080] [Medline: 24706730]
6. Rubin D. Inference and missing data. Biometrika 1976;63(3):581-592 [FREE Full text]
7. Jörnsten R, Ouyang M, Wang HY. A meta-data based method for DNA microarray imputation. BMC Bioinformatics 2007 Mar 29;8:109 [FREE Full text] [doi: 10.1186/1471-2105-8-109] [Medline: 17394658]
8. Beaulieu-Jones BK, Moore JH. Missing data imputation in the electronic health record using deeply learned autoencoders. Pac Symp Biocomput 2017;22:207-218 [FREE Full text] [doi: 10.1142/9789813207813_0021] [Medline: 27896976]
9. Allison P. Missing Data: Sage University Papers Series on Quantitative Applications in the Social Sciences (07-136). Thousand Oaks, CA: Sage; 2001.
10. Beaulieu-Jones B, Lavage D, Snyder J, Moore J, Pendergrass S, Bauer C. Missing data imputation GitHub repository. 2017. URL: https://github.com/EpistasisLab/imputation [accessed 2017-12-01] [WebCite Cache ID 6vOdhupCX]
11. Beaulieu-Jones B, Greene C. Reproducibility of computational workflows is automated using continuous analysis. Nat Biotechnol 2017 Apr;35(4):342-346. [doi: 10.1038/nbt.3780] [Medline: 28288103]
12. Beaulieu-Jones B, Lavage D, Snyder J, Moore J, Pendergrass S, Bauer C. Missing data imputation docker images. 2017. URL: https://hub.docker.com/r/brettbj/ehr-imputation/ [accessed 2017-12-02] [WebCite Cache ID 6vOgQP4YQ]
13. Pedregosa F, Varoquaux G, Gramfort AMV, Thirion B, Grisel O. Scikit-learn: machine learning in Python. J Mach Learning Res 2011;12:2825-2830 [FREE Full text]
14. Rubinsteyn A, Feldman S. fancyimpute: version 0.0.9. 2016. URL: https://github.com/iskandr/fancyimpute [accessed 2018-02-15] [WebCite Cache ID 6xFrmIEfm]
15. Buuren SV, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. J Stat Softw 2011;45(3). [doi: 10.18637/jss.v045.i03]
16. Helfand M, Carson S. Screening for lipid disorders in adults: selective update of 2001 US Preventive Services Task Force Review. U S Prev Serv Task Force Evid Synth 2008 Jun. [Medline: 20722146]
17. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003 Apr;49(4):624-633 [FREE Full text] [Medline: 12651816]
18. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inform Decis Mak 2016 Dec 25;16 Suppl 3:74 [FREE Full text] [doi: 10.1186/s12911-016-0318-z] [Medline: 27454392]
19. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001 Jun;17(6):520-525. [Medline: 11395428]
20. Pestov V. Is the k-NN classifier in high dimensions affected by the curse of dimensionality? Comput Math Appl 2013 May;65(10):1427-1437. [doi: 10.1016/j.camwa.2012.09.011]
21. Stuart EA, Azur M, Frangakis C, Leaf P. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. Am J Epidemiol 2009 May 01;169(9):1133-1139 [FREE Full text] [doi: 10.1093/aje/kwp026] [Medline: 19318618]
22. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011 Feb 20;30(4):377-399. [doi: 10.1002/sim.4067] [Medline: 21225900]
23. Bodner T. What improves with increased missing data imputations? Struct Equ Model 2008 Oct 22;15(4):651-675. [doi: 10.1080/10705510802339072]
24. Robitzsch A, Grund S, Henke T. miceadds: some additional multiple imputation functions, especially for 'mice'. 2017 Dec 18. URL: https://cran.r-project.org/web/packages/miceadds/index.html [accessed 2018-02-15] [WebCite Cache ID 6xFs2gvGZ]
25. Héraud-Bousquet V, Larsen C, Carpenter J, Desenclos J, Le Strat Y. Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data. BMC Med Res Methodol 2012 Jun 08;12:73 [FREE Full text] [doi: 10.1186/1471-2288-12-73] [Medline: 22681630]
26. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. Stat Methods Med Res 2007 Jun;16(3):259-275. [doi: 10.1177/0962280206075303] [Medline: 17621471]

Abbreviations

AUROC: area under the receiver operating characteristic curve
BMI: body mass index
EHR: electronic health record
KNN: k-nearest neighbors
LOINC: Logical Observation Identifiers Names and Codes
MAR: missing at random
MCAR: missing completely at random
MICE: Multivariate Imputation by Chained Equations
MNAR: missing not at random
pmm: predictive mean matching
SVD: singular value decomposition

Edited by G Eysenbach; submitted 14.09.17; peer-reviewed by J Pacheco, I Brooks; comments to author 16.11.17; revised version received 13.12.17; accepted 24.12.17; published 23.02.18.

Please cite as:
Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR
Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis
JMIR Med Inform 2018;6(1):e11
URL: http://medinform.jmir.org/2018/1/e11/
doi:10.2196/medinform.8960
PMID:29475824

Original Paper

Representation of Time-Relevant Common Data Elements in the Cancer Data Standards Repository: Statistical Evaluation of an Ontological Approach

Henry W Chen[1], BS; Jingcheng Du[2], BS; Hsing-Yi Song[2], MPH; Xiangyu Liu[2], BS; Guoqian Jiang[3], MD, PhD; Cui Tao[2], PhD

[1]The University of Texas at Austin, Austin, TX, United States

[2]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States

[3]Mayo Clinic College of Medicine, Rochester, MN, United States

Corresponding Author:
Cui Tao, PhD
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
7000 Fannin, Suite 600 Houston, Texas 77030
Houston, TX,
United States
Phone: 1 7135003981
Email: cui.tao@uth.tmc.edu

Abstract

Background: Today, there is an increasing need to centralize and standardize electronic health data within clinical research as the volume of data continues to balloon. Domain-specific common data elements (CDEs) are emerging as a standard approach to clinical research data capturing and reporting. Recent efforts to standardize clinical study CDEs have been of great benefit in facilitating data integration and data sharing. The importance of the temporal dimension of clinical research studies has been well recognized; however, very few studies have focused on the formal representation of temporal constraints and temporal relationships within clinical research data in the biomedical research community. In particular, temporal information can be extremely powerful to enable high-quality cancer research.

Objective: The objective of the study was to develop and evaluate an ontological approach to represent the temporal aspects of cancer study CDEs.

Methods: We used CDEs recorded in the National Cancer Institute (NCI) Cancer Data Standards Repository (caDSR) and created a CDE parser to extract time-relevant CDEs from the caDSR. Using the Web Ontology Language (OWL)–based Time Event Ontology (TEO), we manually derived representative patterns to semantically model the temporal components of the CDEs using an observing set of randomly selected time-related CDEs (n=600) to create a set of TEO ontological representation patterns. In evaluating TEO's ability to represent the temporal components of the CDEs, this set of representation patterns was tested against two test sets of randomly selected time-related CDEs (n=425).

Results: It was found that 94.2% (801/850) of the CDEs in the test sets could be represented by the TEO representation patterns.

Conclusions: In conclusion, TEO is a good ontological model for representing the temporal components of the CDEs recorded in caDSR. Our representative model can harness the Semantic Web reasoning and inferencing functionalities and present a means for temporal CDEs to be machine-readable, streamlining meaningful searches.

(JMIR Med Inform 2018;6(1):e7) doi:10.2196/medinform.8175

KEYWORDS

common data elements; database management systems; database; time; biomedical ontology

Introduction

Background

With a burgeoning volume of heterogeneous data within the field of health care, health informatics research has focused on finding efficient ways to handle the large influx of new data [1]. One approach is to adopt models to standardize and normalize health care data for efficient data integration and sharing. However, a vast proportion of upwards to 80% of electronic clinical data remains unstructured [2]. Recent efforts on standard terminologies and information models such as Systematized Nomenclature of Medicine—Clinical Terms, *Logical Observation Identifiers Names and Codes*, and OpenEHR archetypes have demonstrated the move toward structuralized electronic health data [3-5].

Semantic interoperability has especially been a key goal of health care systems. Specifically, improvements to the quality and cost of health care are the primary reasons for achieving semantic interoperability within the health care system [6]. Approximately 16% of all reported errors in clinical care are attributed to missing information in patients' electronic health record (EHR) [7]. Additionally, there exists a high level of waste within the health care system [8]. Although a high proportion of the waste comes from the practice of defensive medicine, a significant fraction, constituting $40 million of waste at a single hospital system annually, is the fruit of excessive and unnecessary testing that is the result of the lack of semantic interoperability [9].

Achievement of semantic interoperability has been pursued via representation in the Semantic Web primarily because of its ability to represent the varied features of temporal data. Numerous ontologies have been developed in the recent past, such as CHRONOS, PSI-time ontology, and Resource State/Condition Description Framework ontology [10-12]. Upon reviewing these ontologies, it has been found that overall, these ontologies are lacking in certain key features such as time phase and modality [13]. Additionally, these ontologies were primarily created for general temporal representation and do not specifically address the minutiae of clinical applications. A recently developed ontology, the Time Event Ontology (TEO), addresses the aforementioned shortcomings [14]. TEO, being geared toward temporal annotations in clinical contexts, is utilized and examined in this paper.

There is also a specific need to model temporal relationships within EHRs. In clinical research, time plays an important role in many studies. Temporal reasoning and temporal data management have been identified as two directions of research that are important and relevant to designing architectures for representing the temporal dimension [9]. Temporal reasoning involves the creation of inferred temporal relations between various events. Temporal data maintenance handles the repository of temporal data and the querying of the repository. By modeling temporal relationships with these approaches, study of the time dimension in clinical data becomes possible. For example, careful study of the temporal dimension allows for the elucidation of disease progressions and cause-effect relationships within a clinical setting based on temporal precedent [13].

Current state-of-the-art work in clinical information modeling and extraction includes the HL7 V3 and OpenEHR. Both conform to the ISO 8601 standard as the basis of their syntax [15]. The HL7 V3 represents time based on the following five defined classes: point in time, interval, duration, periodic time, and periodic time as sets [16]. The last class allows HL7 V3 to represent cumulative periodic times. OpenEHR utilizes date, time, date-time, and duration data types [17,18]. OpenEHR allows fields to be missing, allowing for modality to be modeled within the temporal data. These two standardized clinical models can robustly represent temporal data, with each model having its strengths and weaknesses. Unfortunately, these models are only applicable to structured data, leaving out the vast majority of data that is unstructured.

Common data elements (CDEs) have been implemented by the National Cancer Institute (NCI) to answer the need for a standardized format for data collection and storage of clinical trials regarding cancer [19]. Early implementation of CDEs can be observed within the Cancer Informatics Infrastructure [20]. A set of software known as caCORE has been developed to bring together data from various sources to a centralized database. Within caCORE resides the Cancer Data Standards Repository (caDSR), a metadata registry for CDEs. The caDSR is a database supported by the National Cancer Informatics Program that stores these CDEs [21]. Implementation of the caDSR utilizes the ISO/IEC 11179 standard for metadata registries [22]. The ISO/IEC 11179 describes a model for formally associating data model elements with their intended meaning. In the ISO/IEC 11179, a data element is defined as a unit of data for which the definition, identification, representation, and permissible values are specified by means of a set of attributes [22]. The ISO 11179 standard allows the system to determine that two data elements from two different models are alternative representations of the same real world entity [23,24]. The ISO/IEC 11179 specifies an information model by which CDEs are formed and stored within the caDSR by means of a structure based on object and property classes. Although these CDEs provide a useful mechanism to formalize the definitions of intended meaning (ie, data element concepts in the language of ISO/IEC 11179) of a CDE using standard vocabularies (eg, NCI Thesaurus, NCIt), a severe limitation of this representation is the lack of specific semantic relations between the object class annotation and the property class annotation [23,24]. Many times, the object class is simply a plain list, a collection of concept code annotations without semantic relations. This lack of a formal semantic representation presents a problem when attempting to study the temporal relationships associated with a data element concept. As a result, very few studies have focused on the formal representation of temporal relationships associated with a data element concept. Although there exist attempts to represent the temporal relationships within CDEs of caDSR, the lack of standardization still results in ambiguity. For example, ambiguities between the preferred definitions, an abbreviated form of the contents of the CDE, can be seen between CDEs. For CDE 2458736, the preferred definition is PILL_QUANT_DT, whereas for CDE

23 it is OTX_DATE, where DT and DATE both refer to the same meaning. Such ambiguity is highly inconvenient when attempting to study temporal relationships via an ontological approach.

Objective

The primary objective of this research was to represent time-relevant CDEs [22] within the NCI caDSR [25]. Using the Web Ontology Language (OWL) [26] as a technology to model CDEs allows for the leverage of a plethora of reasoning and inference tools available on the Semantic Web. In this paper, we focus on the coverage of patterns developed from the TEO [14], an ontology-based approach to improve semantic representation, on the temporal aspects of CDEs within caDSR.

Methods

Materials

Cancer Data Standards Repository Common Data Elements

The structure of CDEs can be understood by analyzing each component of the CDE. For the purposes of our study, the following fields were useful: (1) DataElement number, (2) PublicID, (3) LongName, (4) PreferredName, (5) PreferredDefinition, and (6) DataElementConcept. The DataElement number and PublicID were used as identifiers for the CDEs. The LongName and PreferredDefinition fields contained information used in generating TEO patterns, which will be explained later in the paper. The PreferredName and DataElementConcept contain current representations of the CDEs in caDSR with NCIt codes.

With the TEO framework, we investigated its usage in representing the temporal components of the CDEs. By using the various OWL classes of TEO, we can generate *building blocks* whereby temporal components of a CDE can be organized and classified. The *building blocks* can simply be described as the representational patterns in the Resource Description Framework (RDF) triples that are built using TEO. This ultimately affords the creation of parsable and therefore, machine-readable, temporal elements of the CDEs within caDSR.

Semantic Web and Web Ontology Language

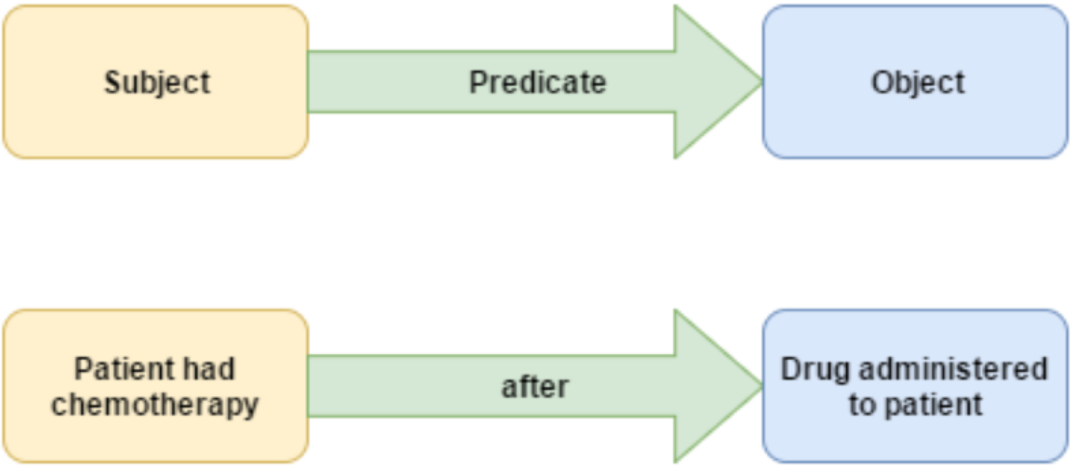
Our efforts focused on addressing the issues regarding (1) giving structure to the vastly unstructured data within the CDEs, (2) capturing temporal relationships between events in the CDEs stored within caDSR, and (3) organizing the data to be machine-readable and processable as opposed to simply being human-readable. To achieve these goals, we took advantage of the Semantic Web and OWL [26]. By representing the temporal dimension with OWL, many of the reasoning capabilities available on the Semantic Web can be leveraged. The temporal relationships themselves can be annotated using an ontology and stored as RDF triples (Figure 1) [27].

Time Event Ontology

TEO is an ontology designed for a formal conceptualization of time-related information (eg, temporal expressions, temporal relations, and granularities of time) in both structured data and textual narratives. The design of TEO was based primarily on its predecessor, the Clinical Narrative Temporal Relation Ontology (CNTRO), a Semantic Web ontology created for representing temporal relationships within clinical narratives [13]. Although CNTRO was primarily focused on annotating clinical narratives, TEO was designed with the goal of annotating a very general category of temporal relationships. In addition, TEO has been refined to cover more semantic features such as finer level of granularity of temporal relations, standard representations of temporal durations, and more sophisticated representations for reoccurred events. The general architecture of TEO can be seen in Figure 2.

To understand how TEO patterns are generated, an elementary understanding of the components of TEO is required. The following section presents a brief overview of the components and their meanings to lay the groundwork for understanding the TEO patterns used to represent the temporal component of the CDEs.

Figure 1. Resource Description Framework (RDF) triple example.



TEO is composed of the following OWL classes: *Event, Time, TimeInstant, TimeInterval, TimePhase, Duration, Granularity,* and *TemporalRelationStatement*. Object properties and data properties are also defined to represent relations and attributes of the classes. Additionally, the TEO framework allows various OWL classes to be interconnected via OWL classes that act as predicates.

The *Event* class is simply defined as any occurrence. Each instance of an *Event* can be related to another instance of *Event* via the *hasTemporalRelation* property or to an instance of the *Time* class via the *hasValidTime* or *hasTemporalRelation* property. The detailed temporal relations in TEO are defined and extended on top of Allen's temporal algebra [28].

The *Time* class is defined as a superclass of the *TimeInstant* and *TimeInterval* classes. A *TimeInstant* can be conceptualized by any event that can be represented by a discrete time point within a given time line. For example, "28 APR 2017" can be represented by *TimeInstant.* The granularity of the *TimeInstant* can be represented using the object property *hasGranularity* with domain *Granularity* that defines a predefined set of temporal granularities, including seconds, minutes, days, etc. A *TimeInterval* can be connected to two instances of *TimeInstant* that represent the start time and end time via the *hasStartTime* and *hasEndTime* properties. Additionally, the duration of the *TimeInterval* can be represented with the *Duration* class. For example, in "Around 06 APR 2017, the infant developed constipation, which persisted as of 28 APR 2017," the time of "constipation" can be represented by a *TimeInterval.* The duration of this *TimeInterval* (22 days) can be represented by a *Duration* class. The *Duration* class is linked with the properties *hasDurationPattern*, which formally defines each duration. For example, we can use "5D10H" to represent "five days and ten hours." It is important to note that an instance of *TimeInterval* is not required to have all three components previously listed. However, to be formally defined as a TimeInterval for reasoning purposes, it is required that two of the three components be defined. This allows for the third missing component to be inferred via a reasoner.

Within the *Time* class, the *TimePhase* class is defined as an extension of the *TimeInterval* class with additional properties. *TimePhase* is a special case of the *TimeInterval* and is composed of multiple instances that reoccur periodically. For example, in "Judy has swum 2 hours a day for 6 months," the "2 hours a day for 6 months" is a *TimePhase* (Textbox 1). The *hasRepeatTime* parameter stores an integer that describes how many times the instances reoccur. The *hasRepeatUnitInterval* property connects to an instance of *Duration* representing the time between two recurring instances. The *hasRepeatUnit* can store either an instance of *Duration* or another *TimePhase*, allowing nesting of multiple *TimePhase* instances. In the above example, the *TimePhase* has the *hasRepeatUnit* property that stores a *Duration* of 2 hours. The *hasPeriod* property connects to an instance of *Duration* representing the sum of the duration between two recurring instances and the duration of the instance. In the above example, the *TimePhase* has the *hasPeriod* property with a *Duration* of 1 day and the *hasDuration* property with a *Duration* of 6 hours. Again, all properties are not required to be specified, but a minimum number is necessary to adequately infer the rest via a reasoner.

Finally, the *TemporalRelationStatement* class is used to add constraints to an RDF triple. It has a built-in *hasApproximation* parameter to account for any temporal uncertainty. For example, in "his constipation may have started before the medication," the RDF triple [constipation][*before has Approximation: True*]["medication"]. Additionally, it can store an instance of *Duration* within the *hasTemporalOffset* parameter to, for example, specify a duration of time after an event occurs.

Textbox 1. Resource Description Framework (RDF) representation of time phase example. Bold font indicates the class, and italic font indicates the property.

<tPhase1> rdf:type **TimePhase**;

hasDuration durat1;

hasRepeatUnit durat2;

hasPeriod durat3;

<durat1> rdf:type **Duration**;

hasDurationPattern 6M;

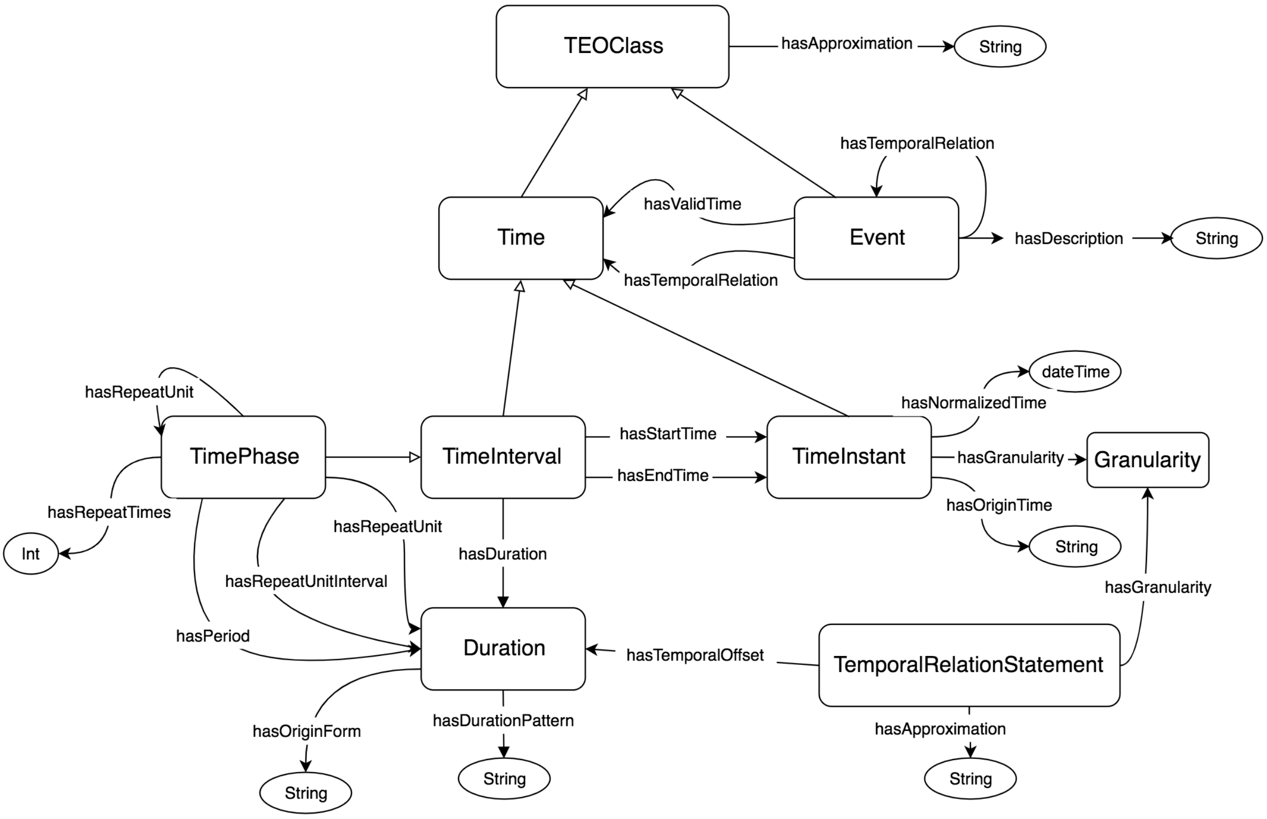
<durat2> rdf:type **Duration**;

hasDurationPattern 2H;

<durat3> rdf:type **Duration**;

hasDurationPattern 1D;

Figure 2. Graphical representation of Time Event Oncology (TEO).



Methods

Identifying Temporal Components Within Common Data Elements

It was necessary to retrieve the CDEs that contain a temporal component from the general population of CDEs. Thus, the CDE parser was created and utilized to accomplish this goal. The NCI offers a CDE browser that we used to obtain the CDEs utilized for our analysis. The CDEs were downloaded in .xml format from the CDE browser as of August 4, 2015.

Of the 42,956 CDEs within caDSR that were downloaded, 7369 were identified to have at least one temporal component. This was accomplished using the CDE parser to target certain keywords within the LongName and PreferredDefinition fields. Each keyword was assigned to a particular TEO class, which will hereafter be referred to as building blocks (Table 1). These building blocks help to inform the annotator of the contents of the CDE. The building blocks are output alongside the parsed CDEs. The keywords are regular expressions that allow for a wide range of temporal information to be captured by one keyword. For example, dates (eg, December 31, 2015) can be easily represented using regular expressions. However, in the case of caDSR, these keywords were not found in any of the CDEs. In the future, regular expressions can easily be added to the CDE parser as the need arises.

Textbox 2 provides an example of a CDE that has been extracted using the CDE parser. The aforementioned fields are present along with the building blocks, marked in bold, that partially compose the CDE. In this particular example, the CDE parser identified the keywords "interval" and "date." These were assigned to the TEO classes TimeInterval and TimeInstant/TimeInterval/Date, respectively. The reason that there are multiple TEO classes assigned to a specific keyword is because of the inherent ambiguous nature of the keywords. These keywords can serve as a guide to the annotator but are primarily used to extract time-relevant CDEs. Because the keywords are simply a guide, this allows the annotator flexibility in assigning the TEO classes and creating the patterns that will be described in the next section of the paper.

Time Event Ontology Pattern Generation

To generate the representation patterns, the temporal aspects of the caDSR CDEs in the observing set were manually annotated using TEO as an ontological basis. The patterns were created by taking into account the building blocks identified by the CDE parser, as well as the LongName and PreferredDefinition fields. By taking into account the PreferredDefinition in conjunction with the LongName, it can be assured that the CDE is assigned an appropriate pattern.

Table 2 provides an example of a CDE that has been annotated using a TEO pattern. The TEO patterns can be constructed by having the annotator first look at the LongName field to get a general idea of the content of the CDE. The PreferredDefinition field can be used to confirm the content of the CDE. In this case, the pattern [Event*] [TemporalRelation] [Event] can be used to represent the temporal aspect of the CDE. The TemporalRelation block stores the *before* temporal relation that relates the "treatment type" to the "surgical procedure." The second Event in the pattern stores the "surgical procedure type." All unasterisked fields are assumed to be static information defined by the CDE. Static information defined by the CDE is assumed to be constant across all instances of the CDE. In Table 2, the surgery and temporal relation of *before* is considered static information because this information is constant for all instances of the CDE. The starred Event stores the treatment that was given before the surgical procedure type. This starred Event is assumed to be variable based on what kind of information is stored within the CDE, which can change among the different instances of CDEs. TEO does not define any subclasses under the Event class with the assumption that each application of the TEO could further define subclasses that are specific to that domain. In this case, we could define the type of *event1* as Treatment, which is a subclass of Event if needed.

Table 1. Keywords represented with regular expressions delimited by commas and their corresponding Time Event Ontology (TEO) class.

Keyword regular expressions	TEO[a] class (building blocks)
Jan(uary)?,Feb(ruary)?,Mar(ch)?,Apr(il)?,May,June,July,Aug(ust)?,Sept(ember)?,Oct(ober)?,Nov(ember)?,Dec(ember)?,today,morning,night,date	TimeInstant
	TimeInterval
	Date
seconds,minutes?,hours?,days?,weeks?,months?,years?	Granularity
	Duration
before,while,prior to, ago,previous(ly)?,post(-)?,subsequent,concurrent(ly)?,meets?,overlaps?,finish(es)?,starts?,during,after,within,until,when	TemporalRelation
	TimeOffset
recurrent,frequent,intermittent,periodic,repeat(ed)?	TimePhase
Interval	TimeInterval

[a]TEO: Time Event Ontology.

Textbox 2. Example of a common data element (CDE) parsed with the CDE parser. Bold font indicates the class.

[TimeInterval, TimeInstant/TimeInterval/Date]

```
<DataElement num="36405">
<PUBLICID>4199738</PUBLICID>
<LONGNAME>QT Interval Medication Administered Last Date</LONGNAME>
<PREFERREDNAME>4199693v1.0:2192181v1.0</PREFERREDNAME>
<PREFERREDDEFINITION>
information related to the date QT interval medication last administered.
</PREFERREDDEFINITION>
<DATAELEMENTCONCEPT>
<PreferredName>4199691v1.0:2233610v1.0</PreferredName>
</DATAELEMENTCONCEPT>
```

Table 2. Common data element (CDE) annotated with a Time Event Ontology (TEO) pattern. Bold font indicates the class, and italic font indicates the property.

Representation type	Content
CDE[a]	**[TemporalRelation/TimeOffset]** <DataElement num="44077"> <LONGNAME>Treatment Given Prior To Surgical Procedure Type</LONGNAME> <PREFERREDDEFINITION>Text term to describe the kind of treatment given to an individual prior to surgery.</PREFERREDDEFINITION>
TEO[b] pattern	**[Event*]** [TemporalRelation] [Event]
Extended TEO pattern	**[Event=Treatment*]** [TemporalRelation=before] [Event=Surgical Procedure Type]
RDF[c] triple representation	<event1> rdf:type **Event (Treatment)**; rdfs:label *****; *before* <event2>; <event2> rdf:type Event; rdfs:label "Surgical Procedure Type";

[a]CDE: common data element.

[b]TEO: Time Event Ontology.

[c]RDF: Resource Description Framework.

Results

Common Data Element Parser Performance

First, it was important for us to analyze the sensitivity and specificity performance of the CDE parser to confirm that the CDEs extracted actually contained a temporal component. True positive denotes the CDEs correctly identified as containing a temporal component. True negative denotes the CDEs correctly excluded from the time-relevant CDEs. False positive denotes the CDEs that do not contain a time component but were retrieved by the CDE parser. False negative denotes the CDEs that have a time component but were not retrieved by the CDEs.

In our analysis of the CDE parser and TEO pattern performance, we performed two iterations of annotation with the first iteration serving as a pilot set to obtain a general idea of CDE parser and TEO pattern performance and to generate a second set of data with a nonarbitrary sample size. To evaluate the CDE parser performance, we used the data from the second, more statistically robust iteration of annotation. Additionally, two sets of data (n=425) were randomly generated from the population of CDEs that were not retrieved by the CDE parser, referred to as complement sets hereafter. These complement sets were used in our analysis to find potential false negatives. In other words, we hoped to identify CDEs with temporal aspects that were not parsed by the CDE parser. Analyses of these complement sets allowed us to identify true negatives and false negatives. As before, two sets of data were used to test for consistency among the complement sets. Three annotators independently examined the complement sets. These two complementary sets of data were used in conjunction with the two test sets (n=425) from earlier. The results are presented in Table 3. Sensitivity values were calculated using the following equation:

Specificity values were calculated using the following equation:

Both sensitivity and specificity parameters exhibit good performance. Interestingly, the sensitivity values are higher on average than the specificity values. This indicates that the performance of the CDE parser could be improved by refining the keywords list to ignore CDEs that do not actually possess a time element.

Table 3. Sensitivity and specificity data of common data element (CDE) parser.

Annotator	Test set	True positive	True negative	False positive	False negative	Sensitivity	Specificity
1	1	398	408	27	17	0.959036	0.937931
	2	404	415	21	10	0.975845	0.951835
2	1	394	418	31	7	0.982544	0.930958
	2	394	412	33	13	0.968059	0.925843
3	1	391	414	34	11	0.972637	0.924107
	2	397	418	28	7	0.982673	0.93722

Time Event Ontology Pattern Evaluation

Interannotator Agreement

Because the test sets were annotated independently by three annotators, it was necessary to examine the interannotator agreement between the patterns assigned by the three annotators (Table 4). In analyzing the interannotator agreement, the CDEs could be categorized into one of three categories: (1) no difference, meaning that all three annotators assigned the same or equivalent pattern to the CDE; (2) one difference, meaning that two annotators assigned the same or equivalent pattern to the CDE, but one annotator assigned a different pattern; or (3) all different, meaning that all three annotators assigned a different pattern to the CDE.

The CDEs assigned to the one difference category are simply assigned to the pattern that two out of the three annotators used. Upon examination of the CDEs that fall under this category, it was found that the intended meaning behind many of these CDEs were very similar. For example, two annotators annotated a CDE as a TimeInstant, whereas one annotated a CDE as a TimeInstant as an end time of a TimeInterval. However, because the patterns were not exactly identical, they are considered to fall under the *one difference* category. At the root of this discrepancy is likely the misinterpretation of the CDE because of the lack of expertise regarding the contents of the CDE. A domain expert or the creator of the CDE would easily solve this ambiguity problem.

With three differences there exists the problem of being unable to assign a pattern to the CDE because of all the annotations being different. These CDEs would require a domain expert to properly annotate them. We see from the data that the vast majority of the CDEs can be assigned a pattern either by having no difference in the pattern assigned by the annotators or by having two differences whereby the pattern that is used by the majority of annotators is used.

Time Event Ontology Pattern Coverage

We were interested in analyzing the coverage of TEO patterns on a randomly generated set of CDEs with a temporal aspect. From the 7369 CDEs identified to have a temporal aspect, we chose to generate a pilot set with an arbitrary size of n=600. The pilot set was randomly partitioned into an observing set of n=300 and three test sets of n=100. Three sets of n=100 were generated to test for consistency of pattern occurrence among the three test sets. The observing set was used to produce a variety of TEO patterns that could be applied to the CDEs in the test sets. CDEs within the test sets could fall into one of four categories: existing pattern, new pattern, not time-related, and nonrepresentable with TEO (Table 5). The approximate proportion of each classification of CDEs within caDSR is also shown in Table 5 based on our analysis. CDEs that can be represented with a pattern generated from the observing set are *existing patterns*. If a CDE cannot be represented by any of the patterns generated in the observing set, but a new pattern can be generated to represent that CDE, then it is classified as a *new pattern*. On the other hand, the other two sections represent CDEs that are either not time-related at all (*not time-related*), a fault of the parser, or nonrepresentable by TEO because of shortcomings in TEO (*TEO cannot represent*).

The initial pilot set of n=600 was used for two reasons: (1) to train annotators on how to annotate the CDEs with TEO and (2) to garner a general idea of how well TEO can represent the various temporal aspects of the CDEs. The analysis from the first pilot set allowed us to generate a new set of data with a nonarbitrary sample size. We chose to look at the robustness of the CDE parser by analyzing the sensitivity of the parsed CDEs. The sensitivity values were then used to calculate the sample size of the second test set. For the purposes of the calculation, the existing pattern and new pattern sections in Table 5 are important. The existing pattern section is denoted as *true positive* because it is for CDEs that are captured by the manually derived existing patterns. The new pattern section is denoted as *false negative* because the pattern would have been designated as nonrepresentable based on the existing patterns. Thus, true positive and false negative can be used to calculate sensitivity. It should be noted here that we did not calculate specificity in the context of TEO pattern coverage. This is because of the four categories presented in Table 5, none of them fall into the category of a *false positive*. This would result in a trivial specificity value of 1 for all test cases. Thus, we chose to simply utilize sensitivity in the TEO pattern coverage analysis. Additionally, we can utilize the *existing pattern*, *new pattern*, and *TEO cannot represent* sections of Table 5 to calculate the coverage rate of TEO for time-related CDEs using the following equation:

The coverage rate is calculated for the final test sets later in the paper.

Table 6 presents the results of three annotators on the three different test sets within the initial pilot set. The margins of error were calculated using the following equation:

where *TP* denotes the number of true positive instances. is the sensitivity expressed in decimals. $Z_{\alpha/2}$ the z-value, which in our case is 1.96, representing a 95% CI. The sensitivity values, as well as the margins of error between the data sets and annotators are insignificantly different. Thus, we were able to use these results to generate a second test set.

Table 4. Interannotator agreement data (N=425).

Test set number	No difference, n (%)	One difference, n (%)	All different, n (%)
1	279 (65.6)	133 (31.2)	13 (3.0)
2	258 (60.7)	146 (34.3)	21 (4.9)

Table 5. Test set common data element (CDE) categorization (N=300).

Category	n (%)
Representable CDEs [a]	
Existing pattern	263 (87.7)
New pattern	9 (2.9)
Nonrepresentable CDEs	
Not time-related	20 (6.8)
TEO[b] cannot represent	8 (2.6)

[a] CDE: common data element.

[b] TEO: Time Event Ontology.

Table 6. Pilot set annotation results.

Annotator	Test set number	Number of TP[a,b]	Number of FN[c,d]	Sensitivity	Margin of error
1	1	85	2	0.977	0.032
	2	82	5	0.943	0.050
	3	83	9	0.902	0.064
2	1	86	5	0.945	0.048
	2	82	7	0.921	0.058
	3	77	11	0.875	0.074
3	1	89	3	0.967	0.037
	2	84	8	0.913	0.060
	3	83	9	0.900	0.064

[a] TP: true positive.

[b] Denotes the number of true positive instances.

[c] FN: false negative.

[d] Denotes the number of false negative instances.

The following equation was used to obtain a sample size from the sensitivity and margin of error values:

$Z_{\alpha/2}$ is, again, the z-value, which in our case is 1.96, representing a 95% CI. , again, is the sensitivity expressed in decimal form. From the previously mentioned equation, *d*, the margin of error was calculated. On the basis of the data in Table 6, the lowest sensitivity and margin of error were used in the equation for calculating sample size. This results in the largest sample size and subsequently a more representative sample of the population of parsed CDEs. The resultant sample size was n=410 and was rounded to n=425 to yield a more rounded number while still preserving the representative sample size. By rounding up, we are able to preserve the representative sample size, whereas rounding down would result in a less representative sample size. An observing set of n=600, determined by doubling the size of the previous observing set, and two test sets of n=425 were randomly generated from the population of time-related CDEs retrieved by the CDE parser. We note here that it is not necessary to determine the size of the observing set through statistics as it is merely collecting patterns for use in the test sets. The annotation process was repeated to gather TEO coverage data from a population of parsed CDEs with a statistically significant sample size [29].

We present the results of the second iteration of annotation in Table 7. The arithmetic mean of the coverage rates is 94.2% (801/850). In the larger, more representative test set with a statistically significant sample size, the coverage rate of TEO for the time-related CDEs was greater than 90% for all test sets, demonstrating TEO's effectiveness at representing the time-related CDEs parsed. Additionally, these values have a low spread and are consistent with each other. This demonstrates that a high proportion of the time-relevant CDEs that were retrieved by the CDE parser are representable by TEO patterns.

Table 7. Statistically significant test set results.

Annotator	Test set number	Coverage rate
1	1	0.950
	2	0.940
2	1	0.949
	2	0.913
3	1	0.964
	2	0.935

Table 8. Most frequently used Time Event Ontology (TEO) patterns used in the observing set of N=600, averaged over three annotators.

Rank	TEO[a] pattern	n (%)
1	[Event (hasValidTime=[TimeInstant (hasGranularity, hasOrigTime*)])]	186 (31.0)
2	[Event* (hasValidTime=[TimeInterval (hasEndTime=[TimeInstant (hasOrigTime)], hasDuration=[Duration (hasDurationPattern)])])]	117 (19.5)
3	[Event (hasValidTime=[TimeInstant (hasNormalizedTime*)])]	90 (15.0)
4	[Event*] [TemporalRelation] [Event]	42 (7.0)
5	[Event (hasModality*)] [TemporalRelation] [Event]	35 (5.9)
6	[Event (hasValidTime=[TimeInterval (hasEndTime=[Time Instant (hasGranularity,hasOrigTime*)])])]	32 (5.4)
7	[Event (hasValidTime=[TimeInterval (hasStartTime=[Time Instant (hasGranularity,hasOrigTime*)])])]	26 (4.4)
8	[Event* (hasModality*,hasValidTime=[TimeInterval(hasEndTime=[TimeInstant(hasOrigTime)], hasDuration=[Duration(hasValue,hasUnit)])])]	25 (4.2)
9	[Event (hasValidTime=[TimeInterval(hasDuration=[Duration(hasDurationPattern*)])])]	17 (2.8)
10	[Event(hasValidTime=[TimeInterval(hasStartTime=[TimeInstant(hasOrigTime*)],hasEndTime=[TimeInstant(hasOrigTime*)])])]	11 (1.8)

[a]TEO: Time Event Ontology.

Pattern Frequency

While annotating the CDEs with TEO patterns, it became quite evident that many of the CDEs could be characterized by a few patterns. Table 8 lists the top ten most-used patterns in the observing set of n=600, accounting for >97% of all CDEs. In conjunction with Table 8 and Table 9 presents a specific example of each pattern with the corresponding RDF format. The first and third most popular patterns are to be expected because many of the CDEs could simply be classified as storing a date or timestamp. The second most frequent pattern stores the many CDEs that store an answer to a question. This pattern is used to represent CDEs that have questions that ask about some occurrence within a past time frame. We believe that this pattern is a testament to the flexibility of the TEO patterns. The different classes of TEO can be manipulated in a variety of ways to represent a wide variety of temporal aspects within CDEs as shown by the top ten most frequent patterns.

Table 9. Specific examples in Resource Description Framework (RDF) format of most frequently used Time Event Ontology (TEO) patterns. Bold font indicates the class, and italic font indicates the property.

Rank	PublicID	CDE[a] LongName	RDF[b] representation	
1	4614514	Stage IV disease progression platinum-based chemotherapy date	<event1>	rdf:type **Event**; rdfs:label "Stage IV Disease Progression Platinum-Based Chemotherapy"; *hasValidTime* <tInstant1>;
			<tInstant1>	rdf:type **TimeInstant**; rdf:label "Date" *hasGranularity* *; *hasOrigTime* *;
2	3191975	Patient reported outcome problem dysuria past week severity score 11 point scale	<event1>	rdf:type **Event**; rdfs:label *; *hasValidTime* <tInterval1>;
			<tInterval1>	rdf:type **TimeInterval**; *hasEndTime* tInstant1;

			<tInstant1>	rdf:type **TimeInstant**;
				hasOrigTime date_of_CDE;
				hasDuration durat1;

			<durat1>	rdf:type **Duration**; *hasDurationPattern* 1 week;
3	3100972	Customer request laboratory final approval date java.util. date	<event1>	rdf:type **Event**; rdfs:label "Customer Request Laboratory Final Approval"; *hasValidTime* <tInstant1>;
			<tInstant1>	rdf:type **TimeInstant**; *hasNormalizedTime* *;
4	2683245	Breast conservation treatment post neoadjuvant therapy not attempt specify	<event1>	rdfs:label *; rdf:type **Event**; *after* < event2>;
			<event2>	rdf:type **Event**; rdfs:label "Neoadjuvant Therapy";
5	3387810	Maintenance therapy prior recurrent disease discontinue indicator	<event1>	rdf:type **Event**; rdfs:label "Maintenance Therapy Discontinue"; *hasModality* *; *before* < event2>;
			<event2>	rdf:type **Event**; rdfs:label "Recurrent Disease";

Rank	PublicID	CDE[a] LongName	RDF[b] representation	
6	2790	Partial response observed end date	<event1>	rdf:type **Event**;
				rdfs:label “Partial Response Observed”;
				hasValidTime <tInterval1>;
			<tInterval1>	rdf:type **TimeInterval**;
				hasEndTime tInstant1;
			<tInstant1>	rdf:type **TimeInstant**;
				hasGranularity *;
				hasOrigTime *;
7	1157	Prior RT begin date	<event1>	rdf:type **Event**;
				rdfs:label “RT”;
				hasValidTime <tInterval1>;
			<tInterval1>	rdf:type **TimeInterval**;
				rdf:label “Prior”;
				hasStartTime tInstant1;
			<tInstant1>	rdf:type **TimeInstant**;
				hasGranularity * **;**
				hasOrigTime *;
8	4609733	FACT-Cog Questionnaire version 3 CogPM1 how true past seven days have been able to remember things score 5 point scale	<event1>	rdf:type **Event**;
				rdfs:label *****;
				hasModality *;
				hasValidTime <tInterval1>;
			<tInterval1>	rdf:type **TimeInterval**;
				hasEndTime tInstant1;
			<tInstant1>	df:type **TimeInstant**;
				r *hasOrigTime* date_of_CDE;
				hasDuration durat1;
			<durat1>	rdf:type **Duration**;
				hasDurationPattern 7 days;
9	3190457	Person clinical study assignment follow-up month duration	<event1>	rdf:type **Event**;
				rdfs:label Personal Clinical Study Assignment Follow-up;
				hasValidTime <tInterval1>;
			<tInterval1>	rdf:type **TimeInterval**;
				hasDuration durat1;
			<durat1>	rdf:type **Duration**;
				*hasDurationPattern******;

Rank	PublicID	CDE[a] LongName	RDF[b] representation	
10	3177036	Adverse event outcome assessment observation performed study activity actual date and time range ISO21090.IVL.TS.DATETIME.v1.0	<event1>	rdf:type **Event**; rdfs:label “Adverse Event Outcome Assessment Observation Performed Study Activity” *hasValidTime* <tInterval1>;

			<tInterval1>	rdf:type **TimeInterval**;
				hasStartTime tInstant1;
				hasEndTime tInstant2;

			<tInstant1>	rdf:type **TimeInstant**; *hasOrigTime**;
			<tInstant2>	rdf:type **TimeInstant**; *hasOrigTime**;

[a] CDE: common data element.

[b] RDF: Resource Description Framework.

Discussion

Comparison With Current Standard Representation in Cancer Data Standards Repository

It is important to note here that TEO is not intended to replace the current standard representation of CDEs within caDSR but rather enhance the representation of temporal components. Although there is a standard representation of the CDEs already implemented, which is stored in the PreferredDefinition field, it does not consistently represent the temporal components of the CDEs [22]. Table 10 demonstrates some inconsistencies within the standard representation that TEO hopes to address. The temporal components within the Preferred Definition and TEO pattern are bolded. It can be seen that given a CDE with the same TEO pattern, the Preferred Definition field uses a different code to represent the CDE. These inconsistencies are resolved by using the TEO patterns. Thus, it can be seen that the TEO patterns are superior at representing the temporal component of the CDEs.

Within caDSR, there also exist inconsistencies between the LongName field and the PreferredDefinition field in some CDEs. Textbox 3 presents an example of a CDE that has a PreferredDefinition field that is inconsistent with the LongName. The LongName implies that the CDE is an indicator for whether a dental procedure known as a post core is used. However, “post” in the LongName is represented as a temporal relation within the PreferredDefinition field. The TEO patterns would allow for clearer representation of these CDEs by allowing the author of the CDE to designate the post core as an Event to clarify any ambiguities.

Table 10. Example standard representation of common data elements (CDEs) versus Time Event Ontology (TEO) patterns.

LongName	PreferredDefinition	TEO[a] pattern
Off treatment date	OTX_DATE	[Event (hasValidTime=[TimeInstant (hasGranularity, hasOrigTime*)])]
Pills quantity date	PILL_QUANT_DT	[Event (hasValidTime=[TimeInstant (hasGranularity, hasOrigTime*)])]
Therapy prior carmustine administered end date	BNCU_ENDDT	[Event (hasValidTime=[TimeInterval (hasEndTime=[TimeInstant (hasGranularity,hasOrigTime*)])])]
Laboratory data inclusion stop date	LAB_INCL_STOP_DT	[Event (hasValidTime=[TimeInterval (hasEndTime=[TimeInstant (hasGranularity,hasOrigTime*)])])]
Breast conservation treatment post neoadjuvant therapy failed performed reason	BCT_P_NEO_FA_PER_RSN	[Event*] [TemporalRelation] [Event]
Lymph node post neoadjuvant therapy response code	LN_NEOADJ_RESP_CD	[Event*] [TemporalRelation] [Event]

[a] TEO: Time Event Ontology.

Textbox 3. Inconsistencies between LongName and PreferredDefinition field.

```
<DataElement num="28726">
<PUBLICID>3250740</PUBLICID>
<LONGNAME>Prior Dental Restoration Post Core Use Yes or No Response</LONGNAME>
<PREFERREDDEFINITION>Earlier in time or order._Replacement or reconstruction of a lost tooth structure._Post;
occuring after._The center of an object; indispensable_Use; put into service; make work or employ (something)
for a particular purpose or for its inherent or natural purpose._A caDSR representation term that is used to
indicate a question with permissible values of yes/no</PREFERREDDEFINITION>
```

Time Event Ontology Limitations

During the annotation process, we discovered limitations with TEO that prevented complete representation of the temporal aspects within the CDEs. To be specific, events listed as an ordinal number of a series are poorly ontologically represented with TEO. This is because of the fact that TEO requires events to be related to each other via a TemporalRelationship, and simply designating the ordinality of the event is not representable with TEO patterns. Due to this requirement of a relationship between events, other temporal relationships found in the CDEs such as “most recent” or “prior” cannot directly be represented with TEO patterns. However, although these temporal aspects cannot be represented ontologically via TEO patterns, they can still be preserved in the RDF label, allowing a human reader or annotator to see these temporal aspects.

Additionally, irregular series of events are not well-represented by TEO. The TimePhase class was built to handle events that reoccur at a regular interval. Although many events within caDSR and a clinical setting in general reoccur at regular intervals, there are a few CDEs that involve events that reoccur irregularly. For example, given caDSR stores many CDEs that are related to cancer, the recurrence of cancers in a patient can be quite sporadic, making it difficult for us to represent the CDE with TEO. Although the current version of TEO cannot adequately represent these temporal aspects, it is anticipated that a future version of TEO will be able to address these shortcomings.

Conclusion and Future Direction

As stated before, we are facing an increasing volume of data within the EHR. To keep up with this exponential growth of data, a machine-readable annotation is necessary. The underlying OWL-based representation of TEO allows for the leveraging of many reasoning tools on the Semantic Web. With respect to the CDEs within caDSR, steps have already been taken toward standard representation of CDEs. However, aforementioned shortcomings of the current representation open the door for improvements. TEO provides a valid solution to improving the representation of the temporal components of the CDEs. Although it is not perfect quite yet, given its inability to represent certain temporal aspects, it improves upon the current standard representation of the temporal dimension. We hope to improve upon TEO to allow it to more completely represent the temporal dimension. Additionally, querying of these TEO patterns using SPARQL Protocol and RDF Query Language , Semantic Web Rule Language, and a TEO querier for TEO is a future goal. Ultimately, we hope to develop methods to automatically match the CDEs with patterns. Improvements on the CDE parser to improve the sensitivity and specificity will aid in assigning a TEO pattern to the time-relevant CDEs within caDSR. This can be done by refining the keywords list that is used to retrieve the CDEs, as well as incorporating standardized concept codes. In conjunction with these improvements, actual implementation into caDSR to represent the time components of CDEs is our ultimate goal. By improving upon the representation of the temporal components of these CDEs, we believe that research involving the temporal aspect of CDEs in caDSR will become more efficient.

Acknowledgments

Research was partially supported by the National Library Of Medicine of the National Institutes of Health under Award Number R01 LM011829, the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number R01 AI130460, the National Cancer Institute of the National Institutes of Health under Award Number U01 CA180940, and the Cancer Prevention Research Institute of Texas Training Grant #RP160015.

Conflicts of Interest

None declared.

References

1. Yang CC, Veltri P. Intelligent healthcare informatics in big data era. Artif Intell Med 2015 Oct;65(2):75-77. [doi: 10.1016/j.artmed.2015.08.002] [Medline: 26306669]

2. insights.datamark. 2013. Unstructured Data in Electronic Health Record (EHR) Systems: Challenges and Solutions URL: http://insights.datamark.net/white-papers/unstructured-data-in-electronic-health-record-systems-challenges-and-solutions [accessed 2017-05-30] [WebCite Cache ID 6qqyBqFpS]
3. Perez-de-Viñaspre O, Oronoz M. SNOMED CT in a language isolate: an algorithm for a semiautomatic translation. BMC Med Inform Decis Mak 2015 Jun;15(Suppl 2):S5 [FREE Full text] [doi: 10.1186/1472-6947-15-S2-S5] [Medline: 26100112]
4. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003 Apr;49(4):624-633 [FREE Full text] [Medline: 12651816]
5. Hwang KH, Chung K, Chung M, Choi D. Review of semantically interoperable electronic health records for ubiquitous healthcare. Healthc Inform Res 2010 Mar;16(1):1-5 [FREE Full text] [doi: 10.4258/hir.2010.16.1.1] [Medline: 21818417]
6. Fickenscher KM. President's column: interoperability--the 30% solution: from dialog and rhetoric to reality. J Am Med Inform Assoc 2013 May 01;20(3):593-594 [FREE Full text] [doi: 10.1136/amiajnl-2013-001768] [Medline: 23579423]
7. Smith PC, Araya-Guerra R, Bublitz C, Parnes B, Dickinson LM, Van VR, et al. Missing clinical information during primary care visits. J Am Med Assoc 2005 Feb 02;293(5):565-571. [doi: 10.1001/jama.293.5.565] [Medline: 15687311]
8. Leslie M. The price of excess DNA. J Cell Biol 2007 Aug 27;178(6):893. [doi: 10.1083/jcb.1786rr5]
9. Shahar Y, Combi C. Timing is everything. Time-oriented clinical information systems. Artif Intell Med 1998 Feb;168(2):105-113 [FREE Full text] [Medline: 9499744]
10. Anagnostopoulos E, Batsakis S, Petrakis EG. CHRONOS: a reasoning engine for qualitative temporal information in OWL. Procedia Comput Sci 2013;22:70-77. [doi: 10.1016/j.procs.2013.09.082]
11. Ermolayev V, Keberle N, Matzke W. An ontology of environments, events, and happenings. 2008 Presented at: Annual IEEE International Computer Software and Applications; 28 July-1 August, 2008; Turku, Finland. [doi: 10.1109/COMPSAC.2008.141]
12. Khriyenko O, Terziyan V. A framework for context-sensitive metadata description. Int J Metadata Semant Ontol 2006;1(2):154. [doi: 10.1504/IJMSO.2006.011011]
13. Tao C, Wei W, Solbrig HR, Savova G, Chute CG. CNTRO: A semantic web ontology for temporal relation inferencing in clinical narratives. AMIA Annu Symp Proc 2010 Nov 13;2010:787-791 [FREE Full text] [Medline: 21347086]
14. SBMI. URL: https://sbmi.uth.edu/ontology/TimeEventOntology.owl [accessed 2017-05-30] [WebCite Cache ID 6qqzaOQpk]
15. Freksa C. Temporal reasoning based on semi-intervals. Artif Intell 1992 Mar;54(1-2):199-227. [doi: 10.1016/0004-3702(92)90090-K]
16. HL7. URL: http://www.hl7.org/ [accessed 2017-09-07] [WebCite Cache ID 6tJ8xyoLa]
17. Beale T. Archetypes and the EHR. Stud Health Technol Inform 2003;96:238-244. [Medline: 15061551]
18. Openehr. OpenEHR Clinical Models Program URL: http://www.openehr.org/programs/clinicalmodels/ [accessed 2017-09-07] [WebCite Cache ID 6tJ9Idw2J]
19. Covitz PA, Hartel F, Schaefer C, De Coronado S, Fragoso G, Sahni H, et al. caCORE: a common infrastructure for cancer informatics. Bioinformatics 2003 Dec 12;19(18):2404-2412. [Medline: 14668224]
20. Silva JS, Ball MJ, Douglas JV. The Cancer Informatics Infrastructure (CII): an architecture for translating clinical research into patient care. Stud Health Technol Inform 2001;84(Pt 1):114-117. [Medline: 11604717]
21. National Cancer Institute. URL: https://cbiit.nci.nih.gov/ncip/biomedical-informatics-resources/interoperability-and-semantics/metadata-and-models [accessed 2017-05-30] [WebCite Cache ID 6qqyq8JFj]
22. metadata-standards. URL: http://metadata-standards.org/11179/ [accessed 2017-05-30] [WebCite Cache ID 6qqyrXxre]
23. Jiang G, Solbrig HR, Chute CG. Quality evaluation of cancer study Common Data Elements using the UMLS Semantic Network. J Biomed Inform 2011 Dec;44(Suppl 1):S78-S85 [FREE Full text] [doi: 10.1016/j.jbi.2011.08.001] [Medline: 21840422]
24. Jiang G, Solbrig HR, Chute CG. Quality evaluation of value sets from cancer study common data elements using the UMLS semantic groups. J Am Med Inform Assoc 2012 Jun;19(e1):e129-e136 [FREE Full text] [doi: 10.1136/amiajnl-2011-000739] [Medline: 22511016]
25. Warzel DB, Andonaydis C, McCurry B, Chilukuri R, Ishmukhamedov S, Covitz P. Common data element (CDE) management and deployment in clinical trials. AMIA Annu Symp Proc 2003:1048 [FREE Full text] [Medline: 14728551]
26. w3.org. OWL Web Ontology Language: Reference W3C Recommendation 10 February 2004 URL: http://www.w3.org/TR/owl-ref [WebCite Cache ID 6qr04nECy]
27. w3.org. World Wide Web Consortium URL: http://www.w3.org/1999/02/22-rdf-syntax-ns [WebCite Cache ID 6qqzbDPBu]
28. Allen JF. Maintaining knowledge about temporal intervals. Commun ACM 1983;26(11):832-843. [doi: 10.1145/182.358434]
29. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform 2014 Apr;48:193-204 [FREE Full text] [doi: 10.1016/j.jbi.2014.02.013] [Medline: 24582925]

Abbreviations

caDSR: Cancer Data Standards Repository
CDE: common data element
CNTRO: Clinical Narrative Temporal Relation Ontology

EHR: electronic health record
FN: false negative
NCI: National Cancer Institute
OWL: Web Ontology Language
RDF: Resource Description Framework
TEO: Time Event Ontology
TP: true positive

Edited by G Eysenbach; submitted 09.06.17; peer-reviewed by H Kim, L Cui, Z He; comments to author 29.07.17; revised version received 17.09.17; accepted 16.11.17; published 22.02.18.

Please cite as:
Chen HW, Du J, Song HY, Liu X, Jiang G, Tao C
Representation of Time-Relevant Common Data Elements in the Cancer Data Standards Repository: Statistical Evaluation of an Ontological Approach
JMIR Med Inform 2018;6(1):e7
URL: http://medinform.jmir.org/2018/1/e7/
doi:10.2196/medinform.8175
PMID:29472179

Original Paper

The Use of Communication Apps by Medical Staff in the Australian Health Care System: Survey Study on Prevalence and Use

Amanda Nikolic[1], MBBS, PGrad Dip Anat, MSurgSci; Nilmini Wickramasinghe[2], PhD, MBA; Damian Claydon-Platt[3], MBBS, PGrad Dip IT/IS, PhD; Vikram Balakrishnan[4], MBBS, BmMedSci; Philip Smart[1,4], MBBS, DMedSci

[1]General Surgery and Gastroenterology Clinical Institute, Epworth Healthcare, Richmond, Australia
[2]Health Information Management, Epworth Healthcare, Richmond, Australia
[3]Information Technology Department, Epworth Healthcare, Richmond, Australia
[4]Department of Surgery, Eastern Health, Box Hill, Australia

Corresponding Author:
Amanda Nikolic, MBBS, PGrad Dip Anat, MSurgSci
General Surgery and Gastroenterology Clinical Institute
Epworth Healthcare
89 Bridge Road
Richmond, 3121
Australia
Phone: 61 294266666
Email: nikolicamanda@gmail.com

Abstract

Background: The use of communication apps on mobile phones offers an efficient, unobtrusive, and portable mode of communication for medical staff. The potential enhancements in patient care and education appear significant, with clinical details able to be shared quickly within multidisciplinary teams, supporting rapid integration of disparate information, and more efficient patient care. However, sharing patient data in this way also raises legal and ethical issues. No data is currently available demonstrating how widespread the use of these apps are, doctor's attitudes towards them, or what guides clinician choice of app.

Objective: The objective of this study was to quantify and qualify the use of communication apps among medical staff in clinical situations, their role in patient care, and knowledge and attitudes towards safety, key benefits, potential disadvantages, and policy implications.

Methods: Medical staff in hospitals across Victoria (Australia) were invited to participate in an anonymous 33-question survey. The survey collected data on respondent's demographics, their use of communication apps in clinical settings, attitudes towards communication apps, perceptions of data "safety," and why one communication app was chosen over others.

Results: Communication apps in Victorian hospitals are in widespread use from students to consultants, with WhatsApp being the primary app used. The median number of messages shared per day was 12, encompassing a range of patient information. All respondents viewed these apps positively in quickly communicating patient information in a clinical setting; however, all had concerns about the privacy implications arising from sharing patient information in this way. In total, 67% (60/90) considered patient data "moderately safe" on these apps, and 50% (46/90) were concerned the use of these apps was inconsistent with current legislation and policy. Apps were more likely to be used if they were fast, easy to use, had an easy login process, and were already in widespread use.

Conclusions: Communication app use by medical personnel in Victorian hospitals is pervasive. These apps contribute to enhanced communication between medical staff, but their use raises compliance issues, most notably with Australian privacy legislation. Development of privacy-compliant apps such as MedX needs to prioritize a user-friendly interface and market the product as a privacy-compliant comparator to apps previously adapted to health care settings.

(JMIR Med Inform 2018;6(1):e9) doi:10.2196/medinform.9526

KEYWORDS

mobile phone; information science; communications media; privacy; interdisciplinary communication; hospital communication systems; communication

Introduction

Due to increased availability, affordability, and functionality, the use of mobile phones to communicate and enhance clinical practice within Australian hospitals is widespread [1-3]. Thousands of apps developed by third parties are available for use on mobile phones to aid in clinical decision making, monitoring of patients, medical education or information, communication, and more [4,5].

Using mobile phones for communication is possible through multiple means including texting (short message service [SMS]), voice and video calling, conferencing email, multimedia messaging, and communication apps such as WhatsApp and Viber [4].

Effective and efficient communication is key to safe and high quality patient care. In hospitals, challenges to communication include large multidisciplinary teams with complex hierarchies guiding patient care, a proliferation of clinical information that is often time critical, and the necessity of staff travel within hospitals and between health care sites. Traditional communication platforms such as paging may be unreliable, and 2-way communication is difficult. The use of communication apps on mobile phones to communicate with colleagues is fast, efficient, portable, and convenient [6-8]. These apps are often free, easily available, and in widespread use. They facilitate rapid communication within teams through conversation and closed-group features, overview and increased involvement by senior clinicians, an enhanced patient handover process, easy communication of patient results, and rapid changes to patient management plans [5,7,8]. With multiple health care staff caring for patients, enhanced communication supports greater efficiency [8]. Therefore, communication app use offers an attractive mode of communication for health care professionals.

The advancement of technology allowing communication on portable devices brings with it a range of legal and ethical issues. In Australia, federal, state, and territory privacy laws regulate the handling of personal information. Consent must be obtained for the use of such information, which can only be used for the purpose consented to. An obligation arises upon entities handling this information to ensure compliance with prescribed privacy principles. For example, security of the information must be catered for to protect it from any unauthorized use or disclosure [7]. Most of the mobile phone apps currently in general usage within the Victorian health care system to communicate clinical information do not comply with these regulations. Consent is often not obtained, data may be accessed from the host device if it is lost or hacked, the data may be stored on an insecure server, which is often overseas or backed up overseas [9]. In Australia, certain apps have been developed to comply with Australian privacy regulations, such as MedX and MyBeepr.

A small number of studies have demonstrated the benefits of communication apps when used as an intervention in clinical practice [4-6]. However, no data exists on whether these apps are being used by medical staff. It is also unclear why medical staff use one app over another. This study aimed to quantify and qualify the use of communication apps among medical staff in clinical situations, their role in patient care, and to elaborate on issues relating to safety, key benefits, potential disadvantages, and policy implications.

Methods

Recruitment

Medical staff across Victorian hospitals from September to October 2017 were sent an email, social media post, or were personally approached to complete an anonymous 33-item online survey administered by SurveyMonkey (Multimedia Appendix 1). The survey was trialed on 2 medical staff at 1 hospital—their responses were not included in the results. “Logic” was used at some questions to guide respondents to further questions based on their previous response. The number of medical staff reached was unable to be calculated given the nature of social media posts and emails that were sent generally to medical staff across multiple departments. The lead researcher and 2 other researchers disseminated the survey. A brief description of the survey, details about anonymity, and intention to publish de-identified data was outlined in the first page of the survey. Consent was obtained with continued participation in the survey beyond the first page signaling consent.

Data Collection

The survey collected data on the following: (1) respondents demographics, (2) use of communication apps in clinical practice, (3) amount and type of communication app use, (4) attitudes towards communication apps, (5) perceived benefits and disadvantages, (6) views on data “safety”, and (7) why one communication app was chosen over others.

Results

Demographics

In total, 118 responses were received, of which 88 (74.6%, 88/118) were complete responses. Of the respondents, 67.8% (80/118) were doctors, with 32.2% (38/118) medical students. The majority of respondents worked in the surgical field (Table 1). Most people (72.2%, 83/115) owned an iPhone.

Communication App Use

Most participants used WhatsApp (85.0%, 96/113) as their main app for communicating clinical information. Most respondents used the app daily (78.4%, 80/102), with the median number of messages sent being 12 per day. A range of patient information was shared on communication apps, both with individual colleagues as well as within clinical teams (Table 2).

Knowledge and Perceptions of Safety and Privacy

Most participants (67%, 60/90) thought communicating patient information on apps was only moderately safe, with 21% (19/90) considering the information safe. Of the participants, 50% (46/90) felt they may “get into trouble” by sharing patient information on apps and 76% (64/86) did not know that if communicated data was stored on overseas servers it breached Australian privacy legislation. Only 45% (5/11) of participants were aware of a hospital policy regarding the use of apps. Most participants were aware that patient consent was required to share the information. The majority of participants (94%, 85/90) considered that consent was required prior to "taking a photograph of a wound to send to a plastic's registrar". However, only half considered this consent needed to be documented in the patient notes or entered onto a hospital consent form.

Perceptions on the Benefits and Disadvantages of Using Apps for Clinical Purposes

All respondents stated benefits relating to the use of apps in clinical practice. These included prompt communication, reduction in interruptions, portability, easier access to senior clinicians and other team members who were only intermittently available (ie, when registrars or consultants are scrubbed in theater), and enhanced communication relating to patient progress, results, and education. Other benefits noted included the ability to create "groups" correlating to clinical teams, being able to view who has seen comments, and the ability to mute conversations when not at work. Most respondents (78%, 71/90) also noted disadvantages relating to the use of apps. The main disadvantage noted by 94% (64/68) of participants was the potential risk to patient confidentiality. Less commonly viewed disadvantages included 1-sided communication, missing aspects of a conversation thread, and the expectation that all members of the group were equally informed about patient information shared on the app, even if members of the group were not present in the workplace. In addition, there was a concern that use of apps needed to be commensurate with the clinical situation, with face to-face or voice interaction required for more time critical situations.

Preferences

Respondents were more likely to use an app that was free, easy to login, in wide usage with other colleagues, enabled the establishment of discrete "groups," permitted the sharing of multiple data formats (ie, text, images, tables, video), and complied with privacy requirements. Only 13% (11/85) of participants were aware of the communication app MedX; however, none used it as their main communication app. Most frequently cited reasons for this included (1) a difficult login process and user interface; (2) widespread use of WhatsApp as an alternative; (3) and the perception that messages degrade on MedX.

The key findings of the surveys are shown in Textbox 1.

Table 1. Demographics of respondents.

Demographics	n (%)
Position (N=115)	
Medical student	38 (32.2)
Intern	19 (16.1)
Resident	17 (14.4)
Registrar	28 (23.7)
Fellow	8 (6.8)
Consultant	8 (6.8)
Department (N=71)	
Medicine	39 (35)
Surgery	51 (45)
Pediatrics	1 (1)
Obstetrics and Gynecology	2 (2)
Radiology	3 (3)
Emergency	9 (12)
Phone (N=115)	
iPhone	83 (72.2)
Android	32 (27.8)

Table 2. Communication app use, types of information shared, and group communications.

Characteristics	n (%)
Main app used for clinical purposes (N=105)	
WhatsApp	89 (84.7)
Viber	0
MedX	0
Slack	1 (0.9)
Other	15 (14.3)
Quantity/use of app for clinical purposes (N=100)	
Daily	78 (78)
<Daily	22 (22)
Type of information sent via communication app	
Patient management details	78 (80)
Patient results	77 (79)
Details that facilitate clinical handover	63 (65)
Questions to colleagues about management	69 (79)
Answers to colleagues about management	65 (75)
Pictures of bradma labels	49 (56)
Patient name and unit record numbers (unique patient identifier)	62 (71)
Pathology results	66 (76)
Admission notes	38 (44)
Imaging reports	54 (61)
Pathology reports	51 (44)
Microbiology reports	35 (50)
Radiological pictures	49 (56)
Interventional reports[a]	45 (52)
Electrocardiogram images	33 (38)
Participants belong to a communication or team group (N=96)	87 (91)

[a] Examples of interventional reports include reports and operation notes.

Textbox 1. Key findings of the surveys.

- All medical staff owned a mobile phone.
- The majority of medical staff used apps for clinical purposes on a daily basis.
- WhatsApp was the most commonly-used app.
- All staff shared patient data.
- All staff considered apps to enhance clinical practice and communication to improve patient care.
- Confusion existed regarding what consent was required when sharing patient information.
- Most medical staff were concerned about the privacy implications of apps for clinical purposes.
- Staff were more likely to use an app if it was in widespread use, and was free, reliable, and easy to use.

Discussion

Communication App Use and Benefits

The use of apps for clinical purposes by medical staff is widespread, with most using them to enhance communication with colleagues and share clinical information to enhance patient outcomes. All participants saw the benefit of using apps in clinical situations, considering them to be efficient, portable, and a less obtrusive means of facilitating patient handover, communicating within teams, integrating patient information, and optimizing patient management plans. It is clear the role of these apps in Victorian clinical practice is well-established and offers benefits over more traditional forms of communication (ie, paging systems, voice calling, and face to face meetings). In particular, they are suited to the unique challenges faced by health care teams including large multidisciplinary teams where senior staff may only be intermittently available, responding to time critical issues, and optimizing team interaction in geographically-diverse health care centers.

Knowledge and Perceptions of Safety and Privacy

Despite the widespread use of apps, there was confusion about privacy implications and consent. For example, most participants knew that consent was required when taking a photograph on a mobile phone to share with a colleague; however, only half considered that documenting patient consent in their notes was also required. Added complexity exists when considering ongoing discussion and sharing of a wide range of patient information to facilitate daily patient care by medical staff on communication apps. Privacy legislation in Australia states that patient information can only be used for the purpose to which it was collected and consented for by the patient [7]. Medico-legal providers recommend documenting consent in patient notes when sharing images on mobile phones [10]. This raises the question of whether patient consent should also be obtained prior to sharing their information on communication apps. Given the number of patients often admitted under clinical teams, this requirement may serve to reduce the speed, efficacy, and attractiveness of these apps. Another option may be to require consent on admission by patients for the specified and agreed purpose of discussing their care via these apps.

Respondents associated subjective risks with sharing patient data on communication apps, with half suspecting they may get “in trouble” for doing so, and 97% (64/68) acknowledging privacy concerns. However, the majority of medical staff continue to use these apps on a daily basis despite recognizing potential non-compliance with privacy laws. This may be considered a case of convenience and familiarity trumping privacy, or alternatively clinicians becoming reliant on their mobile phones and these types of communication technologies for patient-related communication. Medical staff were also unaware that these apps store data, including identifiable patient information, on overseas servers which contravenes Australian privacy legislation. Although there have not been any legal cases against medical staff regarding the use of these apps, given their non-compliance with data safety legislation, a case may be made against medical staff if data shared on these apps was compromised. The Medical Board of Australia does not currently have guidelines that address the use of communication apps. However, in 2015 they highlighted the risks related to communication with patients via electronic messaging and recommended that medical practitioners be aware of privacy legislation [11].

Simple steps may be taken by medical staff to decrease the risk to patient data safety when using mobile phones for communication. These include obtaining consent from the patient, having pin numbers on devices to prevent unauthorized access, and deleting information once it is not longer required. The Australian Medical Association (AMA) policy on the use of clinical photography on mobile phones is a good practical guide that can be used in the clinical setting [12]. Hospitals may also play a role in increasing awareness relating to privacy by establishing their own policies.

The increasing development, use, and benefits of communication apps needs to be balanced against the risk to patient safety and confidentiality. Developing apps which comply with privacy legislation, protect patient data with encryption, and are resistant to cyber crime may facilitate the use of these apps without risking patient data security. Guidelines do not currently exist which address the use of communication apps in clinical practice. These need to be developed to guide medical staff on their safe use, at a pace which mirrors their adoption by clinical staff.

Perceptions of Disadvantages Relating to the Use of Mobile Phone Apps

Barriers and disadvantages relating to the use of apps were acknowledged by 80% (72/90) of participants with the most commonly-cited being the risk to patient confidentiality. Other disadvantages were less commonly noted, and did not appear to deter medical staff from using these apps.

Preferences

Respondents were more likely to use an app if it was free to access, already in widespread use, had an easy and reliable interface, and was easy to use. WhatsApp was the most frequently used app. Most staff were unaware of a purpose-built communication app called MedX, which complies with Australian privacy regulations. Medical staff that were aware of MedX did not use the app due to a difficult login process, user interface, and low use pattern. The widespread use of other apps, mainly WhatsApp, will likely render the introduction of these compliant apps problematic. A simpler login process and user interface needs to be developed, local policies prioritizing these apps, appropriate advertising, and even incentives may be required to shift established usage to privacy-compliant apps. Given the intersection of app usage with federal, state, and territory laws, this issue may also be worth considering by the Victorian Boards Ministerial Advisory Committee, and perhaps even the Australian Health Ministers’ Advisory Council (AHMAC) and COAG Health Council (CHC).

Limitations

The limitations of this study include the small sample population and simple survey framework.

Conclusions

The use of communication apps by medical personnel in Victorian hospitals is pervasive, with WhatsApp the most commonly used. These apps play a role in optimizing communication between medical staff to deliver better health outcomes for patients. The major disadvantage arising from the use of apps is the non-compliance of apps currently in widespread usage with Australian privacy legislation. However, this does not appear to limit their use despite the majority of medical staff acknowledging risks to patient privacy. Development of privacy-compliant apps such as MedX needs to prioritize those features that currently engage user interest in non-compliant apps. A coordinated effort is also required in a regulatory and policy sense to ensure the transition to privacy-compliant apps. This is an initiative worthy of consideration by the Victorian Boards Ministerial Advisory Committee, and perhaps even the AHMAC and CHC.

Acknowledgments

This study was supported by Epworth Research Institute Major Research Grant # 11.952.000.80982.

Conflicts of Interest

VB is a developer of the medical communication app MyBeepr.

Multimedia Appendix 1

The SurveyMonkey questionnaire used in this study.

[PDF File (Adobe PDF File), 247KB - medinform_v6i1e9_app1.pdf]

References

1. Mosa A, Yoo I, Sheets L. A systematic review of healthcare applications for smartphones. BMC Med Inform Decis Mak 2012 Jul 10;12:67 [FREE Full text] [doi: 10.1186/1472-6947-12-67] [Medline: 22781312]
2. O'Connor P, Byrne D, Butt M, Offiah G, Lydon S, Mc IK, et al. Interns and their smartphones: use for clinical practice. Postgrad Med J 2014 Feb;90(1060):75-79. [doi: 10.1136/postgradmedj-2013-131930] [Medline: 24243966]
3. Charani E, Castro-Sánchez E, Moore L, Holmes A. Do smartphone applications in healthcare require a governance and legal framework? It depends on the application!. BMC Med 2014 Feb 14;12:29 [FREE Full text] [doi: 10.1186/1741-7015-12-29] [Medline: 24524344]
4. Ventola CL. Mobile devices and apps for health care professionals: uses and benefits. P&T 2014 May;39(5):356-364 [FREE Full text] [Medline: 24883008]
5. Ozdalga E, Ozdalga A, Ahuja N. The smartphone in medicine: a review of current and potential use among physicians and students. J Med Internet Res 2012 Sep 27;14(5):e128 [FREE Full text] [doi: 10.2196/jmir.1994] [Medline: 23017375]
6. Patel B, Johnston M, Cookson N, King D, Arora S, Darzi A. Interprofessional communication of clinicians using a mobile phone app: a randomized crossover trial using simulated patients. J Med Internet Res 2016 Apr 06;18(4):e79 [FREE Full text] [doi: 10.2196/jmir.4854] [Medline: 27052694]
7. Johnston MJ, King D, Arora S, Behar N, Athanasiou T, Sevdalis N, et al. Smartphones let surgeons know WhatsApp: an analysis of communication in emergency surgical teams. Am J Surg 2015 Jan;209(1):45-51. [doi: 10.1016/j.amjsurg.2014.08.030] [Medline: 25454952]
8. Khanna V, Sambandam S, Gul A, Mounasamy V. "WhatsApp"ening in orthopedic care: a concise report from a 300-bedded tertiary care teaching center. Eur J Orthop Surg Traumatol 2015 Jul;25(5):821-826. [doi: 10.1007/s00590-015-1600-y] [Medline: 25633127]
9. Drake TM, Claireaux HA, Khatri C, Chapman SJ. WhatsApp with patient data transmitted via instant messaging? Am J Surg 2016 Jan;211(1):300-301. [doi: 10.1016/j.amjsurg.2015.04.004] [Medline: 26092444]
10. Kunde L, McMeniman E, Parker M. Clinical photography in dermatology: ethical and medico-legal considerations in the age of digital and smartphone technology. Australas J Dermatol 2013 Aug;54(3):192-197. [doi: 10.1111/ajd.12063] [Medline: 23713892]
11. Medical Board of Australia. 2015 Feb. Update: Medical Board of Australia Februrary 2015 URL: http://www.medicalboard.gov.au/News/Newsletters/February-2015.aspx#latest [accessed 2017-12-17] [WebCite Cache ID 6vnSqYgFJ]
12. Australian Medical Association. 2014. Clinical images and the use of personal mobile devices URL: https://ama.com.au/article/clinical-images-and-use-personal-mobile-devices [accessed 2017-10-26] [WebCite Cache ID 6uVsCJydW]

Abbreviations

AHMAC: Australian Health Ministers' Advisory Council
CHC: COAG Health Council

Edited by G Eysenbach; submitted 30.11.17; peer-reviewed by V Khanna, M Abdelhamid, R Pankomera; comments to author 16.12.17; revised version received 18.12.17; accepted 23.12.17; published 09.02.18.

Please cite as:
Nikolic A, Wickramasinghe N, Claydon-Platt D, Balakrishnan V, Smart P
The Use of Communication Apps by Medical Staff in the Australian Health Care System: Survey Study on Prevalence and Use
JMIR Med Inform 2018;6(1):e9
URL: http://medinform.jmir.org/2018/1/e9/
doi:10.2196/medinform.9526
PMID:29426813

Original Paper

Patient and Health System Experience With Implementation of an Enterprise-Wide Telehealth Scheduled Video Visit Program: Mixed-Methods Study

Rhea E Powell[1], MPH, MD; Danica Stone[1], BA; Judd E Hollander[1], MD

Thomas Jefferson University, Philadelphia, PA, United States

Corresponding Author:
Judd E Hollander, MD
Thomas Jefferson University
1025 Walnut Street Suite 300
Philadelphia, PA, 19107
United States
Phone: 1 2155035591
Email: judd.hollander@jefferson.edu

Abstract

Background: Real-time video visits are increasingly used to provide care in a number of settings because they increase access and convenience of care, yet there are few reports of health system experiences.

Objective: The objective of this study is to report health system and patient experiences with implementation of a telehealth scheduled video visit program across a health system.

Methods: This is a mixed methods study including (1) a retrospective descriptive report of implementation of a telehealth scheduled visit program at one large urban academic-affiliated health system and (2) a survey of patients who participated in scheduled telehealth visits. Health system and patient-reported survey measures were aligned with the National Quality Forum telehealth measure reporting domains of access, experience, and effectiveness of care.

Results: This study describes implementation of a scheduled synchronous video visit program over an 18-month period. A total of 3018 scheduled video visits were completed across multiple clinical departments. Patient experiences were captured in surveys of 764 patients who participated in telehealth visits. Among survey respondents, 91.6% (728/795) reported satisfaction with the scheduled visits and 82.7% (628/759) reported perceived quality similar to an in-person visit. A total of 86.0% (652/758) responded that use of the scheduled video visit made it easier to get care. Nearly half (46.7%, 346/740) of patients estimated saving 1 to 3 hours and 40.8% (302/740) reported saving more than 3 hours of time. The net promoter score, a measure of patient satisfaction, was very high at 52.

Conclusions: A large urban multihospital health system implemented an enterprise-wide scheduled telehealth video visit program across a range of clinical specialties with a positive patient experience. Patients found use of scheduled video visits made it easier to get care and the majority perceived time saved, suggesting that use of telehealth for scheduled visits can improve potential access to care across a range of clinical scenarios with favorable patient experiences.

(JMIR Med Inform 2018;6(1):e10) doi:10.2196/medinform.8479

KEYWORDS

telemedicine; video visit; primary care; specialty; patient satisfaction

Introduction

Telehealth video visits, or real-time remote face-to-face visits between patients and providers, have been implemented in a number of settings in recent decades. Video visits have a well-established track record of use in rural and health shortage service areas, where the availability of providers may be limited [1,2]. Applications of various forms of telehealth including video visits have been studied in a number of settings, including behavioral health care [3], dermatology [4], genetic counseling [5], rheumatology [6], and pain management [7]. Real-time remote video visits have been shown to be an acceptable alternative to patients and providers in a number of settings and have the potential to reduce costs [8,9].

Although telehealth video visit use for scheduled routine visits are increasingly implemented in various health care settings, there are few published reports of health system experiences implementing telehealth programs that include scheduled video visits. In the United States, some information is available about system-wide implementation of clinical video telemedicine programs from the Veteran's Administration (VA) [8]. The VA has reported experiences with cost savings related to widespread video visit use [10], which can inform other health systems seeking to implement system-wide video visit programs. A number of cost and reimbursement factors are unique to the VA setting, however, and may not be shared by other US health systems looking to adopt video visit programs. Other experiences can be gleaned from the international community. One health system report from a tertiary hospital in Australia described processes and outcomes of introducing of a centralized coordination for telehealth service [11], with a resulting increase in availability of telehealth services. This work focuses primarily on health system factors, and does not include patient experiences of implementation. An understanding of health system and patient experiences with implementation of telehealth visits is needed to improve design and delivery of telehealth for scheduled visits.

We report experiences of one large urban health system in implementing an enterprise-wide scheduled video visit program across various disciplines and specialties, with a focus on the impact on access, experience, and effectiveness of care. These three domains of care represent three of the four domains that inform the National Quality Forum (NQF) telehealth measures framework [12].

Methods

Study Design and Setting

This is a mixed methods study evaluating the JeffConnect scheduled visit program at a large urban academic-affiliated health system, Jefferson Health (Jefferson), located in Philadelphia, PA, USA. The study includes a retrospective descriptive evaluation of implementation of the program and a survey of patients who participated in scheduled telehealth visits.

Selection of Survey Participants

Patients aged 18 years and older with an existing relationship with a Jefferson provider who was trained on telehealth use were eligible to participate in scheduled telehealth visits. Patients were informed of the option of a telehealth scheduled visit by their provider, an administrator, or learned about it through marketing notifications. All patients who participated in a telehealth visit were eligible to participate in the survey. Patients were contacted by email the second week following their scheduled visit to provide feedback via survey.

Program Description

Jefferson Health provides hospital-based and outpatient-based services to patients across its four hospital systems, including the academic medical center at Thomas Jefferson University. In 2015, Jefferson initiated JeffConnect, an enterprise-wide telehealth program that offers video visits with a Jefferson health care provider via Web or mobile app, allowing patients to follow up with providers virtually as an alternative to returning to the office in-person. Patients and providers schedule appointments the same way they would for in-person visits, and visits are performed via real-time face-to-face remote video.

The JeffConnect team comprises a program and project manager and five telehealth coordinators. The telehealth coordinators are trained to be responsible for coordinating clinical services and enhancing patient engagement via telehealth. Telehealth coordinators are not medical providers. They are college educated and have completed an American Telehealth Association-accredited telehealth facilitator certificate program [13]. Coordinators use videoconferencing technologies and scheduling software to coordinate and connect staff, patients, and providers in the manner effective to delivery of services, patient care, education, and training. Scheduled video visits were initially piloted with the institution's covered employees who are also patients, then offered widely to all established Jefferson patients in all specialties.

Training Program

During its initial implementation phase, the JeffConnect team conducted more than 50 two-hour in-person group education sessions training providers, schedulers, and staff. The training session covered topics including program description, legal and regulatory information relevant to providing care via telehealth, how to use the telehealth platform for conducting video visits, and value to patients. Presently, the in-person group telehealth classes are now individually offered virtually through webinar.

Scheduling Visits

Patients schedule video visits using the same processes that are used for scheduling in-person visits, either by calling a centralized health system scheduler, calling the office, or by requesting a telehealth visit online. Many patients were referred to schedule via telehealth by their provider, and appropriateness for telehealth visit was determined by provider for all visits. All patients at the health system receive an automated reminder phone call 2 days prior to the visit, which is in place for on-site and telehealth visits. Additionally, for telehealth visits, the patient receives a phone call the day before from a telehealth coordinator to review processes for log-on, check that any necessary steps such as app download and registration are completed, and test the connection.

Conducting Visits

On the day and time of the scheduled video visit, patients log on to their password-protected JeffConnect account using a mobile phone or tablet app, or via a laptop or desktop browser equipped with a webcam and microphone. Providers log on from their health system location using a tablet app or Web browser, also with a webcam and microphone. Visits include real-time video and audio. Providers have access to the electronic medical record to review the patient's prior records and document the visit.

Data Collected

Data were collected with a focus on the impact on access, experience, and effectiveness of care, because these three domains of care represent three of the four domains that inform the NQF telehealth measures framework [12].

Access

Health system measures of access to care via telehealth scheduled visit included number of providers trained to use telehealth for scheduled visits, the number of downloads and registrations of the app, and the number of completed visits. Metrics for each department were collected and reported monthly, indicating how many visits were completed and by which provider. Providers were categorized by specialty, including dermatology, emergency medicine, family medicine, medical subspecialties (allergy, cardiology, endocrinology, gastroenterology, hematology, infectious disease, nephrology, oncology, pulmonology, and rheumatology), neurology, obstetrics and gynecology, psychiatry, radiation oncology, radiology, rehabilitation medicine, and surgical and related subspecialties (anesthesia, general surgery, neurosurgery, oral maxillofacial surgery, otolaryngology, preadmission testing, and urology).

Patient-reported measures of access to care included reported ease of use and impact on the ability to receive care when and where needed (both on a five-point Likert scale), as well as patient estimates of time saved through use of the telehealth visit.

Experience

Patient experience was assessed using a series of questions including overall patient satisfaction, reasons for dissatisfaction (if noted), if the patient would use JeffConnect for a scheduled telehealth visit again, and if the patient would recommend it to a family member or friend. Experience was also assessed through calculation of net promoter score, a measure of willingness to recommend to others.

Effectiveness

Effectiveness of care was assessed using health system data including qualitative responses from division directors. Patient responses relevant to effectiveness of care included patients' perspectives of whether level of care received via telehealth was equal to level of care received via in-person visit, and whether patients had adequate time with the provider, assessed on five-point Likert scale.

Data Analysis

Data are presented descriptively as absolute numbers and percent frequency of occurrence.

Results

Access

The Jefferson telehealth program trained 746 providers, including physicians and advanced practice providers, to perform scheduled video visits. A summary of total completed visits between January 2015 and December 2016 are presented in Table 1.

All the clinical care departments that provide outpatient care had physicians capable of delivering telehealth. There were 32,234 registrations and downloads of the JeffConnect app, and 3018 scheduled outpatient video visits were completed during the 18-month implementation period.

Of the 3018 completed video visits, 764 patients responded to the after-visit survey. Patient survey responses are summarized in Table 2. Most patients (84.8%, 646/762) surveyed had no prior experience with telehealth video visits. The majority (86.0%, 652/758) agreed or strongly agreed that JeffConnect made it easier to get care.

Experience

Among survey participants, 91.3% (728/797) reported satisfaction with their scheduled telehealth video visit. Among the 67 participants who reported they were not satisfied, 57 of those cited technical issues and five reported they did not like interacting on video. A total of 86.7% (656/757) agreed or strongly agreed it was easy to use and 90.9% (686/755) would use it again. The net promoter score, a reflection of patient willingness to recommend scheduled visits, was 52, consistent with high likelihood of recommending the service.

Table 1. Scheduled video visits completed from January 2015 to December 2016 by department.

Department	Visits by physicians, n	Visits by advanced practice providers, n
Dermatology	32	3
Emergency medicine	88	0
Family medicine	32	0
Medical subspecialties	734	233
Neurology	10	0
Obstetrics & gynecology	40	9
Psychiatry	240	40
Radiation oncology	55	5
Radiology	60	0
Rehabilitation medicine	50	0
Surgical subspecialties	908	479
Total	2249	769

Table 2. Scheduled visit patient survey responses (N=764).

Question and response	n (%)
How did you hear about JeffConnect?	
Email	42 (5.2)
Postal mail	1 (0.1)
Friend or family	24 (3.0)
Health care provider	554 (69.7)
Jefferson website	39 (4.9)
Print advertisement	5 (0.6)
Online advertisement	1 (0.1)
Other	129 (16.2)
Have you ever had a telehealth video visit before this visit?	
Yes	116 (15.2)
No	646 (84.8)
Do you use social media?	
Yes	63 (71.6)
No	25 (28.4)
Have you recommended JeffConnect to your friends or family?	
Yes	307 (43.6)
No	397 (56.4)
Overall, were you satisfied with your most recent visit?	
Yes	728 (91.6)
No	67 (8.4)
What is the reason you were unsatisfied with your visit (check all that apply)[a]	
I experienced technical issues	53 (83.1)
I didn’t like interacting on video	5 (7.9)
I was not happy with the physician	0 (0.0)
Other	27 (42.8)
How much time do you think JeffConnect saved you?	
None	31 (4.1)
Less than 1 hour	61 (10.7)
1-3 hours	346 (45.5)
More than 3 hours	302 (39.7)

[a] Respondents had the option to identify more than one response.

Effectiveness

Use cases for scheduled visits varied by department. Many of the clinical departments used scheduled video visits for routine follow-up to assess an ongoing episode of care, chronic condition management, medication updates, and to engage families in outpatient care. Anesthesiology used scheduled video visits for some components of preadmission testing before surgery and for postoperative pain management. Surgical specialties (urology, otolaryngology, and oral maxillofacial surgery) employed scheduled telehealth visits for postop follow-up. Rehabilitation medicine used telehealth scheduled video visits for transitions of care visits after hospital discharge, wound care visits, prosthesis monitoring, and physical therapy follow-up. Obstetrics and gynecology use cases included family planning visits.

Among patient responses with regard to effectiveness, 91.0% (691/759) reported having had enough time with the provider and 82.7% (628/759) perceived the same level of care as in in-person visits. More than 87.6% (648/740) perceived at least 1 hour of time saved by converting outpatient visit to a scheduled telehealth visit, and nearly 40.8% (302/740) perceived more than 3 hours of time saved.

Discussion

This study reports the initial implementation of a scheduled video visit program at one large academic health system, including completion of 3018 scheduled telehealth visits across all clinical departments in the enterprise. Our findings demonstrate that use of telehealth for scheduled visits increases potential and realized access to health care across a range of clinical scenarios, and is associated with favorable patient experiences.

The NQF report establishing a framework for measuring quality of care provided through telehealth focuses on access, experience, effectiveness, and financial impact of care [12]. Access to care includes access for patients and families, access for the care team, and access to information. The Anderson and Aday [14] conceptual model for understanding access to care considers potential (resources that allow patients to seek care) and realized access (actual use of care). This study provides input on the ability of a large health system to increase potential access to care by enabling providers in every clinical department to potentially provide care and facilitating availability of telehealth scheduled visits to patients who have registered and downloaded the app. The study also demonstrates the impact on realized access to care across a range of clinical departments through the completed visits.

Improving access to health care has been touted as a primary value added by telehealth in health care [15-17], and policy recommendations for improving access to care include integrating telehealth into care [18]. We add to the existing body of literature around improved access through telehealth with evidence of a large health system's experience implementing scheduled video visits into routine care of existing patients. Measuring access to care under the NQF framework for telehealth will importantly include access for patients and family, and will also include access for the care team and access to information (electronic health records and health information). Although this study does not directly evaluate access for the care team or access to information, we note that clinicians providing care via scheduled video visit have continuous access to the electronic medical record while engaging in the video visit.

This work also builds on existing literature suggesting that patient's report favorable experience with telehealth video visits [7,19,20]. We add to this work, and add to it with the use of the net promoter score, to assess patient satisfaction with telehealth services. The net promoter score is a metric to estimate how likely an individual is to recommend a service. Initially used in marketing [21,22], and more recently adapted for use in health care [23], the net promoter score allows for categorization of survey respondents either as a "promoter," "passive," or "detractor." The score is calculated by the percentage of promoters minus the percentage of detractors, and ranges from –100 to 100. A positive score of 52, such as we found among our patients, reflects high likelihood of recommending to friends and family.

There is a broad and growing body of literature surrounding the impact of telehealth video visits on access and experience of care, but the effectiveness of telehealth scheduled video visits for routine care and the financing of scheduled video visits are incompletely understood. The clinical use cases reported were wide-ranging and varied significantly by department, making a uniform assessment of quality of care provided challenging. These findings demonstrate that the majority of patients surveyed across a heterogeneous group of clinical scenarios felt they had received the same level of quality as they would have during on-site in-person visits.

Limitations

This study reports on experiences with initial implementation of an enterprise-wide scheduled video visit program. Patients who participated in scheduled visits self-selected to use video visits to connect with their providers. Although very few of these patients had any prior experience with video visits for health care, they were nevertheless the early adopters of this application of telehealth at our health system. Their perceptions may not be generalizable to other populations who did not choose to use telehealth.

Additionally, the perspective of providers and staff are not captured by these data. Provider and staff engagement are essential to the success of a system-wide program. Implementation of comprehensive scheduled video visit programs require communicating the value of telehealth to providers, compensating accordingly, keeping information technology applications and workflows simple, recognizing the workload that providers handle, and investing in a culture where providers are trained and rewarded for providing high-quality care that includes telehealth visits [24]. Future work should address the experience of the care team in widespread implementation of scheduled video visits.

Finally, we were unable to evaluate the financing of a scheduled video visit program with this work. During our initial implementation period, telehealth scheduled visits were not compensated by most payers and patients were not billed for this uncovered benefit; as such, evaluating the financial implications was not possible. Health systems considering system-wide implementation of telehealth program should identify motivations and barriers of all stakeholders for telehealth scheduled visits among patients, providers, administrators, and payers [25], and they will need information on how a scheduled visit program impacts care access, experience, effectiveness, and financing.

Conclusions

Health care delivery is in a state of flux, shifting from traditional in-person, visit-based, fee-for-service models toward care delivery that is patient-centered, efficient, and lower cost. Effective use of telehealth video visits can facilitate meeting these goals, but requires broad adoption and integration into clinical care. Our experiences implementing an enterprise-wide telehealth program at one large urban multihospital health system demonstrate the promise that scheduled telehealth video visits hold for improving access, supporting a positive patient experience and providing effective care.

Conflicts of Interest

None declared.

References

1. Marcin JP, Ellis J, Mawis R, Nagrampa E, Nesbitt TS, Dimand RJ. Using telemedicine to provide pediatric subspecialty care to children with special health care needs in an underserved rural community. Pediatrics 2004 Jan;113(1 Pt 1):1-6. [Medline: 14702439]
2. Menon P, Stapleton R, McVeigh U, Rabinowitz T. Telemedicine as a tool to provide family conferences and palliative care consultations in critically ill patients at rural health care institutions: a pilot study. Am J Hosp Palliat Care 2015 Jun;32(4):448-453. [doi: 10.1177/1049909114537110] [Medline: 24871344]
3. Bashshur RL, Shannon GW, Bashshur N, Yellowlees PM. The empirical evidence for telemedicine interventions in mental disorders. Telemed J E Health 2015 Dec:1 [FREE Full text] [doi: 10.1089/tmj.2015.0206]
4. Bashshur R, Shannon G, Tejasvi T, Kvedar J, Gates M. The empirical foundations of teledermatology: a review of the research evidence. Telemed J E Health 2015 Dec;21(12):953-979 [FREE Full text] [doi: 10.1089/tmj.2015.0146] [Medline: 26394022]
5. Buchanan AH, Datta SK, Skinner CS, Hollowell GP, Beresford HF, Freeland T, et al. Randomized trial of telegenetics vs in-person cancer genetic counseling: cost, patient satisfaction and attendance. J Genet Couns 2015 Dec;24(6):961-970 [FREE Full text] [doi: 10.1007/s10897-015-9836-6] [Medline: 25833335]
6. Piga M, Cangemi I, Mathieu A, Cauli A. Telemedicine for patients with rheumatic diseases: systematic review and proposal for research agenda. Semin Arthritis Rheum 2017 Mar:1 [FREE Full text] [doi: 10.1016/j.semarthrit.2017.03.014]
7. Hanna GM, Fishman I, Edwards DA, Shen S, Kram C, Liu X, et al. Development and patient satisfaction of a new telemedicine service for pain management at Massachusetts General Hospital to the island of Martha's Vineyard. Pain Med 2016 Sep;17(9):1658-1663. [doi: 10.1093/pm/pnw069] [Medline: 27121891]
8. Wennergren J, Munshi I, Fajardo A, George V. Implementation of clinical video telemedicine (CVT) within a VA medical center is cost effective and well received by veterans. IJCM 2014;5(12):711-716. [doi: 10.4236/ijcm.2014.512097]
9. Totten A, Womack DM, Eden KB, McDonagh MS, Griffin JC, Grusing S, et al. Telehealth: mapping the evidence for patient outcomes from systematic reviews. Report No: 16-EHC034-EF. In: AHRQ Comparative Effectiveness Technical Briefs. Rockville, MD: Agency for Healthcare Research and Quality; Jun 2016.
10. Russo J, McCool R, Davies L. VA Telemedicine: an analysis of cost and time savings. Telemed J E Health 2016 Mar;22(3):209-215. [doi: 10.1089/tmj.2015.0055] [Medline: 26305666]
11. Martin-Khan M, Fatehi F, Kezilas M, Lucas K, Gray LC, Smith AC. Establishing a centralised telehealth service increases telehealth activity at a tertiary hospital. BMC Health Serv Res 2015 Dec 03;15:534 [FREE Full text] [doi: 10.1186/s12913-015-1180-x] [Medline: 26630965]
12. National Quality Forum. Telehealth framework to support measure development 2016-2017 URL: http://www.qualityforum.org/Telehealth_2016-2017.aspx [accessed 2018-01-31] [WebCite Cache ID 6wt2s0F6w]
13. Thomas Jefferson University. Institute of Emerging Health Professions Telehealth Facilitator Certificate URL: http://www.jefferson.edu/university/emerging-health-professions/programs/telehealth-facilitator-certificate.html [accessed 2018-02-02] [WebCite Cache ID 6wvyi8VDQ]
14. Andersen R, Aday LA. Access to medical care in the US: realized and potential. Med Care 1978 Jul;16(7):533-546. [Medline: 672266]
15. Uscher-Pines L, Mehrotra A. Analysis of Teladoc use seems to indicate expanded access to care for patients without prior connection to a provider. Health Aff (Millwood) 2014 Feb;33(2):258-264. [doi: 10.1377/hlthaff.2013.0989] [Medline: 24493769]
16. Ashwood JS, Mehrotra A, Cowling D, Uscher-Pines L. Direct-to-consumer telehealth may increase access to care but does not decrease spending. Health Aff (Millwood) 2017 Mar 01;36(3):485-491. [doi: 10.1377/hlthaff.2016.1130] [Medline: 28264950]
17. Marcin JP, Shaikh U, Steinhorn RH. Addressing health disparities in rural communities using telehealth. Pediatr Res 2015 Oct 14;79(1-2):169-176 [FREE Full text] [doi: 10.1038/pr.2015.192]
18. Committee on Pediatric Workforce, Marcin J, Rimsza M, Moskowitz W. The use of telemedicine to address access and physician workforce shortages. Pediatrics 2015 Jul;136(1):202-209 [FREE Full text] [doi: 10.1542/peds.2015-1253] [Medline: 26122802]
19. Dixon R, Stahl J. 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]
20. 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]

21. Reichheld FF. Harvard Business Review. 2003 Dec. The one number you need to grow URL: https://hbr.org/2003/12/the-one-number-you-need-to-grow [accessed 2018-01-31] [WebCite Cache ID 6wt32AaXn]
22. Keiningham TL, Cooil B, Wallin Andreassen T, Aksoy L. Longitudinal examination of net promoter and firm revenue growth. J Marketing 2007;71:39-51.
23. Hamilton D, Lane J, Gaston P, Patton J, Macdonald DJ, Simpson A, et al. Assessing treatment outcomes using a single question: the net promoter score. Bone Joint J 2014 May;96-B(5):622-628. [doi: 10.1302/0301-620X.96B5.32434] [Medline: 24788496]
24. Pearl R. NEJM Catalyst. 2016 Mar 29. Engaging physicians in telehealth URL: https://catalyst.nejm.org/engaging-physicians-in-telehealth/ [accessed 2018-01-31] [WebCite Cache ID 6wt3b99Kb]
25. Menachemi N, Burke DE, Ayers DJ. Factors affecting the adoption of telemedicine-a multiple adopter perspective. J Med Syst 2004 Dec;28(6):617-632. [Medline: 15615290]

Abbreviations

NQF: National Quality Forum
VA: Veteran's Administration

Edited by G Eysenbach; submitted 17.07.17; peer-reviewed by Y Liu, J Clemensen; comments to author 24.11.17; revised version received 15.12.17; accepted 16.12.17; published 13.02.18.

Please cite as:
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;6(1):e10
URL: http://medinform.jmir.org/2018/1/e10/
doi:10.2196/medinform.8479
PMID:29439947

Original Paper

Experiences of Indian Health Workers Using WhatsApp for Improving Aseptic Practices With Newborns: Exploratory Qualitative Study

Parika Pahwa[1], BHMS, MBA; Sarah Lunsford[2], PhD; Nigel Livesley[1], MD

[1]University Research Co, LLC, Delhi, India

[2]EnCompass LLC, Chevy Chase, MD, United States

Corresponding Author:
Sarah Lunsford, PhD
EnCompass LLC
5404 Wisconsin Ave
Chevy Chase, MD, 20815
United States
Phone: 1 6177849008
Email: ssmith@urc-chs.com

Abstract

Background: Quality improvement (QI) involves the following 4 steps: (1) forming a team to work on a specific aim, (2) analyzing the reasons for current underperformance, (3) developing changes that could improve care and testing these changes using plan-do-study-act cycles (PDSA), and (4) implementing successful interventions to sustain improvements. Teamwork and group discussion are key for effective QI, but convening in-person meetings with all staff can be challenging due to workload and shift changes. Mobile technologies can support communication within a team when face-to-face meetings are not possible. WhatsApp, a mobile messaging platform, was implemented as a communication tool by a neonatal intensive care unit (NICU) team in an Indian tertiary hospital seeking to reduce nosocomial infections in newborns.

Objective: This exploratory qualitative study aimed to examine experiences with WhatsApp as a communication tool among improvement team members and an external coach to improve adherence to aseptic protocols.

Methods: Ten QI team members and the external coach were interviewed on communication processes and approaches and thematically analyzed. The WhatsApp transcript for the implementation period was also included in the analysis.

Results: WhatsApp was effective for disseminating information, including guidance on QI and clinical practice, and data on performance indicators. It was not effective as a platform for group discussion to generate change ideas or analyze the performance indicator data. The decision of who to include in the WhatsApp group and how members engaged in the group may have reinforced existing hierarchies. Using WhatsApp created a work environment in which members were accessible all the time, breaking down barriers between personal and professional time. The continual influx of messages was distracting to some respondents, and how respondents managed these messages (eg, using the silent function) may have influenced their perceptions of WhatsApp. The coach used WhatsApp to share information, schedule site visits, and prompt action on behalf of the team.

Conclusions: WhatsApp is a productive communication tool that can be used by teams and coaches to disseminate information and prompt action to improve the quality of care, but cannot replace in-person meetings.

(JMIR Med Inform 2018;6(1):e13) doi:10.2196/medinform.8154

KEYWORDS

quality improvement; mobile apps; communication; patient care team

Introduction

Interprofessional teamwork is an essential component of effective quality improvement (QI) in health care [1]. Health care staff must collaborate to make workflows and processes of care more efficient and improve patient outcomes. Successful teamwork in QI requires functional communication structures in which team members can participate [2]. Inclusive leadership and communication can create an environment in which staff feel valued, appreciated, and empowered to contribute to improvement efforts [3]. Communication failures in interprofessional teams can be attributed to different communication styles by cadre, hierarchical structures, and a culture in which mistakes are perceived as personal failings [4]. With high patient loads and conflicting schedules, it can be a challenge to find time for face-to-face communication with all or most of a QI team at regular intervals. Mobile technologies present an opportunity for team members to remain connected in between or in lieu of in-person meetings.

mHealth, “medical and public health practice supported by mobile devices” [5], has been applied in the health sector in low- and middle-income countries (LMICs) predominantly to educate and promote behavior change among patients. mHealth approaches have also been applied to medical imaging, collecting and transmitting patient-level data, and providing support to health workers in clinical tasks, communication with patients, and supply chain management [6]. A recent framework on mHealth as a health systems strengthening tool presented 12 common applications of mHealth, including to improve adherence to clinical protocols [7]. Another systematic review identified five uses of mobile technologies by frontline workers in LMICs [8]. Notably absent in the research from LMICs is the use of mobile platforms to support communication among health workers as a mechanism for improving care systems and processes.

WhatsApp is a messaging platform that allows users to send text messages, photographs, documents, and videos and make phone calls using a smartphone. The app allows for group chats with up to 256 members. There are an estimated one billion WhatsApp users worldwide, with about 160 million of those in India [9]. In health care in high-resource settings, WhatsApp has been employed as a means of communication in laboratory services [10] and emergency surgery [11], and in LMICs, it has been used to facilitate supervision of community health workers [12].

In spite of its use in health care, we found no published research on the application of WhatsApp or similar platforms to facilitate the work of QI teams in any setting. We sought to explore how a QI team in one Indian hospital communicated with each other and with a coach via WhatsApp while implementing modern improvement methods to improve adherence to aseptic protocols in the neonatal intensive care unit (NICU).

Methods

Study Site

The study hospital is a tertiary-level care hospital in Delhi providing free services to a population of 300,000 with nearly 7000 deliveries every year. The 15-bed NICU has a bed occupancy rate of 50% and is staffed by 5 pediatricians, 6 general doctors, 12 nurses, and 2 paramedical staff who provide round-the-clock services for an average of 120 newborn admissions per month.

The hospital had experience implementing modern QI methods with support from the United States Agency for International Development Applying Science to Strengthen and Improve Systems (USAID ASSIST) Project in the gynecology department, but this was the first QI activity implemented in the NICU. The positive impact of QI implementation in other departments and results from other QI projects in local hospitals were instrumental in motivating the NICU staff to take up this project. The NICU improvement team consisted of 21 staff, including 1 head of department, 1 senior specialist, 3 senior residents, 6 junior residents, 2 senior staff nurse, 1 sister in charge, 5 staff nurse, and 2 technicians. In total, 12 team members joined when the team was formed and formed the core team; the remaining 9 participated in the improvement activity as they were able.

The NICU team observed that the QI approach involved the following 4 steps: (1) forming a team to work on a specific aim, (2) analyzing the reasons for current poor performance, (3) developing changes that could improve care and testing these changes using plan-do-study-act (PDSA) cycles, and (4) implementing successful changes to sustain improvements. Analysis of unit data revealed that babies’ length of stay was increasing due to nosocomial infections, as aseptic protocols during intravenous (IV) procedures were not properly followed. In May 2016, the team began a QI activity on following asepsis protocol while performing IV procedures and collecting data regarding number of blood draws done on daily basis.

The QI team tested and implemented the following changes: prearranging blood sampling trays with all necessary items required to perform aseptic procedures; preparing a checklist to evaluate performance; and establishing a cardboard dropbox into which observers deposited completed checklists to make the evaluation process anonymous. The team continued to do PDSA cycles to eliminate unnecessary steps in the blood sampling process to make it simpler and adaptable to hospital staff. Within 12 weeks of starting the improvement activity, the team followed aseptic protocols in 80% of the blood samples taken and decided to expand their activities to improve central lines procedures.

Over the course of 18 weeks, the QI team received 8 in-person coaching visits from an external advisor with 3 years of experience in helping health workers use QI approaches. The coach was also in the WhatsApp group and engaged in regular communication. Not all in-person visits were equally fruitful due to medical emergencies or scheduling that prevented the team from coming together. The coach guided the team through the learning process of collecting, analyzing, and interpreting data to identify gaps in quality and generating changes to test to address those gaps.

Study Design

An exploratory, qualitative case study design was used to examine the role of WhatsApp in QI team communication and coaching. Ten team members were purposively selected for interviews to represent different cadres and tenures at the facility (Table 1). The coach who supported the team was also included in the interview sample to present an alternate perspective on WhatsApp for coaching and improvement.

Table 1. Respondents by title, sex, and role in quality improvement team. NICU: neonatal intensive care unit.

Pseudonym	Designation	Sex	Role in quality improvement team
Dr Manish	Senior Consultant, NICU	Male	Team leader
Dr Vijay	Senior Resident	Male	Team member
Dr Arun	Consultant	Male	Associate team leader
Dr Nisha	Senior Resident	Female	Team member
Dr Rajiv	Junior Resident	Male	Team member: took the lead in finalizing the standard operating procedure for asepsis protocol; used flowchart as tool for identifying the process
Anita	Senior Staff Nurse	Female	Team member: prepared checklist for evaluation and dropbox; main communicator for passing on all information to the unit people; active member on WhatsApp
Manjeet	Senior Staff Nurse	Male	Team member
Seema	Staff Nurse	Female	Team member: prepared sample collecting tray; was not on WhatsApp despite being an active team member
Kajal	Staff Nurse	Female	Team member
Raman	Technician	Male	Team member: responsible for data collection
Anay	Coach	Male	Coach

Data Collection

Semistructured interviews of team members were conducted by the first author. The coach who provided support to the team was interviewed by the second author. Questions on communication methods and the use of WhatsApp were part of a longer interview on team-based QI. Interviews were conducted at the hospital in Hindi. Audio recordings were transcribed and translated into English for analysis. In all, 18 weeks of the WhatsApp transcript (May-August 2016) were accessed, including images and files that were shared in the group. Communication in the WhatsApp group was predominantly in English.

Data Analysis

Interviews were coded using in vivo, process, and structural coding strategies. An initial coding scheme was developed by the first two authors; it was refined by the second author and applied to all interviews. Codes were then aggregated into categories and themes during discussions among all authors. Frequency of type and form of messages (quantification) of the WhatsApp transcript was conducted. Analysis was done using NVivo 11 (QSR International, Burlington, MA, USA).

Ethics approval was granted by University Research Co., LLC (USA). The purpose and procedures of this research were explained to all respondents, and informed consent was obtained. Pseudonyms have been used to protect confidentiality.

Results

Setting Up and Managing a WhatsApp Group

The WhatsApp group was established by Dr Manish, team leader, who was having difficulty scheduling a meeting to share a new standard operating procedure and thought a communications app could be of use, especially considering staff had already used mobile technology to communicate among themselves. Dr Manish had the perception that WhatsApp was universally used among the staff, an opinion shared by 3 other respondents, so it would be easy to employ the app in the workplace:

> *I think everyone is using WhatsApp for everything. So I thought let’s make a WhatsApp group and there I will ask about project details. I made a group.*

The real or imagined hierarchy in the facility influenced who was added to the WhatsApp group. Dr Manish recognized some hesitation on behalf of the nurses to share their mobile numbers with their superiors so they could be added to the group:

> *Yes I took help; I didn’t have contacts of all. Anita was there so I had to ask her to add few numbers. Nurses are not so comfortable to give their numbers to everyone especially doctors, so there was hesitation.*

From the more junior staff members' perspective, the inclusion of staff in the WhatsApp group was the team leader’s decision alone. Seema (nurse), who praised Dr Manish for his leadership in setting up the group, but was not a part of the group, stated:

> *I have left it on the sir [Dr. Manish]. He might have felt that I am not required in the group and I also never said anything to him about it.*

Messages Sent in WhatsApp

Over the course of 18 weeks, 279 messages were exchanged in WhatsApp. Although 9 members sent messages, almost three-quarters of all messages were sent from the leader of the team or the coach. The majority (73.8%, 206/279) were text messages, 22.9% (64/279) were images, and the remaining messages were documents or videos. The images were photos of completed data forms, resource and reference materials, documentation of PDSA cycles, and physical spaces within the hospital. The frequency of messages followed a pattern, increasing when a coach site visit was nearing, which remained high immediately following a coaching visit, and then waning until the coach or team leader began scheduling the next site visit.

Much of the perceived value of WhatsApp was the ability to quickly disseminate information or ideas to a group of people:

> *If we have an idea in our mind, we can communicate easily and everyone comes to know about the issue together. You are in a group, you send information and you don't have to tell everyone separately. You conveyed the message to all at one go.* [Seema]

It also provided a medium for clarifying clinical or QI guidance that was shared with the whole team at the same time (9 messages [3.2%, 9/279] containing materials on clinical guidance; 4 messages [1.4%, 4/279] containing materials on general QI guidance; and 14 messages [5.0%, 14/279] pertaining to the team's improvement activity such as their aim, standard operating procedure, and flow chart). Flow charts were used by the team to help analyze the steps of carrying out IV procedures to identify where the aseptic technique was not being followed:

> *They were not understanding the flow chart, so Doctor made the flow chart again and sent in WhatsApp...Everything was discussed in [WhatsApp].* [Anita]

Finally, the app was used to share data on compliance with the aseptic technique, revealing the gaps in and raising awareness of compliance (26 messages [9.3%, 26/279] on the team's data):

> *We came to know our mistakes, like 10 samples happened and we washed hands properly in 7 only and so on. Then there comes a time that if 10 pricks happened then among all 10 handwashing was followed properly. This increased awareness.* [Dr Vijay]

Accessibility All the Time

Respondents both praised and critiqued WhatsApp for facilitating access at all times. The key benefit was being able to disseminate information without having to convene a meeting with all staff, which proved difficult with staff shifts and workloads. WhatsApp provided a medium for staff to both share ideas or information at any time and engage with the information at any time:

> *Sometimes one is not active and cannot listen properly in-person meetings but on WhatsApp he can at least read whenever he is feeling comfortable. If you are discussing something and someone is very tired, not able to pay attention at that time, they can always read the messages anytime they are comfortable.* [Dr Manish]

Despite these benefits of WhatsApp, being accessible all the time came with two notable drawbacks. Logistically, there was the expectation that facility staff use their own personal data packages to engage in the WhatsApp group, including downloading documents and watching videos. Additionally, being accessible all the time resulted in little separation or balance between work and personal life. One respondent stated she was able to remain up to date on work through the WhatsApp communication while she was on leave. She viewed remaining informed as an asset, but it did not allow her to separate herself from work while away.

Engaging (or Not) in WhatsApp

Across respondents, it was generally agreed that there were between 15 and 20 members in the WhatsApp group, but only 3 to 5 active participants. Although the WhatsApp transcript indicated that 9 members shared messages, most messages came from the team leader and coach, highlighting the variable levels of participation.

Several reasons for not actively participating in the WhatsApp group were presented. The hierarchical structure of staffing contributed to fear and hesitation around being judged by peers and superiors for one's remarks, comments, or suggestions:

> *So, they were taking it as an order from senior authorities and we have to follow it. Like if we are working and you have got information from senior authorities so generally you don't question back, you just follow it and if they are having any problem so they prefer to discuss it personally rather than on group.* [Dr Arun]

The same respondent (Dr Arun) posited the inverse as well, that in a hypothetical group of nurses a new nurse:

> *...might open up or maybe a [nurse with 20 years of experience] might not open up because if she had done any mistake then she is under fear that everybody will get to know about that.*

WhatsApp was also a means of maintaining a record of messages and a means of promoting accountability:

> *This gets documented. In personal meetings, people forgot whatever you said, it will never be documented but on WhatsApp you can see that what has been said and by whom. It is like committing yourself when you are. When you discuss personally then there are so many suggestions which are coming but none of them ever came on WhatsApp. So there, people can talk freely as nobody is making video of anybody that who I speaking, but when you are typing then it becomes a proof that you have said this and then it's his responsibility to do that task.* [Dr Manish]

This sentiment, expressed by a senior staff person, encapsulated one reason why group members were reluctant to actively participate. According to Dr Vijay, they did not want to be held accountable for any specific suggestion or idea:

> *If you put anything there then you will be bound to do that.*

Anay, the coach, in contrast, viewed WhatsApp as an *informal mode of communication* and did not perceive the enhanced accountability attribute of the technology.

Manjeet, echoed by Dr Arun and Kajal, offered another interpretation for the lack of responses, suggesting there was no value in contributing via WhatsApp when in-person meetings and discussion were going to take place, indicating that the real value of WhatsApp, in the professional setting, was to disseminate and share information rather than discuss.

Raman, Dr Nisha, and Seema went further, stating that WhatsApp was *irritating* and *worthless* because:

> *It consumes our time, like we are busy in our work and then we got a message and then we had to check it and have to give reply so it wastes our time. So our attention is on WhatsApp rather than on patient care.* [Raman]

Similarly, it was argued that there were so many messages being sent via WhatsApp that it was impossible to read them all. This was countered by other respondents who silenced their mobile devices so they were not notified of new messages:

> *[While] doing something for which I don't require my phone to buzz, you can always put it on silent.* [Dr. Nisha]

It was suggested that had group members been told there was an expectation of discussion in the group, they would have participated more actively; however, that expectation was not clear. According to Anay, the technology was able to report who had received and opened a message, but it could not prompt someone to reply, so it was never clear if someone was not responding because they were not able to or did not want to engage.

Finally, Anay shared that language may have presented a barrier to participation. Some of the group members were not as confident in English, which was the dominant language used in WhastsApp communication.

In spite of the variable participation in the WhatsApp group, one respondent pointed to the documented improvements in clinical practice, noting that whether staff responded in the app was not essential if they were implementing improved practices in their patient care:

> *We used to discuss that it is working because we were seeing through WhatsApp that number of pricks are being counted daily so we were discussing that something has kicked on, something has happened. Positive discussions started and through WhatsApp group we were seeing that positive results are coming then we realized that we might not be active but things are happening really good and that gives us motivation. A famous saying that motivation is more important than inspiration.* [Dr Vijay]

WhatsApp Versus Other Communication Methods

Although WhatsApp was viewed overall as a valuable resource for sharing information, it was not a replacement to in-person meetings and discussion, rather a supplement. In-person meetings were needed to discuss the documents shared via the app:

> *Face-to-face is of course the best. WhatsApp is like you are kicking a board but the board will not kick you back. We cannot underestimate the power of personal interaction. WhatsApp is like we can be aware of all things.* [Dr Vijay]

> *Things cannot be understood clearly on WhatsApp but in meetings we can discuss it in detail and as a result we can understand it much better as we are getting both theoretical and practical information clearly, because you will get to know about it completely when you are seeing and discussing those situations practically but in WhatsApp this is not possible.* [Raman]

In-person meetings were also viewed as essential for building team work, which could not be done via WhatsApp.

There was some discussion of the value of conference calls to share information, but this was generally agreed to be an ineffective communication medium due to scheduling, cost, the limitations of having a remotely delivered lecture, and the challenges of having equal participation if too many people were on the call. Anay viewed conference calls as a new mode of communication for health care staff; although WhatsApp was not new, it was *tapping into something already existing*.

Coaching via WhatsApp

Anay, the coach, indicated that there were benefits of coaching via WhatsApp as well as drawbacks (Textbox 1). He was able to share materials on QI methods via WhatsApp, so he did not have to bring hard copies to in-person meetings and so team members could review materials in preparation for a coaching visit. WhatsApp also functioned as a coaching management tool for setting up visits, which was particularly helpful with the team, given the scheduling challenges they experienced. Documentation in WhatsApp allowed the coach to stay up to date on the activities and efficiently tailor coaching to that specific activity and team needs. The coach could also observe when enthusiasm appeared to be lagging among team members and purposively schedule a visit to re-energize the group. Finally, the coach was able to review data and observe that data collection was ongoing, but that the team was not able to learn from the data.

From the coach's perspective, a limitation of using an exclusively in-person coaching approach was the inability to remain a part of the improvement process when not at the facility. Anay explained that with in-person coaching, the coach would visit the facility, help plan PDSA cycles, but then not know if the facility QI team followed through on the "do" phase and frequently would not "study" the impact of the change until the coach returned for another visit. Via WhatsApp, however, Anay could prompt the team to "do" and "study" by posing questions about the results to the group. Thus, WhatsApp provided an avenue to *keep them on track through small-scale testing and correcting them when I saw them going astray*. For example, one message Anay sent emphasized the importance of doing small PDSA cycles to test proposed changes, such as a checklist:

> *It helps to do things in small bits...We might learn as we use this tool and may want to modify it.*

Textbox 1. Elements of coaching that were and were not made easier by using WhatsApp.

Made easier by WhatsApp

- Sharing materials
- Scheduling site visits
- Identifying gaps in quality improvement (QI) knowledge
- Prompting action related to analyzing problems and testing changes
- Observing ongoing team dynamics

Made more difficult by WhatsApp

- Gaining initial understanding of team dynamics
- Facilitating learning from data
- Performing direct observation of and feedback on clinical practice

Following the test for the checklist, Anay reminded the group of the questions they need to think about to determine if the checklist was successful:

> _Excited to know the results of the first test! We are seeking answers to 1. What is the current performance level on different steps of asepsis? 2. How easy or difficult is it to fill the checklist? 3. Any side effects of this change in how we draw blood samples? And finally *What do we do next?*_ [Anay, WhatsApp chat transcript]

Anay shared that WhatsApp is a good complement to in-person meetings, but getting to know the team can only happen in-person:

> _It's very difficult to understand the team dynamics from a WhatsApp conversation, at least with a new team it is quite difficult. Now that I know these team members, who is who, even with the tone of conversation I can make out a lot about team dynamics._ [Anay]

Anita noted that the coach had on one or two occasions _lowered his support_ but _in between through WhatsApp he used to be in touch._

Other Uses of WhatsApp

Outside of the application of WhatsApp to improve processes and systems of care, the messaging platform was used to communicate about other aspects of clinical care. Both nurses and doctors sent patients images and other information to colleagues for input on diagnosis and treatment. This form of communication was also used to confront some of the hierarchy present in the staffing structure.

One respondent, a nurse, recounted recent experiences in which a doctor was not physically present but did not agree with her diagnosis or suggestion for treatment. She was able to take a video or photograph of the patient and share it via WhatsApp with the doctor and receive confirmation on the diagnosis and treatment plan:

> _Sometimes doctors are not there in the room and caesarean is there and we are telling this happened, so he tells that now baby is fine how come it was caesarean? It did not come in front of him and you will explain everything but he will not believe because it was not in front of his eyes. After that we thought to make video at that time because it happened with me personally. Two days ago, one baby's x-ray was bad we took the picture and sent because I thought the senior resident does not understand because he was new. He told me the x-ray is fine but I took the picture of the x-ray and sent it to sir and I told him that x-ray is not fine, do something. The video making thing is helpful because at least you are having some proof that you are telling right._ [Anita]

Other uses of WhatsApp included scheduling or discussing nonwork-related issues such as celebrations.

Discussion

WhatsApp was applied as a communications platform by a QI team in a NICU in one hospital in India. An improvement team aimed to improve compliance with aseptic protocols to reduce the risk of nosocomial infections. Using WhatsApp, QI team members shared materials, data on compliance with protocols, and, to a lesser extent, changes to test.

The principal use of WhatsApp among this team was to disseminate information, which, although an integral part of QI, is only part of the process. Dynamic discussion among team members and with the coach is an essential step in analyzing data and generating changes to test. Yet, this medium did not facilitate sharing among this team.

Perceptions of quality communication among improvement team members can yield improved patient outcomes [13] and team cohesion [14]. Mobile technology can facilitate communication between health workers [15]. However, the impact of frequent messaging on workload and delivering services was a concern of some of our respondents and has been documented in other research [10]. It is also possible that through the frequent messages too much information was being shared, overloading team members. Such information and communication overload can negatively impact productivity [16] and should be managed to keep QI team members engaged in the improvement work. How team members managed the influx of messages (eg, using the silent function) may have influenced their perception of WhatsApp and its application in the improvement work. Related was the sense of always being available and engaged in work, which may create conflict between personal and professional lives or may allow for greater flexibility in both [17,18]. The expectation of always being available, therefore, is a double-edged sword.

Hierarchy played a role in the relationships between senior and junior residents and between doctors and nurses. This manifested itself in who was included in the WhatsApp group and their levels of participation and likely was present in other forms of interprofessional interaction; thus, it is possible that WhatsApp reinforced existing hierarchies. Nurses who feel like they contribute to decision making with the doctors are less likely to leave their jobs; conversely, nurses who do not feel they have positive relationships with doctors experience greater levels of professional stress [19,20]. It is important to move toward a culture of respect and equity among staff to improve job satisfaction and health worker retention. Mentoring or other similar supportive relationships may aid in flattening the hierarchical structure.

A benefit of WhatsApp was the ability to send patient images to other clinical staff for review and input on treatment approaches. There is mixed evidence on the utility of using websites and mobile technologies for transmitting images for diagnosis, review, and feedback from experts [6,15] and as a mechanism for building capacity of providers in low- and middle-income settings [21]. Although we did not ask about concerns of patient privacy, it should be taken into consideration when deciding whether and how to implement a messaging platform like WhatsApp in clinical services [22]. During the course of this activity, WhatsApp implemented end-to-end encryption, which may provide adequate security for maintaining patient privacy; however, the ethical and legal implications should be examined thoroughly.

WhatsApp not only facilitated communication among team members but was a useful, though limited, tool for QI coaching. Specifically, WhatsApp provided a real-time way of identifying when the team was having problems in teamwork and participation in the QI activity, problem and data analysis, and closing PDSA cycles. WhatsApp was very helpful in providing guidance on doing PDSA in real time. Establishing an improvement aim, forming a cohesive team, and analyzing data were much harder to support using a messaging platform and required in-person site visits. This study did not examine the sustainability of improvements made by the QI team, but we would expect that a coach could remotely inspire continued enthusiasm among a team without needing to visit in-person.

WhatsApp and other similar mobile messaging platforms have been underutilized and under-researched in LMICs to facilitate communication around health care improvement. Our study is a small-scale pilot that offers valuable insight, but does not offer evidence on the effectiveness or cost-effectiveness of these technologies. Thus, the implementation of mobile messaging platforms and other mHealth interventions needs to be scaled up and rigorously evaluated to better understand how these technologies impact health outcomes. This study was also limited to the WhatsApp-based communications and did not capture other formal or ad hoc communications that were part of the QI activity.

This exploration in the application of WhatsApp to aid in communication within a QI team shows the platform's promise and highlights some areas of consideration before implementation. First, the decision to use a mobile communication app should be discussed and agreed upon by all team members, and all team members should be invited to participate equally. Building on this, issues of hierarchy within the staff and QI team structure should be addressed both in-person and via mobile technologies to better engage all staff in improvement activities. Second, the volume of information shared should be managed to allow staff to review and reflect on the information. Similarly, a culture of use should be fostered that creates expectations of participation even if not in real time and includes ground rules on appropriate and inappropriate use and patient privacy issues. Team leaders should keep track of participation and react if people are not engaging in discussion in WhatsApp. For coaching teams, other tools, such as short videos, should be prepared to aid in improving team work, using analysis tools, and analyzing data that could be shared in WhatsApp.

Acknowledgments

The authors thank the quality improvement team members for their continued work on improving the quality of care for patients and for sharing their experiences. The authors thank Edward Broughton, Lani Marquez, M Rashad Massoud, and Lisa Maniscalco for comments on early drafts of the manuscript. This work was carried out by the USAID Applying Science to Strengthen and Improve Systems Project, which is funded by the American people through USAID's Bureau for Global Health, Office of Health Systems (Cooperative Agreement Number AID-OAA-A-12-00101).

Conflicts of Interest

None declared.

References

1. Thomas EJ. Improving teamwork in healthcare: current approaches and the path forward. BMJ Qual Saf 2011 Aug;20(8):647-650. [doi: 10.1136/bmjqs-2011-000117] [Medline: 21712372]

2. Proudfoot J, Jayasinghe UW, Holton C, Grimm J, Bubner T, Amoroso C, et al. Team climate for innovation: what difference does it make in general practice? Int J Qual Health Care 2007 Jun;19(3):164-169. [doi: 10.1093/intqhc/mzm005] [Medline: 17337517]
3. Vogelsmeier A, Scott-Cawiezell J. Achieving quality improvement in the nursing home: influence of nursing leadership on communication and teamwork. J Nurs Care Qual 2011;26(3):236-242. [doi: 10.1097/NCQ.0b013e31820e15c0] [Medline: 21278595]
4. Leonard M, Graham S, Bonacum D. The human factor: the critical importance of effective teamwork and communication in providing safe care. Qual Saf Health Care 2004 Oct;13(Suppl 1):i85-i90 [FREE Full text] [doi: 10.1136/qhc.13.suppl_1.i85] [Medline: 15465961]
5. World Health Organization. mHealth: New horizons for health through mobile technologies. Geneva: World Health Organization; 2011.
6. Hall CS, Fottrell E, Wilkinson S, Byass P. Assessing the impact of mHealth interventions in low- and middle-income countries--what has been shown to work? Glob Health Action 2014;7:25606 [FREE Full text] [Medline: 25361730]
7. Labrique AB, Vasudevan L, Kochi E, Fabricant R, Mehl G. mHealth innovations as health system strengthening tools: 12 common applications and a visual framework. Glob Health Sci Pract 2013 Aug;1(2):160-171 [FREE Full text] [doi: 10.9745/GHSP-D-13-00031] [Medline: 25276529]
8. Agarwal S, Perry HB, Long L, Labrique AB. Evidence on feasibility and effective use of mHealth strategies by frontline health workers in developing countries: systematic review. Trop Med Int Health 2015 Aug;20(8):1003-1014 [FREE Full text] [doi: 10.1111/tmi.12525] [Medline: 25881735]
9. Gadgets 360. WhatsApp Has Over 160 million monthly active users in India, its Biggest Market URL: https://gadgets.ndtv.com/apps/news/whatsapp-now-has-over-160-million-monthly-actiove-users-in-india-1625558 [accessed 2018-02-14] [WebCite Cache ID 6xETeTG4f]
10. Dorwal P, Sachdev R, Gautam D, Jain D, Sharma P, Tiwari AK, et al. Role of WhatsApp messenger in the laboratory management System: a boon to communication. J Med Syst 2016 Jan;40(1):14. [doi: 10.1007/s10916-015-0384-2] [Medline: 26573651]
11. Johnston MJ, King D, Arora S, Behar N, Athanasiou T, Sevdalis N, et al. Smartphones let surgeons know WhatsApp: an analysis of communication in emergency surgical teams. Am J Surg 2015 Jan;209(1):45-51. [doi: 10.1016/j.amjsurg.2014.08.030] [Medline: 25454952]
12. Henry JV, Winters N, Lakati A, Oliver M, Geniets A, Mbae SM, et al. Enhancing the supervision of community health workers with WhatsApp mobile messaging: qualitative findings From 2 low-resource settings in Kenya. Glob Health Sci Pract 2016 Jun 20;4(2):311-325 [FREE Full text] [doi: 10.9745/GHSP-D-15-00386] [Medline: 27353623]
13. Arling PA, Abrahamson K, Miech EJ, Inui TS, Arling G. Communication and effectiveness in a US nursing home quality-improvement collaborative. Nurs Health Sci 2014 Sep;16(3):291-297. [doi: 10.1111/nhs.12098] [Medline: 24256620]
14. Mickan SM, Rodger SA. Effective health care teams: a model of six characteristics developed from shared perceptions. J Interprof Care 2005 Aug;19(4):358-370. [doi: 10.1080/13561820500165142] [Medline: 16076597]
15. Free C, Phillips G, Watson L, Galli L, Felix L, Edwards P, et al. The effectiveness of mobile-health technologies to improve health care service delivery processes: a systematic review and meta-analysis. PLoS Med 2013 Jan;10(1):e1001363 [FREE Full text] [doi: 10.1371/journal.pmed.1001363] [Medline: 23458994]
16. Karr-Wisniewski P, Lu Y. When more is too much: operationalizing technology overload and exploring its impact on knowledge worker productivity. Comput Human Behav 2010;26(5):1061-1072. [doi: 10.1016/j.chb.2010.03.008]
17. Roy G. Impact of mobile communication technology on the work life balance of working women - a review of discourses. JCMR 2016;10(1):79-101.
18. Wright KB, Abendschein B, Wombacher K, O'Connor M, Hoffman M, Dempsey M, et al. Work-related communication technology use outside of regular work hours and the work life conflict: the influence of communication technologies on perceived work life conflict, burnout, job satisfaction, and turnover intentions. Manag Commun Q 2014 May 14;28(4):507-530 [FREE Full text] [doi: 10.1177/0893318914533332]
19. Lakshman S. Nurse turnover in India: factors impacting nurses' decisions to leave employment. SAJHRM 2016 Nov 04;3(2):109-128 [FREE Full text] [doi: 10.1177/2322093716657470] [Medline: 22973420]
20. Sharma P, Davey A, Davey S, Shukla A, Shrivastava K, Bansal R. Occupational stress among staff nurses: controlling the risk to health. Indian J Occup Environ Med 2014 May;18(2):52-56 [FREE Full text] [doi: 10.4103/0019-5278.146890] [Medline: 25568598]
21. Swanson JO, Plotner D, Franklin HL, Swanson DL, Lokomba BV, Lokangaka A, et al. Web-based quality assurance process drives improvements in obstetric ultrasound in 5 low- and middle-income countries. Glob Health Sci Pract 2016 Dec 23;4(4):675-683 [FREE Full text] [doi: 10.9745/GHSP-D-16-00156] [Medline: 28031304]
22. Adesina AO, Agbele KK, Februarie R, Abidoye AP, Nyongesa HO. Ensuring the security and privacy of information in mobile health-care communication systems. S Afr J Sci 2011;107(9-10):26-32. [doi: 10.4102/sajs.v107i9/10.508]

Abbreviations

IV: intravenous
LMIC: low- and middle-income country
NICU: neonatal intensive care unit
PDSA: plan-do-study-act
QI: quality improvement
USAID: United States Agency for International Development

Edited by G Eysenbach; submitted 07.06.17; peer-reviewed by S Buchholz, L Vasudevan; comments to author 22.11.17; revised version received 26.12.17; accepted 03.01.18; published 01.03.18.

Please cite as:
Pahwa P, Lunsford S, Livesley N
Experiences of Indian Health Workers Using WhatsApp for Improving Aseptic Practices With Newborns: Exploratory Qualitative Study
JMIR Med Inform 2018;6(1):e13
URL: http://medinform.jmir.org/2018/1/e13/
doi:10.2196/medinform.8154
PMID:29496651

Original Paper

Effect of Seasonal Variation on Clinical Outcome in Patients with Chronic Conditions: Analysis of the Commonwealth Scientific and Industrial Research Organization (CSIRO) National Telehealth Trial

Ahmadreza Argha[1*], BE, ME, PhD; Andrey Savkin[1*], BE, PhD; Siaw-Teng Liaw[2*], MBBS, PhD, FRACGP; Branko George Celler[1,3*], BSc, BE, PhD

[1]Biomedical Systems Research Laboratory, University of New South Wales, Kensington, Australia

[2]Ingham Institute of Applied Medical Research, School of Public Health and Community Medicine, University of New South Wales, Kensington, Australia

[3]eHealth Research Program, Commonwealth Scientific and Industrial Research Organisation, Marsfield, Australia

[*]all authors contributed equally

Corresponding Author:
Branko George Celler, BSc, BE, PhD
Biomedical Systems Research Laboratory
University of New South Wales
Electrical Engineering and Telecommunications
Kensington, 2052
Australia
Phone: 61 0418228297
Email: b.celler@unsw.edu.au

Abstract

Background: Seasonal variation has an impact on the hospitalization rate of patients with a range of cardiovascular diseases, including myocardial infarction and angina. This paper presents findings on the influence of seasonal variation on the results of a recently completed national trial of home telemonitoring of patients with chronic conditions, carried out at five locations along the east coast of Australia.

Objective: The aim is to evaluate the effect of the seasonal timing of hospital admission and length of stay on clinical outcome of a home telemonitoring trial involving patients (age: mean 72.2, SD 9.4 years) with chronic conditions (chronic obstructive pulmonary disease coronary artery disease, hypertensive diseases, congestive heart failure, diabetes, or asthma) and to explore methods of minimizing the influence of seasonal variations in the analysis of the effect of at-home telemonitoring on the number of hospital admissions and length of stay (LOS).

Methods: Patients were selected from a hospital list of eligible patients living with a range of chronic conditions. Each test patient was case matched with at least one control patient. A total of 114 test patients and 173 control patients were available in this trial. However, of the 287 patients, we only considered patients who had one or more admissions in the years from 2010 to 2012. Three different groups were analyzed separately because of substantially different climates: (1) Queensland, (2) Australian Capital Territory and Victoria, and (3) Tasmania. Time series data were analyzed using linear regression for a period of 3 years before the intervention to obtain an average seasonal variation pattern. A novel method that can reduce the impact of seasonal variation on the rate of hospitalization and LOS was used in the analysis of the outcome variables of the at-home telemonitoring trial.

Results: Test patients were monitored for a mean 481 (SD 77) days with 87% (53/61) of patients monitored for more than 12 months. Trends in seasonal variations were obtained from 3 years' of hospitalization data before intervention for the Queensland, Tasmania, and Australian Capital Territory and Victoria subgroups, respectively. The maximum deviation from baseline trends for LOS was 101.7% (SD 42.2%), 60.6% (SD 36.4%), and 158.3% (SD 68.1%). However, by synchronizing outcomes to the start date of intervention, the impact of seasonal variations was minimized to a maximum of 9.5% (SD 7.7%), thus improving the accuracy of the clinical outcomes reported.

Conclusions: Seasonal variations have a significant effect on the rate of hospital admission and LOS in patients with chronic conditions. However, the impact of seasonal variation on clinical outcomes (rate of admissions, number of hospital admissions, and LOS) of at-home telemonitoring can be attenuated by synchronizing the analysis of outcomes to the commencement dates for the telemonitoring of vital signs.

Trial Registration: Australian New Zealand Clinical Trial Registry ACTRN12613000635763; https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=364030&isReview=true (Archived by WebCite at http://www.webcitation.org/ 6xLPv9QDb)

(JMIR Med Inform 2018;6(1):e16) doi:10.2196/medinform.9680

KEYWORDS

telehealth; telemonitoring; seasonal variation; clinical trial; vital signs; chronic disease

Introduction

Telehealth systems in at-home, primary care, and hospital-based settings have been extensively investigated for more than 20 years [1-5]. Large health care organizations, such as the Veterans Administration in the United States and the National Health Service in the United Kingdom, have already adopted a range of telehealth solutions [6]. Employment of telehealth services for the management of patients with chronic conditions has progressively increased in recent years because of population aging and the increasing burden of chronic disease, along with the availability of low-cost monitoring technology.

Several trials have been carried out to analyze clinical, service, and economic benefits of telehealth systems [7,8]. These analyses are crucial to encourage wide-scale implementation of telehealth services. However, to the authors' best knowledge, no study has examined the impact of seasonal timing of hospital admission and length of stay (LOS) on clinical outcomes of a home telemonitoring trial.

This paper discusses the possible effect of seasonal variations on the clinical outcomes of a recently completed Commonwealth Scientific and Industrial Research Organization (CSIRO) trial of home monitoring for chronic disease management, carried out at several locations along the east coast of Australia [9]. The aim of this trial was to investigate health care outcomes as well as clinical and economic benefits of telehealth systems by introducing a telehealth model of service based on at-home telemonitoring of vital signs to patients with a range of chronic conditions supervised in either in hospital-based or community-based settings. The clinical protocols for the trial [9], the data architecture design [10], decision support and statistical trend analysis of vital signs data [11], and the impact of telemonitoring on health care expenditure, hospital admissions, and LOS [8] have been published previously.

In this paper, we introduce a novel method to estimate seasonal trends in hospitalization data of 136 patients with cardiovascular disease, respiratory disease, and diabetes over 3 years (January 1, 2010 to December 31, 2012), recruited in the CSIRO National Telehealth Trial. The final hypothesis in this paper is to show that seasonal variations can be minimized to have little or no significant influence on the clinical outcomes reported for the CSIRO National Telehealth Trial.

Methods

Research Ethics Committee Approval

The CSIRO Human Research Ethics Committee (HREC) as well as five other local HRECs approved the clinical trial protocol for this study (Approval Number: 13/04, March 25, 2013).

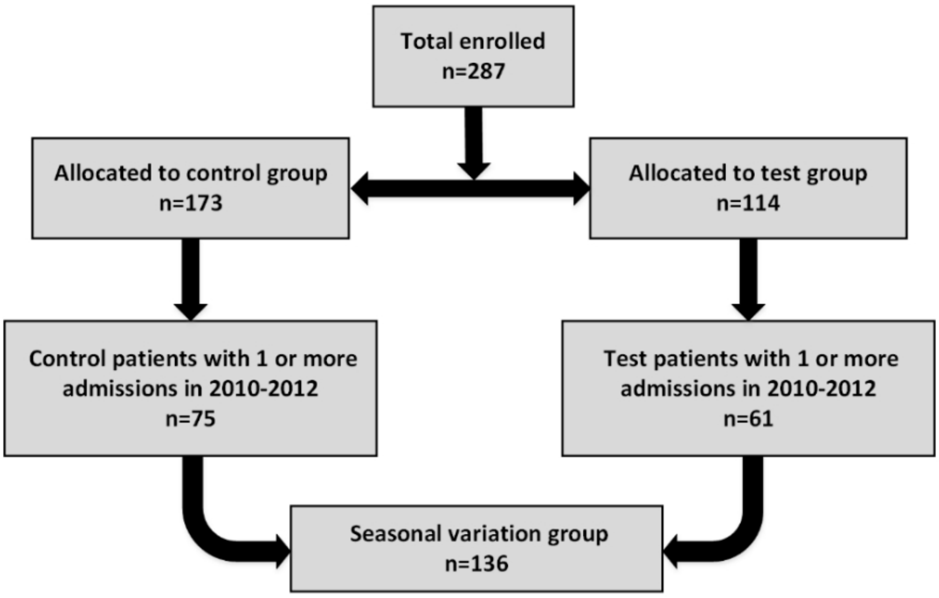
Patient Selection

In this trial, 1429 eligible patients from hospital lists provided by local health districts and patients known to clinical staff formed a Master Register. The local health districts were located in the states of Queensland, New South Wales, the Australian Capital Territory, and Tasmania. Inclusion criteria were thoroughly described in a previous article [9]; for convenience, we briefly summarize them here: age 50 years and older; at least two unplanned acute admissions during the previous 12 months or at least four unplanned acute admissions during the previous 5 years, with a principal diagnosis of coronary artery disease, congestive heart failure, hypertensive diseases, chronic obstructive pulmonary disease (COPD), asthma, or diabetes. Patients with compromised cognitive function, a neuromuscular disease, cancer, or a psychiatric condition were excluded from the trial. From the Master Register, 479 were deemed eligible and were contactable following individual screening.

Of the 479 eligible patients, only 287 could commence the trial, of which 114 were allocated to the telemonitoring test group and 173 were allocated to the control group [8]. The test patients were supplied with a telemonitoring system and trained on its use at installation, whereas the control group received only normal care through their primary care physician. Of the 287 patients monitored in this trial, we only considered patients who had one or more admissions in the years 2010 to 2012. With this additional inclusion criterion, 61 patients from the test group and 75 patients from the control group were selected to estimate the seasonal variation trends in the three years of 2010 to 2012. Figure 1 summarizes the patient selection process for this study.

However, because the telemonitoring intervention was only experienced by the 61 test patients, only these patients could be considered when evaluating the effect of seasonal variations on outcome variables, noting that the start of the intervention was synchronized for all test patients, considering each of the three climate subgroups separately.

Figure 1. Final cohort of seasonal variation group.


Total enrolled
n=287
Allocated to control group
n=173
Allocated to test group
n=114
Control patients with 1 or more admissions in 2010-2012
n=75
Test patients with 1 or more admissions in 2010-2012
n=61
Seasonal variation group
n=136

Definition of Seasons

The temperate zone along the eastern seaboard of Australia occupies the coastal hinterland of New South Wales, much of Victoria, and Tasmania. Hence, the four test sites in this trial are located in the temperate zone where seasons, in terms of European seasons applied to the southern hemisphere, are described as follows: summer (December to February), autumn (March to May), winter (June to August), and spring (September to November).

However, the fifth trial site, Townsville in Queensland, is within a subtropical zone, which is dominated by two distinct seasons: the wet season in summer (November to April) and the dry season in winter (May to October). Summer months in this city are generally hot and humid with day temperatures often around 29°C to 31°C and night temperatures around 20°C to 24°C. Winter months are generally warm to mild with day temperatures often around 25°C to 29°C and night temperatures around 13°C to 18°C.

Regression Modeling

The number of hospital admissions and LOS were analyzed as outcome variables in this study. As discussed in a previous article [8], all the outcome variables of this trial, including admission rates and LOS, were expected to increase over time because the patients involved in this trial were chronically ill and aging. To remove the aging effects from the 3 years' of hospitalization data, we fitted a linear model with the calendar months of the year as variables (3 years=36 months).

To implement this, the admission rate and LOS were summed for all patients within each calendar month of the year. The monthly time course of 3 years' of data was then modeled using linear regression to identify statistically significant differences in admission rates and LOS slopes.

We used the "fit" command in the MATLAB (The MathWorks Inc) statistics toolbox to carry out linear regression. To obtain 95% prediction intervals, the command "predObs" was used to plot 95% prediction intervals. A 95% prediction interval is an estimate of an interval in which future observations will fall, with 95% probability, given what has already been observed.

Moreover, different standard goodness-of-fit measures, including the coefficient of determination (R^2), the R^2 value adjusted for degrees of freedom, the residual sum of squares, and the standard error or root mean square error were considered.

Following the derivation of a linear model (baseline) in this study, we replaced the absolute vales of outcome variables (rate of admission, number of hospital admissions, and LOS) with percentage deviation from the baseline. The deviation from baseline was defined as the distance between each observations point (outcome variable at each month) and its corresponding point at baseline, which was estimated by the linear regression line of best fit . Then, a yearly seasonal variation trend was obtained by averaging over each calendar month the percentage deviation from baseline over the 3 years' of data available.

Seasonal Effects on Outcome Variables

During interventions, the 61 test patients were monitored for a mean 481 (SD 77) days with no significant difference between average monitoring durations for female patients (mean 498, SD 82 days) and male patients (mean 463, SD 67 days). Of the 61 test patients, 87% (53/61) were monitored for periods exceeding 12 months.

As proposed in a previous article [8], a possible method to minimize the impact of seasonal variation on the outcomes of the telemonitoring trial is to synchronize medical, pharmaceutical, and hospital data to the date when the telemonitoring commenced, thus effectively smoothing the effect of seasonal variations. However, no accurate analysis was given in that article on the effect of time synchronization compared to using actual monitoring durations.

Hence, to compare these two methods, we compensated for the effect of seasonal variation on each patient monthly data point by using the yearly seasonal variation model obtained previously from 3 years' of hospital data.

Statistical Analysis

To determine the statistical significance of the differences between subgroups, a two-sample *t* test was performed for continuous variables and the Wilcoxon rank sum test was carried out for skewed variables. For describing baseline characteristics, we used means and standard deviations for continuous symmetrical variables and medians and 95% confidence intervals for skewed data.

Categorical variables are also presented as counts and percentages. All statistical tests were two-tailed, and *P*<.05 was considered statistically significant. Statistical analyses were performed using MATLAB (R2016b) and Microsoft Excel.

Results

Findings

Basic demographics of seasonal variation group in the study are given in Table 1. There were no significant differences in age between test patients in each of the five sites and between male and female patients. Test patients continued to be monitored in their own home, and no patient requested a relocation of their telemonitoring equipment to another location during the trial.

In this study, 58.1% (79/136) of the patients were male and 41.9% (57/136) were female. Most patients included in this study had more than one condition listed as a primary diagnosis, but for simplicity, primary disease conditions were grouped in the broad categories of cardiovascular disease (n=55), respiratory disease (n=66), and diabetes (n=15).

The 136 combined test and control patients were admitted to hospital 817 times during 2010 to 2012, with a total LOS of 3627 days. Adopting the same definition of season for the site in Queensland, the highest number of patients were admitted during spring (218/817, 26.7%) followed by winter (216/817, 26.4%), autumn (207/817, 25.3%), and summer (176/817, 21.5%). Excluding winter, the number of hospital admissions was significantly higher in spring compared to other seasons (*P* values versus winter, autumn, and summer were .86, .04, and .02, respectively).

However, LOS was higher during winter (1007/3627, 27.76%) followed by spring (990/3627, 27.30%), and autumn (916/3627, 25.26%). Overall LOS was shorter during summer (714/3627, 19.69%). Except for spring, LOS was significantly longer in winter compared to other seasons (*P* values versus spring, autumn, and summer were .62, .03, and .01, respectively).

We considered three different groupings because of substantially different climates: (1) Queensland (subtropical), (2) Australian Capital Territory and Victoria (temperate), and (3) Tasmania (colder). Regarding Queensland, because only two distinct seasons (winter and summer) are notable, we compared the variables for these two seasons. The number of hospital admission and overall LOS were significantly higher and longer in winter (hospital admissions=157, LOS=633 days) compared to summer (hospital admissions=98, *P*=.01; LOS=338 days, *P*=.02).

Table 1. Basic demographics of seasonal variation patients.

Location	Demographics					
	Patients, n	Patient age, mean (SD)	Male, n	Male patient age, mean (SD)	Female, n	Female patient age, mean (SD)
Tasmania						
Test	23	70.3 (9.4)	14	70.5 (11.0)	9	69.9 (6.6)
Control	45	72.7 (9.0)	25	73.5 (8.1)	20	71.6 (10.1)
Australian Capital Territory						
Test	11	71.1 (7.9)	8	70.4 (8.4)	3	73.1 (7.4)
Control	13	75.8 (7.8)	7	73.3 (7.4)	6	77.3 (8.4)
Victoria						
Test	6	65.4 (6.1)	3	68.6 (4.5)	3	62.1 (6.4)
Control	1	76.1 (0)	0	—	1	76.1 (0)
Queensland						
Test	21	70.9 (10.5)	12	68.7 (8.9)	9	73.8 (12.4)
Control	16	74.3 (7.5)	10	74.5 (7.9)	6	73.9 (7.6)
Total						
Test	61	70.2 (9.2)	37	69.7 (9.1)	24	70.8 (9.5)
Control	75	73.5 (8.4)	42	73.7 (7.7)	33	73.2 (9.3)
All	136	72.0 (8.9)	79	71.9 (8.6)	57	72.2 (9.4)

For Australian Capital Territory and Victoria, the maximum number of admissions was in spring (43/119, 36.1%) followed by winter (34/119, 28.6%), autumn (21/119, 17.7%), and summer (21/119, 17.7%). Although hospital admissions were not significantly higher in spring compared to winter (*P*=.48), they were significantly higher than for autumn (*P*=.009) and summer (*P*=.01). Furthermore, the longest LOS was during spring (192/478, 40.2%), followed by winter (138/478, 28.9%), summer (75/478, 15.5%), and autumn (73/478, 15.1%). LOS was significantly longer in spring compared to autumn (*P*=.004) and summer (*P*=.01), but not significantly different from winter (*P*=.32), thus matching the results obtained for the number of hospital admissions. No significant differences were observed for Tasmania between seasons (hospital admissions: 107, 109, 114, 113; LOS: 522, 554, 645, 457 for winter, spring, autumn, and summer, respectively). These results are summarized in Table 2.

Obtaining Seasonal Trends

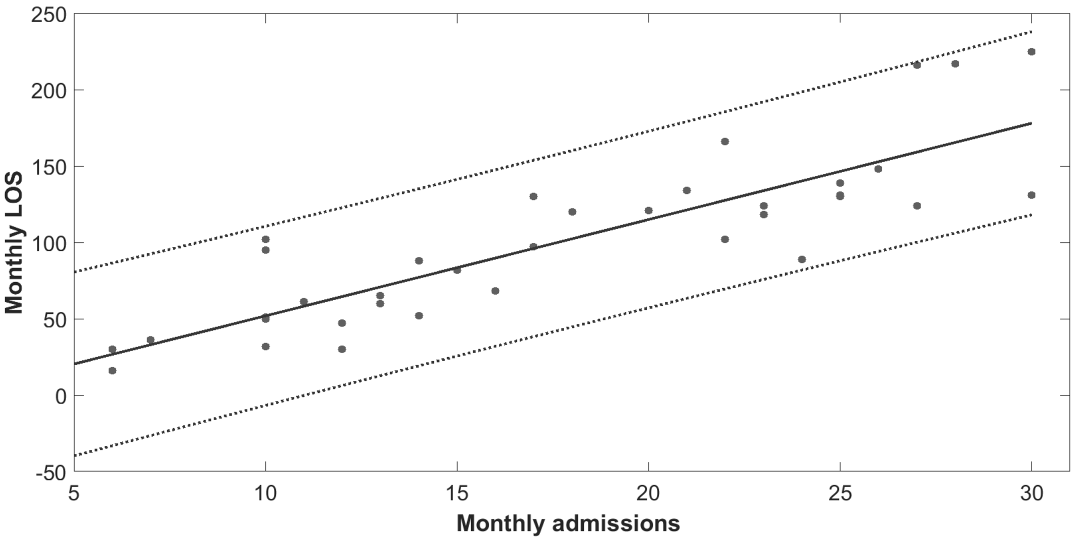
The objective here is to explain the procedure for obtaining seasonal trends. As mentioned earlier, three different climate subgroups were considered: (1) Queensland (subtropical), (2) Australian Capital Territory and Victoria (temperate), and (3) Tasmania (colder). As a result, three subtrends were obtained for the three climate subgroups, accordingly. Note also that due to space constraints, we only show the analysis for one subgroup (Queensland) here. Other seasonal trends for other subgroups were obtained using the same method. Additionally, because there was a strong positive correlation between number of hospital admissions and LOS (Figure 2), we only studied LOS as the outcome variable. The data in Figure 2 suggest that, for these patients, each admission resulted in an average LOS of 6.31 days.

Table 2. Seasonal variation in hospital admissions and length of stay (LOS).

Location	Season, n (%)				Total, n (%)	*P* (winter vs)[a]		
	Summer	Autumn	Winter	Spring		Summer	Autumn	Spring
Tasmania								
Admissions	113 (25.5)	114 (25.7)	107 (24.2)	109 (24.6)	443 (54.2)	.51	.38	.74
LOS	457 (21.0)	645 (29.6)	522 (24.0)	554 (25.4)	2178 (60.0)	.38	.62	.59
Australian Capital Territory and Victoria								
Admissions	21 (17.7)	21 (17.7)	34 (28.6)	43 (36.1)	119 (14.6)	.08	.06	.48
LOS	75 (15.7)	73 (15.3)	138 (28.9)	192 (40.2)	478 (13.2)	.14	.04	.32
Queensland								
Admissions	98 (38.4)	—	157 (61.6)	—	255 (31.2)	.01	—	—
LOS	338 (34.8)	—	633 (65.2)	—	971 (26.8)	.02	—	—

[a] *P* values were calculated using the Wilcoxon rank sum test.

Figure 2. Length of stay (LOS) versus hospital admissions for 136 (test and control) patients in years 2010 to 2012. Correlation coefficient=0.86. Solid line is the linear regression line (slope=6.31, intercept=–11.17, R2=.73), and dotted lines are 95% prediction bounds (slope=4.98, 7.63 and intercept=–36.59, 14.24).



Removing Aging Effects Via Linear Regression Analysis

Figure 3 shows the LOS summed over each calendar month of the 3 years before the intervention for the Queensland subgroup. To estimate and remove the aging effects from the 3 years' of hospitalization data, a linear model including the calendar months of the year as variables (3 years=36 months) was fitted. The solid line in Figure 3 is the linear regression line (LOS: slope=3.93, intercept=28.06 with R^2=.60).

Percentage of Deviation From Baseline

We then replaced the absolute values of LOS with the percentage deviation from the baseline trend line as shown in Figure 4. The seasonal variation trend (Figure 5) for LOS was then derived by averaging the 3 years' of values in Figure 4.

Distribution of Commencement Dates

Figure 6 shows the wide distribution of commencement dates for test patients supplied with a device for the daily monitoring their vital signs.

Influence of Synchronization of Commencement Days on Seasonal Variation

Seasonal annual variation for LOS is shown in Figure 7, calculated from the average trend of the previous 3 years prior to the start of intervention, for the three subgroups (Queensland, Tasmania, and Australian Capital Territory and Victoria). The seasonal variation in LOS shows that the maximum deviation from baseline was 101.7% (SD 42.2%), 60.6% (SD 36.4%), and 158.3% (SD 68.1%) for the Queensland, Tasmania, and Australian Capital Territory and Victoria subgroups, respectively.

Figure 3. Length of stay for 37 (test and control) patients of Queensland subgroup in years 2010 to 2012. Solid line is the linear regression line (slope=1.12, intercept=6.32, R2=.33), and dotted lines are 95% prediction bounds (slope=0.56, 1.67 and intercept=–5.41, 18.06).

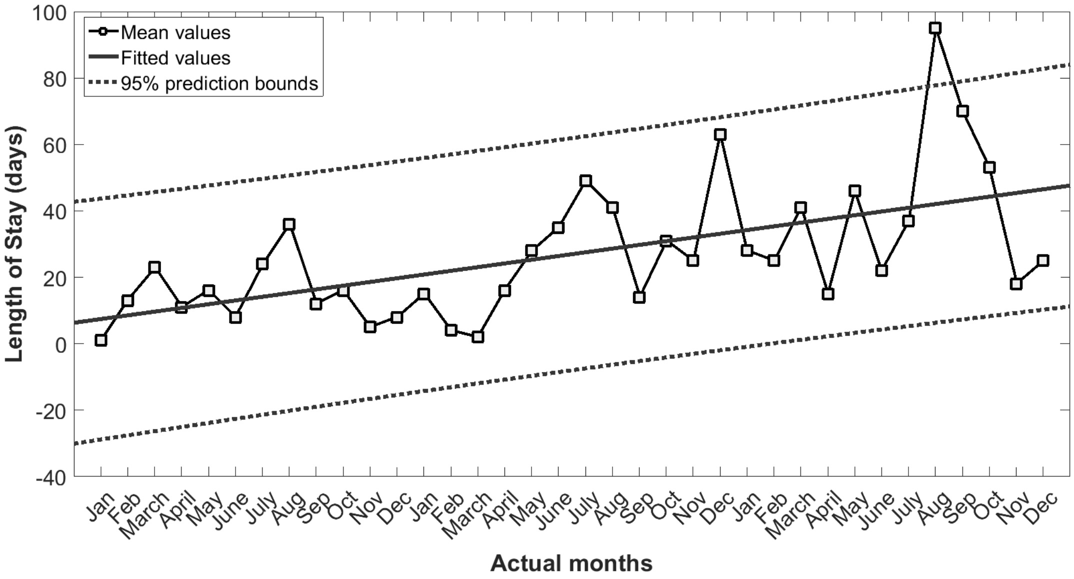


Figure 4. Deviation from baseline (fitted values) in length of stay of 37 (test and control) patients of Queensland subgroup in years 2010 to 2012.

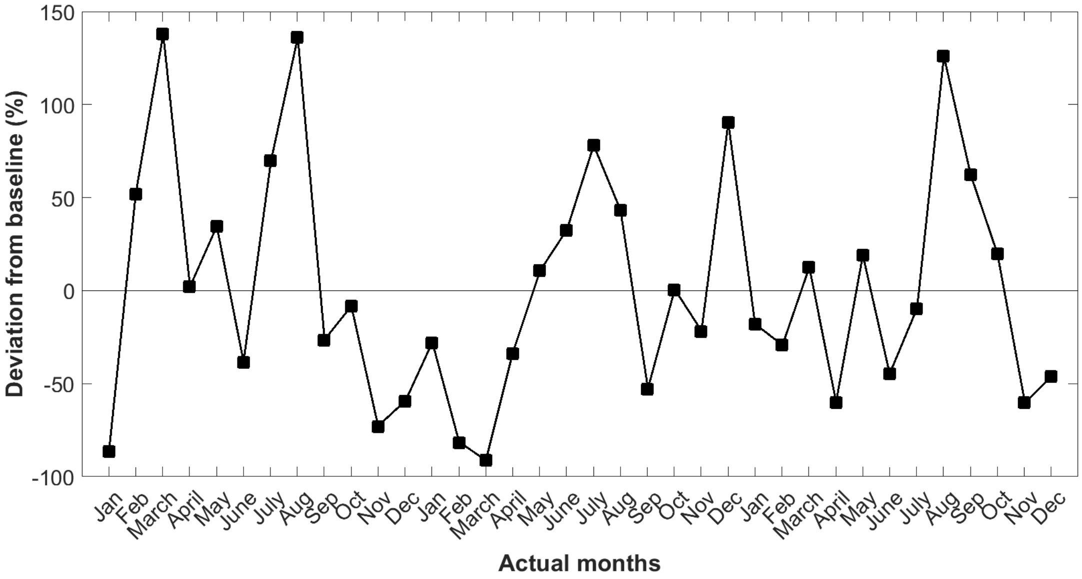


Figure 5. Average deviation from baseline: seasonal variation trend of length of stay in hospital for Queensland patients.

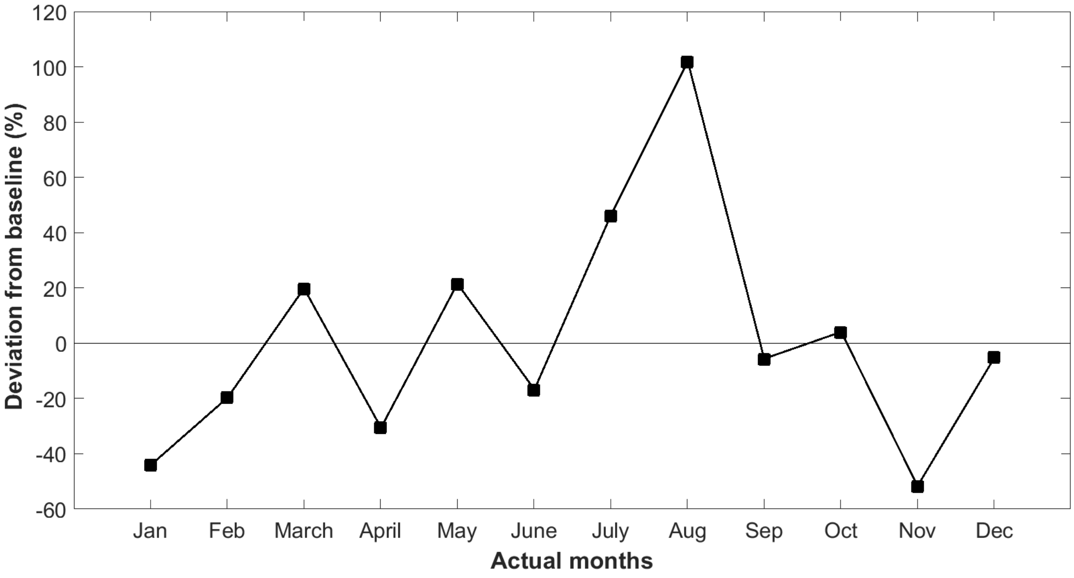
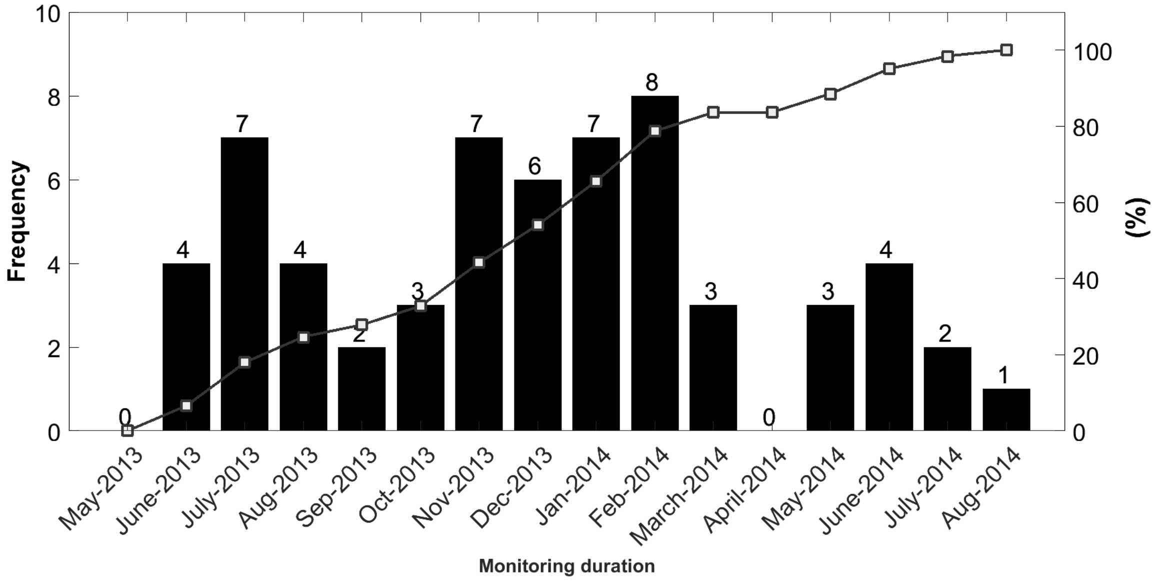


Figure 6. Distribution of commencement dates for monitoring of vital signs.



The synchronized profile shown in Figure 7 was derived by averaging the LOS calculated for the actual calendar month when monitoring began for each of the subsequent 12 calendar months. Note that month 1 after the start of monitoring for one patient could be March, whereas it could be September for another, as shown in Figure 6. Thus, for example, the value of the synchronized profile in the *i* th (*i*=1,...,12) month results from the ratio of the sum of the corresponding values of the obtained seasonal profiles at the first month of monitoring (ie, March or September) for all patients, and the number of patients monitored at the *i* th month. Similarly, this value can be calculated for subsequent months. In summary, the synchronized profile can be obtained by the formula in Figure 8.

In Figure 8, $S_p(i)$ denotes synchronized profile at *i* th month, N_j denotes the number of patients in *j* th (*j*=1,2,3) subgroups (ie, Tasmania, Queensland, and Australian Capital Territory and Victoria), and P_j is the seasonal profile achieved for *j* th subgroups.

As evident from Figure 7, by synchronizing the data to the start of monitoring, the impact of seasonal variation in LOS is greatly reduced to a peak of 9.5% (SD 7.7%), thus minimizing the impact of seasonal variations on the time course of LOS and other output variables.

Let us assume that the recruitment distribution is a Poisson distribution with lambda as the rate (mean) parameter (ie, the average number of patients recruited in a month). To identify lambda from the actual distribution of recruitments in the trial, we used the “poissfit” command of MATLAB, which returns the maximum likelihood estimate of the Poisson distribution, with lambda given by the data. The estimated value of lambda was 6.5.

Figure 7. Estimated seasonal variation impact on length of stay with synchronized commencement days at different trial sites in Australia. QLD: Queensland, TAS: Tasmania, ACT: Australian Capital Territory, VIC: Victoria.

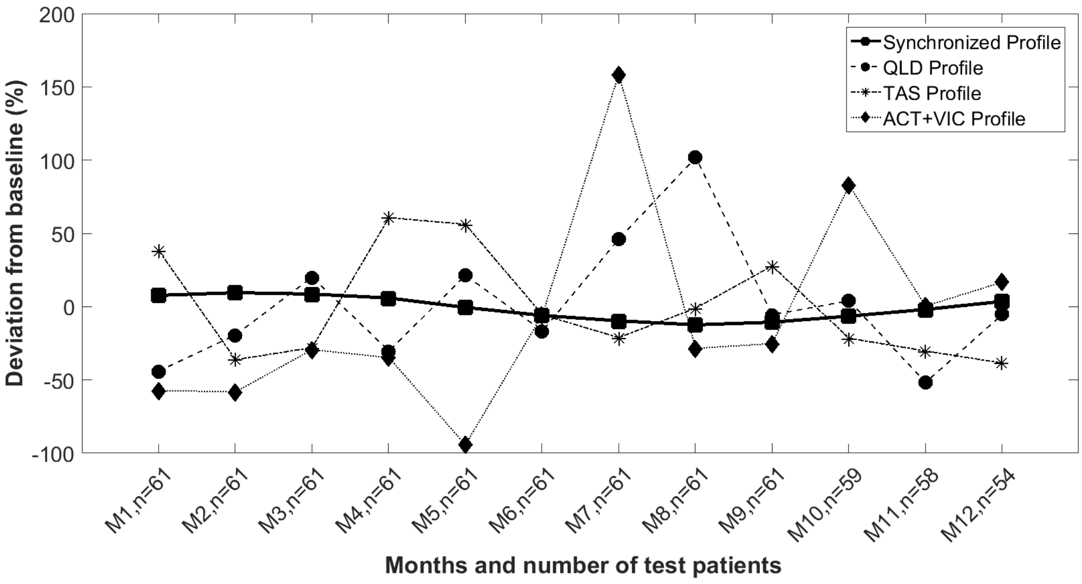


Figure 8. The formula to obtain the synchronized seasonal profile.

$$S_p(i) = \frac{\sum_{j=1}^{3}\sum_{k=1}^{N_j} P_j(Start\ of\ monitoring\ for\ kth\ Patient\ in\ Actual\ Calender + i - 1)}{Number\ of\ Patients\ at\ month\ i\ of\ interventions}$$

Using the lambda estimate, we created 100 sets of random numbers following the Poisson distribution by using *poissrnd(6.5,1,61)*, where 61 was the number of test subjects, and the achieved average maximum deviation from baseline, after synchronizing the analysis of LOS to the commencement of the intervention, was 9.0% (SD 7.5%). These values are quite close to the ones derived when the actual recruitment distribution was used, showing that the actual recruitment distribution is very close to a Poisson distribution.

Let us now assume that the recruitment distribution is a discrete uniform distribution (ie, the recruitment of patients is equally likely to occur during the whole duration of the intervention). Again, we created 100 sets of random numbers following a discrete uniform distribution by using *unidrnd(15,1,61)*, where 15 here is the total length of intervention in month.

Synchronizing the analysis of LOS to the commencement of the intervention, the average maximum deviation from baseline was obtained as 4.9% (SD 3.0%), which is significantly smaller than the one obtained from the previous Poisson distribution. In other words, the impact of seasonal variation on the outcome variables of a telemonitoring trial can be minimized by evenly distributing recruitment over the entire monitoring duration and synchronizing the analysis of outcome variables to the commencement of the intervention.

Discussion

The existence of seasonal variation in incidence of stroke, blood pressure, sudden death, myocardial ischemia, acute myocardial infarction, pulmonary embolism, lung function, and symptoms in COPD has been widely documented [12-17]. Seasonal variation also has an impact on the hospitalization rate of patients with a range of cardiovascular diseases [18,19], as well as acute myocardial infarction and angina in a western Sicily (Italy) hospital [20-21]. From these references, most admissions occur in the winter season for patients with cardiovascular disease from increased hypertension [22], ischemia [23,24], and recurrent infections [25].

The relationship between COPD exacerbation and seasonality has also been investigated in the Towards a Revolution in COPD Health [26] and Prevention of Exacerbations with Tiotropium in COPD [27] trials, both large international studies with more than 13,000 patients. These studies showed an increase in COPD exacerbations as well as an increase in hospitalization rate during the winter months. However, no association was observed in the tropics. This could be due to the fact that respiratory viruses are more prevalent in the cold months of temperate countries [28].

In the CSIRO National Telehealth Trial, a winter-spring predominance was evident in the seasonal variation in hospitalization and LOS both in the overall patient cohort as well as the Queensland and Australian Capital Territory and

Victoria subgroups. This finding is partially in-line with the results of several other studies [12-17] performed in different countries with COPD and congestive heart failure patients showing a peak in winter.

The average LOS per admission was in summer (4.06 days) followed by autumn (4.42 days), spring (4.54 days), and winter (4.66 days). This reveals that an increase in the hospitalization rate coincided with a longer average LOS in cold months.

In Launceston, Tasmania, admissions were below average in spring and summer, increased rapidly in autumn (which coincided with high rainfall periods), and then dropped off again in winter before increasing again quite rapidly as winter ended.

The LOS broadly matches this pattern of admissions except for unexpectedly high LOS and below average rates of admissions in January. This anomaly is explained by local circumstances at Launceston base hospital, where new and inexperienced medical staff arrive in January to replace more experienced clinicians on leave, and have a tendency to keep patients in hospital longer as a precautionary measure.

Townsville in Queensland is a subtropical area with a rainy season in January, February, and March. Admissions were below average during the wet season, but increased rapidly following the end of the wet season, possibly due to increased pollen counts. However, LOS remained fairly static until August, at the end of winter, when they almost doubled. This is difficult to explain because peak average temperatures in winter are around 25°C in Townsville versus 12°C in Tasmania.

This study confirms the existence of a significant seasonal variation in hospital admissions as well as LOS in a recently completed CSIRO national trial of home telemonitoring of patients with chronic conditions, carried out at five locations along the east coast of Australia. Climactic and environmental conditions can also change year by year as shown in Figures 3 and 4, making the analysis of seasonal impacts more difficult to interpret.

We have shown that by synchronizing analysis of outcomes to the start of the intervention, the effect of seasonal variation on clinical outcome of an at-home telemonitoring trial can be significantly attenuated. To the authors’ knowledge, this is the first study to introduce a method to attenuate the effect of seasonal variations on the time course of outcome variables by synchronizing the analysis to the start of intervention for each patient.

This method recognizes that difficulties of recruitment of test patients in clinical trials are common, and patients may often be recruited over many months. This study suggests that by evenly distributing recruitment over the intervention duration and synchronizing the analysis of outcome variables to the commencement of the intervention, the confounding impact of seasonal variations can be minimized.

Acknowledgments

This project was funded by the Australian Commonwealth Government National Broadband Network Telehealth Pilot Program with additional financial support from the CSIRO and in-kind contributions from health service providers at the five trial sites. These health service organizations include Australian Capital Territory (Department of Health), Townsville-MacKay Medicare Local, Launceston Hospital, Nepean Blue Mountains Local Health District, and Djerriwarrh Health services. Industry and research partners, Telemedcare Pty Ltd, Samsung Australia Pty Ltd, and iiNET Pty Ltd, are also acknowledged. Additional support for ongoing data analysis was provided through an Australian Research Council Linkage Grant (LP160101202).

Authors' Contributions

All authors made a significant contribution to data analysis and the drafting of the manuscript.

Conflicts of Interest

There was no conflict of interest during the planning and execution of the project. Six months after its completion, BGC, Chief Investigator and Project Director, was appointed to a part-time position at Telemedcare Pty Ltd as Director of Research.

References

1. Bashshur RL, Shannon GW, Smith BR, Alverson DC, Antoniotti N, Barsan WG, et al. The empirical foundations of telemedicine interventions for chronic disease management. Telemed J E Health 2014 Sep;20(9):769-800 [FREE Full text] [doi: 10.1089/tmj.2014.9981] [Medline: 24968105]
2. Bashshur RL, Howell JD, Krupinski EA, Harms KM, Bashshur N, Doarn CR. The empirical foundations of telemedicine interventions in primary care. Telemed J E Health 2016 May;22(5):342-375 [FREE Full text] [doi: 10.1089/tmj.2016.0045] [Medline: 27128779]
3. Steventon A, Bardsley M, Billings J, Dixon J, Doll H, Hirani S, Whole System Demonstrator Evaluation Team. BMJ. 2012 Jun 21. Effect of telehealth on use of secondary care and mortality: findings from the Whole System Demonstrator cluster randomised trial URL: https://sso.lib.uts.edu.au/cas/login?service=https%3A%2F%2Fwww.lib.uts.edu.au%2Fgoto%3Fqurl%3Dhttps%253a%252f%252fdoi.org%252f10.1136%252fbmj.e3874%26_casCheck%3Dtrue [accessed 2018-02-26] [WebCite Cache ID 6xWt9AoOR]
4. Brown EM. The Ontario Telemedicine Network: a case report. Telemed J E Health 2013 May;19(5):373-376. [doi: 10.1089/tmj.2012.0299] [Medline: 23301768]

5. Paré G, Jaana M, Sicotte C. Systematic review of home telemonitoring for chronic diseases: the evidence base. J Am Med Inform Assoc 2007 May;14(3):269-277 [FREE Full text] [doi: 10.1197/jamia.M2270] [Medline: 17329725]
6. Maeder A, Poultney N, Morgan G, Lippiatt R. Patient compliance in home-based self-care telehealth projects. J Telemed Telecare 2015 Dec;21(8):439-442. [doi: 10.1177/1357633X15612382] [Medline: 26556057]
7. Shany T, Hession M, Pryce D, Roberts M, Basilakis J, Redmond S, et al. A small-scale randomised controlled trial of home telemonitoring in patients with severe chronic obstructive pulmonary disease. J Telemed Telecare 2017 Aug;23(7):650-656. [doi: 10.1177/1357633X16659410] [Medline: 27464957]
8. Celler B, Varnfield M, Nepal S, Sparks R, Li J, Jayasena R. Impact of at-home telemonitoring on health services expenditure and hospital admissions in patients with chronic conditions: before and after control intervention analysis. JMIR Med Inform 2017 Sep 08;5(3):e29 [FREE Full text] [doi: 10.2196/medinform.7308] [Medline: 28887294]
9. Celler BG, Sparks R, Nepal S, Alem L, Varnfield M, Li J, et al. Design of a multi-site multi-state clinical trial of home monitoring of chronic disease in the community in Australia. BMC Public Health 2014 Dec 15;14:1270 [FREE Full text] [doi: 10.1186/1471-2458-14-1270] [Medline: 25511206]
10. Nepal S, Jang-Jaccard J, Celler B, Yan B, Alem L. Data architecture for Telehealth services research: a case study of home tele-monitoring. 2013 Oct 20 Presented at: 9th International Conference Conference on Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom); Oct 20, 2013; Austin, TX p. 458-467. [doi: 10.4108/icst.collaboratecom.2013.254220]
11. Sparks R, Celler B, Okugami C, Jayasena R, Varnfield M. Telehealth monitoring of patients in the community. J Intell Syst 2016;25(1):37-53. [doi: 10.1515/jisys-2014-0123]
12. Tsementzis SA, Gill JS, Hitchcock ER, Gill SK, Beevers DG. Diurnal variation of and activity during the onset of stroke. Neurosurgery 1985 Dec;17(6):901-904. [Medline: 4080122]
13. Brennan PJ, Greenberg G, Miall WE, Thompson SG. Seasonal variation in arterial blood pressure. Br Med J (Clin Res Ed) 1982 Oct 02;285(6346):919-923 [FREE Full text] [Medline: 6811068]
14. Muller JE, Ludmer PL, Willich SN, Tofler GH, Aylmer G, Klangos I, et al. Circadian variation in the frequency of sudden cardiac death. Circulation 1987 Jan;75(1):131-138. [Medline: 3791599]
15. Rocco MB, Barry J, Campbell S, Nabel E, Cook EF, Goldman L, et al. Circadian variation of transient myocardial ischemia in patients with coronary artery disease. Circulation 1987 Feb;75(2):395-400. [Medline: 3802443]
16. Manfredini R, Gallerani M, Boari B, Salmi R, Mehta RH. Seasonal variation in onset of pulmonary embolism is independent of patients' underlying risk comorbid conditions. Clin Appl Thromb Hemost 2004 Jan;10(1):39-43. [Medline: 14979403]
17. Almagro P, Hernandez C, Martinez-Cambor P, Tresserras R, Escarrabill J. Seasonality, ambient temperatures and hospitalizations for acute exacerbation of COPD: a population-based study in a metropolitan area. Int J Chron Obstruct Pulmon Dis 2015;10:899-908 [FREE Full text] [doi: 10.2147/COPD.S75710] [Medline: 26056439]
18. Martínez-Sellés M, García RJ, Prieto L, Serrano JA, Muñoz R, Frades E, et al. Annual rates of admission and seasonal variations in hospitalizations for heart failure. Eur J Heart Fail 2002 Dec;4(6):779-786 [FREE Full text] [Medline: 12453550]
19. Gallerani M, Boari B, Manfredini F, Manfredini R. Seasonal variation in heart failure hospitalization. Clin Cardiol 2011 Jun;34(6):389-394 [FREE Full text] [doi: 10.1002/clc.20895] [Medline: 21538387]
20. Abrignani MG, Corrao S, Biondo GB, Renda N, Braschi A, Novo G, et al. Influence of climatic variables on acute myocardial infarction hospital admissions. Int J Cardiol 2009 Oct 02;137(2):123-129. [doi: 10.1016/j.ijcard.2008.06.036] [Medline: 18694607]
21. Abrignani MG, Corrao S, Biondo GB, Lombardo RM, Di Girolamo P, Braschi A, et al. Effects of ambient temperature, humidity, and other meteorological variables on hospital admissions for angina pectoris. Eur J Prev Cardiol 2012 Jun;19(3):342-348. [doi: 10.1177/1741826711402741] [Medline: 21450571]
22. Youn J, Rim S, Park S, Ko Y, Kang S, Choi D, et al. Arterial stiffness is related to augmented seasonal variation of blood pressure in hypertensive patients. Blood Press 2007;16(6):375-380. [doi: 10.1080/08037050701642618] [Medline: 18058455]
23. Danet S, Richard F, Montaye M, Beauchant S, Lemaire B, Graux C, et al. Unhealthy effects of atmospheric temperature and pressure on the occurrence of myocardial infarction and coronary deaths. A 10-year survey: the Lille-World Health Organization MONICA project (Monitoring trends and determinants in cardiovascular disease). Circulation 1999 Jul 06;100(1):E1-E7 [FREE Full text] [Medline: 10393689]
24. Spencer FA, Goldberg RJ, Becker RC, Gore JM. Seasonal distribution of acute myocardial infarction in the second National Registry of Myocardial Infarction. J Am Coll Cardiol 1998 May;31(6):1226-1233 [FREE Full text] [Medline: 9581712]
25. Stewart S, McIntyre K, Capewell S, McMurray JJ. Heart failure in a cold climate. Seasonal variation in heart failure-related morbidity and mortality. J Am Coll Cardiol 2002 Mar 06;39(5):760-766 [FREE Full text] [Medline: 11869838]
26. Jenkins CR, Celli B, Anderson JA, Ferguson GT, Jones PW, Vestbo J, et al. Seasonality and determinants of moderate and severe COPD exacerbations in the TORCH study. Eur Respir J 2012 Jan;39(1):38-45 [FREE Full text] [doi: 10.1183/09031936.00194610] [Medline: 21737561]
27. Rabe KF, Fabbri LM, Vogelmeier C, Kögler H, Schmidt H, Beeh KM, et al. Seasonal distribution of COPD exacerbations in the Prevention of Exacerbations with Tiotropium in COPD trial. Chest 2013 Mar;143(3):711-719. [doi: 10.1378/chest.12-1277] [Medline: 23188489]

28. Wedzicha JA, Seemungal TA. COPD exacerbations: defining their cause and prevention. Lancet 2007 Sep 1;370(9589):786-796. [doi: 10.1016/S0140-6736(07)61382-8] [Medline: 17765528]

Abbreviations

COPD: chronic obstructive pulmonary disease
CSIRO: Commonwealth Scientific and Industrial Research Organization
HREC: Human Research Ethics Committee
LOS: length of stay

Edited by M Focsa; submitted 18.12.17; peer-reviewed by E Krupinski, R Pankomera; comments to author 28.12.17; revised version received 02.01.18; accepted 04.01.18; published 16.03.18.

Please cite as:
Argha A, Savkin A, Liaw ST, Celler BG
Effect of Seasonal Variation on Clinical Outcome in Patients with Chronic Conditions: Analysis of the Commonwealth Scientific and Industrial Research Organization (CSIRO) National Telehealth Trial
JMIR Med Inform 2018;6(1):e16
URL: http://medinform.jmir.org/2018/1/e16/
doi:10.2196/medinform.9680
PMID:29549068

Original Paper

Automating Quality Measures for Heart Failure Using Natural Language Processing: A Descriptive Study in the Department of Veterans Affairs

Jennifer Hornung Garvin[1,2,3,4,5], MBA, PhD; Youngjun Kim[2,6], MS; Glenn Temple Gobbel[7,8], PhD, DVM; Michael E Matheny[7,8], MPH, MD, MS; Andrew Redd[2,4], PhD; Bruce E Bray[2,3], MD; Paul Heidenreich[9], MD, MS; Dan Bolton[2,4], PhD; Julia Heavirland[2], MA; Natalie Kelly[2], MBA; Ruth Reeves[7,8], PhD; Megha Kalsy[2,3], PhD; Mary Kane Goldstein[10,11], MD, MS; Stephane M Meystre[2,6], MD, PhD

[1]Health Information Management and Systems Division, School of Health and Rehabilitation Sciences, The Ohio State University, Columbus, OH, United States

[2]IDEAS 2.0 Health Services Research and Development Research Center, Salt Lake City Veterans Affairs Healthcare System, Department of Veterans Affairs, Salt Lake City, UT, United States

[3]Department of Biomedical Informatics, School of Medicine, University of Utah, Salt Lake City, UT, United States

[4]Division of Epidemiology, Department of Medicine, University of Utah, Salt Lake City, UT, United States

[5]Geriatric Research, Education and Clinical Center, Salt Lake City Veterans Affairs Healthcare System, Department of Veterans Affairs, Salt Lake City, UT, United States

[6]Translational Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC, United States

[7]Geriatric Research, Education and Clinical Center, Tennessee Valley Healthcare System, Department of Veterans Affairs, Nashville, TN, United States

[8]Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, TN, United States

[9]Palo Alto Geriatric Research, Education and Clinical Center, Veterans Affairs Palo Alto Health Care System, Department of Veterans Affairs, Stanford University, Palo Alto, CA, United States

[10]Medical Service, Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, United States

[11]Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States

Corresponding Author:
Jennifer Hornung Garvin, MBA, PhD
Health Information Management and Systems Division
School of Health and Rehabilitation Sciences
The Ohio State University
453 W 10th Ave
Columbus, OH, 43210
United States
Phone: 1 2156203390
Email: jennifer.garvin@hsc.utah.edu

Abstract

Background: We developed an accurate, stakeholder-informed, automated, natural language processing (NLP) system to measure the quality of heart failure (HF) inpatient care, and explored the potential for adoption of this system within an integrated health care system.

Objective: To accurately automate a United States Department of Veterans Affairs (VA) quality measure for inpatients with HF.

Methods: We automated the HF quality measure Congestive Heart Failure Inpatient Measure 19 (CHI19) that identifies whether a given patient has left ventricular ejection fraction (LVEF) <40%, and if so, whether an angiotensin-converting enzyme inhibitor or angiotensin-receptor blocker was prescribed at discharge if there were no contraindications. We used documents from 1083 unique inpatients from eight VA medical centers to develop a reference standard (RS) to train (n=314) and test (n=769) the Congestive Heart Failure Information Extraction Framework (CHIEF). We also conducted semi-structured interviews (n=15) for stakeholder feedback on implementation of the CHIEF.

Results: The CHIEF classified each hospitalization in the test set with a sensitivity (SN) of 98.9% and positive predictive value of 98.7%, compared with an RS and SN of 98.5% for available External Peer Review Program assessments. Of the 1083 patients available for the NLP system, the CHIEF evaluated and classified 100% of cases. Stakeholders identified potential implementation facilitators and clinical uses of the CHIEF.

Conclusions: The CHIEF provided complete data for all patients in the cohort and could potentially improve the efficiency, timeliness, and utility of HF quality measurements.

(JMIR Med Inform 2018;6(1):e5) doi:10.2196/medinform.9150

KEYWORDS

natural language processing (NLP); heart failure; left ventricular ejection fraction (EF); informatics; quality measures

Introduction

Heart failure (HF) is associated with substantial morbidity, mortality, and consumption of medical resources. HF affects approximately five million Americans and is the number one reason for discharge for Veterans treated within the United States Department of Veterans Affairs (VA) health care system [1,2]. The cost of HF care is high, and will remain a significant concern for the US health care system with high mortality; 50% of Medicare beneficiaries do not survive three years after an HF hospitalization [3,4].

The cost of treating HF in the United States is estimated to increase from US $31 billion in 2012 to US $70 billion by 2030 [5-7]. Despite decreased HF hospitalizations between 2001 and 2009, the presence of HF as a secondary condition in hospitalizations increased over the same period [7], with research suggesting that 55% of acute exacerbations were preventable [8]. HF was the fourth most common diagnosis for hospitalization in 2014 [9] and prevalence figures indicate that 6.6 million American adults 18 years of age or older (2.8%) have HF [10]. It is estimated that an additional 3 million adults (25% increase) will be diagnosed with HF by 2030 [3,5], and it is important to implement evidence-based, guideline-concordant care that can improve HF symptoms, prolong life, and reduce readmissions [3,6,11-15].

The VA HF quality measure known as Congestive Heart Failure Inpatient Measure 19 (CHI19) describes how often guideline-concordant medical therapy, in the form of angiotensin-converting enzyme inhibitor (ACEI) or angiotensin-receptor blocker (ARB) use, is provided for patients with left ventricular ejection fraction (LVEF) of <40% at the time of discharge, unless there are contraindications. The same information is currently collected for outpatients using the Congestive Heart Failure Outpatient Measure 7 (CHF7): HF-Outpatient Left Ventricular Failure (LVF) documented and Congestive Heart Failure Outpatient Measure 14 (CHF14): HF-Outpatient LVEF Less Than 40 on ACEI or ARB measures. The measurement of this information is used for accountability within the VA. The use of these measures provides key feedback to patients (through public reporting), providers, and local or regional areas, including the VA's Veterans Integrated Service Networks [16,17]. The measures used by the VA are in alignment with Medicare and are reported publicly [18].

Our primary goal was to develop an efficient and accurate method of obtaining quality data by automating the CHI19 measure, as it is an accountability measure that has been widely used in the VA for many years, and currently requires time-consuming chart abstraction to determine through the External Peer Review Program (EPRP). EPRP provides peer review for the VA through an external medical professional association that abstracts the charts manually to populate a dashboard [19]. Additional HF measures abstracted by EPRP include Congestive Heart Failure Inpatient Measure 10 (CHI10) and Congestive Heart Failure Inpatient Measure 20 (CHI20). CHI10 refers to HF patients who were assessed for LVF at discharge or patients for whom such an assessment was planned, whereas CHI20 refers to patients who had LVEF <40% and were taking an ACEI or ARB before being admitted as inpatients.

Using automated methods to share data and measure quality for provider feedback and public reporting is a key goal of the incentives provided by the Centers for Medicare and Medicaid Services, so that certified electronic health records (EHRs) of “meaningful use” criteria can be attained [20]. Some quality measures that use only structured data from the EHR are relatively easy to automate. A challenge for automating the computation of CHI10, CHI19, and CHI20 is that, unlike quality measures that use only structured data [21], these measures require data regarding LVEF and contraindications to medications, which in the VA are primarily in free-text health record documents and are therefore more difficult to extract.

Prior research in informatics in VA showed that health information technology and the use of explicit conceptual models can not only contribute to increasing well-formed and well-grounded health informatics research [22], but can also facilitate evidence-based practice [23] through usability testing, good research design, and implementation methodology [24]. Importantly, prior research indicates that end-user considerations, including where and when the technology is required as well as stakeholder needs and goals, must be identified for successful implementation [25-30]. To this end, we initiated development of an automated natural language processing (NLP) system capable of efficient data capture that could meet end-user needs and generate data for other informatics applications, such that the system would be positioned for adoption and implementation by the VA or other health care organizations.

Methods

Setting and Context

For the system's clinical basis, we used the American Heart Association/American College of Cardiology level 1A clinical evidence, which recommends assessing the left ventricular systolic function and use of ACEI or ARB if the ejection fraction (EF) is <40%, if there are no contraindications [6,31]. We used the VA Informatics and Computing Infrastructure [32] for NLP development and analysis of EHR patient data from the VA's Corporate Data Warehouse (CDW) [33].

Patient Cohort Identification and Document and Structured Data Acquisition

We obtained a listing of EPRP abstracted cases involving HF patients discharged from eight VA medical centers. To approximate the general VA patient population, we selected facilities which in total were representative of the VA population in terms of race, ethnicity, and rurality in the fiscal year 2008 to serve as our study cohort. The patient cohort was randomized and split into training and test sets based on the sampling strategy described below. We obtained the associated text integration utilities (TIUs) notes for each patient. The TIUs software processes free-text clinical notes so they can be saved in the Veterans Health Information Systems and Technology Architecture files. We also obtained structured data from the Pharmacy Benefit Management (PBM) software to determine each patient's medications, and International Classification of Diseases, 9thRevision, Clinical Modification (ICD-9-CM) codes, and laboratory data to identify reasons the patient was not prescribed medications (reasons no medications; RNM) for each patient in the cohort. Acquiring these data allowed for comparison of the concepts found in free text through NLP with VA structured data for determination of each patient's medications or RNM.

Sampling Strategy for Natural Language Processing Development

We used a power analysis that accounted for differences in the prevalence of clinical concepts within notes across the medical centers. We selected the sample size that involved the largest number of patients to determine the test set, in order to accommodate the rarest event (contraindications to ACEIs and ARBs) which was estimated to be 14.9% based on the HF literature [34]. We determined a sample size of 769 patients for the test set for system performance evaluation, and the remaining patients in the EPRP abstraction set (n=314) served as a separate set for training the NLP system.

Reference Standard Development for Natural Language Processing Development

We used Knowtator Protégé plug-in software [35] to annotate the training and evaluation (test) sets, to create a Reference Standard (RS) to undertake an accurate performance evaluation [36] of the NLP system at both the concept (eg, EF, medications, RNM) and patient (eg, overall determination or classification of a patient meeting the equivalent of CHI19) levels. We developed annotation guidelines that provided explicit examples of concepts (data) to be identified, which documents were preferred for each concept (eg, most recent echocardiogram for EF, and discharge medication reconciliation form for ACEIs and ARBs), annotation at the document level, and how to use the document-level annotations to determine the patient classification with resulting patient-level annotation [37]. We annotated 100% of the unique patients in our cohort for NLP training and testing. Two annotators independently reviewed the text documents. We measured percent agreement between the annotators across all concepts. The patient- and document-level annotations, as well as differences between concept-level annotations, were resolved via consensus determination by the two annotators with assistance from a subject matter expert (SME) cardiologist who was part of the study team. The annotators were required to achieve 90% interannotator agreement (IAA) at the concept level, and were assessed for accuracy before annotating the RS. A cardiologist (SME) reviewed and adjudicated differences when needed. We created the final RS after all differences were resolved. All cases were successfully classified by the annotators with cardiology oversight.

We used two software tools to assist annotators by preannotating concepts for subsequent verification. The first software tool, based on the Apache Unstructured Information Management Architecture (UIMA) framework [38,39], was designated *Capture with UIMA of Needed Data using Regular Expressions for Ejection Fraction* [40] and used to preannotate EF information. The second tool, the Extensible Human Oracle Suite of Tools [41], was used to preannotate ACEI/ARB medications. Preannotated concepts were read into the Knowtator software for annotators to review and finalize. Annotators reviewed preannotations as well as all other information in the document, based on the annotation guidelines.

Natural Language Processing System Development for Information Extraction

We based target concepts for NLP development on clinical guidelines, VA policy, and what was currently collected manually through the EPRP process [6,31]. These target concepts also served as elements in an algorithm for calculating VA CHI19 at the time of discharge. We developed an application called the Congestive Heart Failure Information Extraction Framework (CHIEF) [42-44], based on the Apache UIMA framework, to provide robustness and scalability [38]. As depicted in Figure 1, the CHIEF includes modules for (1) clinical text preprocessing (eg, detecting sentences and tokens as well as conducting syntactic analyses), (2) extracting mentions of EF as well as quantitative values, and (3) extracting mentions of medications (eg, ACEIs and ARBs). RNM were extracted with another NLP application called RapTAT [45], and the resulting data were integrated into the CHIEF.

Figure 1. Congestive heart failure information extraction framework (CHIEF).

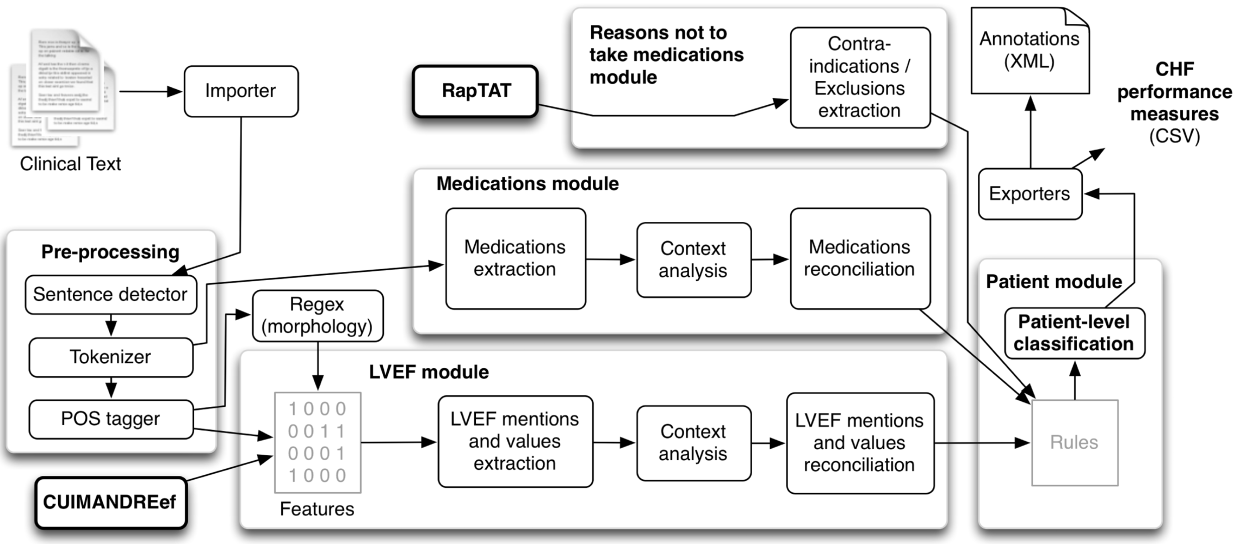
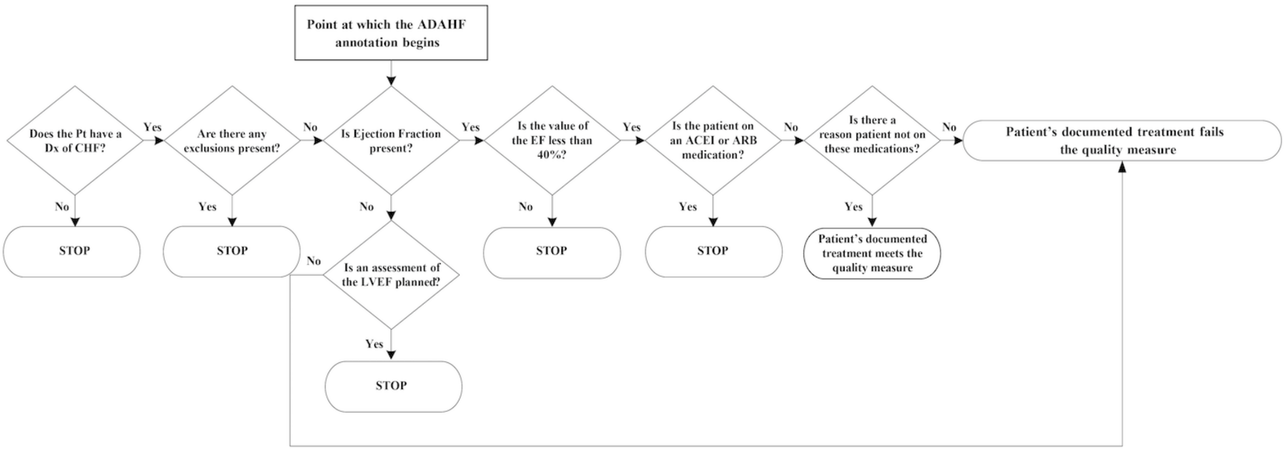


Figure 2. Algorithm to classify the patient as meeting the measure.


Point at which the ADAHF annotation begins
Does the Pt have a Dx of CHF?
Yes
No
STOP
Are there any exclusions present?
No
Yes
STOP
Is Ejection Fraction present?
Yes
No
Is an assessment of the LVEF planned?
No
Yes
STOP
Is the value of the EF less than 40%?
Yes
No
STOP
Is the patient on an ACEI or ARB medication?
No
Yes
STOP
Is there a reason patient not on these medications?
No
Yes
Patient's documented treatment meets the quality measure
Patient's documented treatment fails the quality measure

Finally, all extracted information at the concept and patient levels was compared and combined using a set of rules to classify HF treatment performance measures automatically for each patient (see Figure 2). For example, the NLP system assessed whether the patient had an EF recorded (and if the answer was yes, was it <40%?). If the EF was <40%, then the system determined if the patient was on an ACEI or ARB. If the patient was not, then the system determined if there were RNM. The patient met the measure: if the EF was present but not <40%; if the EF was <40% and there was an active prescription for an ACEI or ARB; or if the patient had an EF <40%, was not on an ACEI or ARB, and had RNM (see Figure 2).

Key Informant and Subject Matter Expert Interviews

To inform the design of our automated system and to facilitate adoption, we interviewed both key informants and SMEs. We conducted 15 interviews consisting of four key informants and a convenience sample of 11 additional SMEs. The four key informants that were interviewed were VA quality measurement experts with national roles and VA-wide knowledge about inpatient HF quality measurements and the use of quality measurements for HF in the VA. Based on a snowball sampling design, the key informants recommended the inclusion of 11 additional VA-based SMEs. We recruited and interviewed these SMEs, who were responsible for receiving and interpreting quality monitoring data, and included VA cardiologists and HF quality experts with extensive experience in making decisions regarding the quality measures to be used and presentation of the results of quality assessments. The key informants' and SMEs' experience in the VA ranged from 2 to 35 years, and from 2 to 33 years in quality management.

To develop our interview questions, we drew upon the Promoting Action on Research Implementation in Health Sciences (PARiHS) framework [46,47], which postulates that evidence, context, and facilitation are central to implementation. We complemented the PARiHS framework with the Socio-Technical Model of Health Information Technology to focus on the information technology context of potential implementation [48]. We studied the potential of integrating an automated quality measurement system in the VA through these interviews, and will detail our applied thematic analysis in a future manuscript.

Measurements

We compared the CHIEF system output to the human annotator-created RS to compute performance at the concept level and for the patient-level binary classification of meeting or not meeting the CHI19 measure. We calculated sensitivity (SN), specificity (SP), and positive predictive value (PPV) in addition to the F-measure, which is the harmonic mean of the SN and PPV [49]. We also computed the SN of the NLP test set based on the results of the EPRP review at the patient level for target concepts, and the overall binary classification of meeting the CHI19 measure. We computed Cohen's kappa [50] parameter to determine concordance between the structured prescription data from the PBM package to determine patient medications, and both the human-annotated RS and the NLP output. Similarly, we compared ICD-9-CM codes and laboratory results to both the human-annotated RS and the NLP output to find RNM. We then summarized the interview findings to complement the system development.

Institutional Review Board Approval

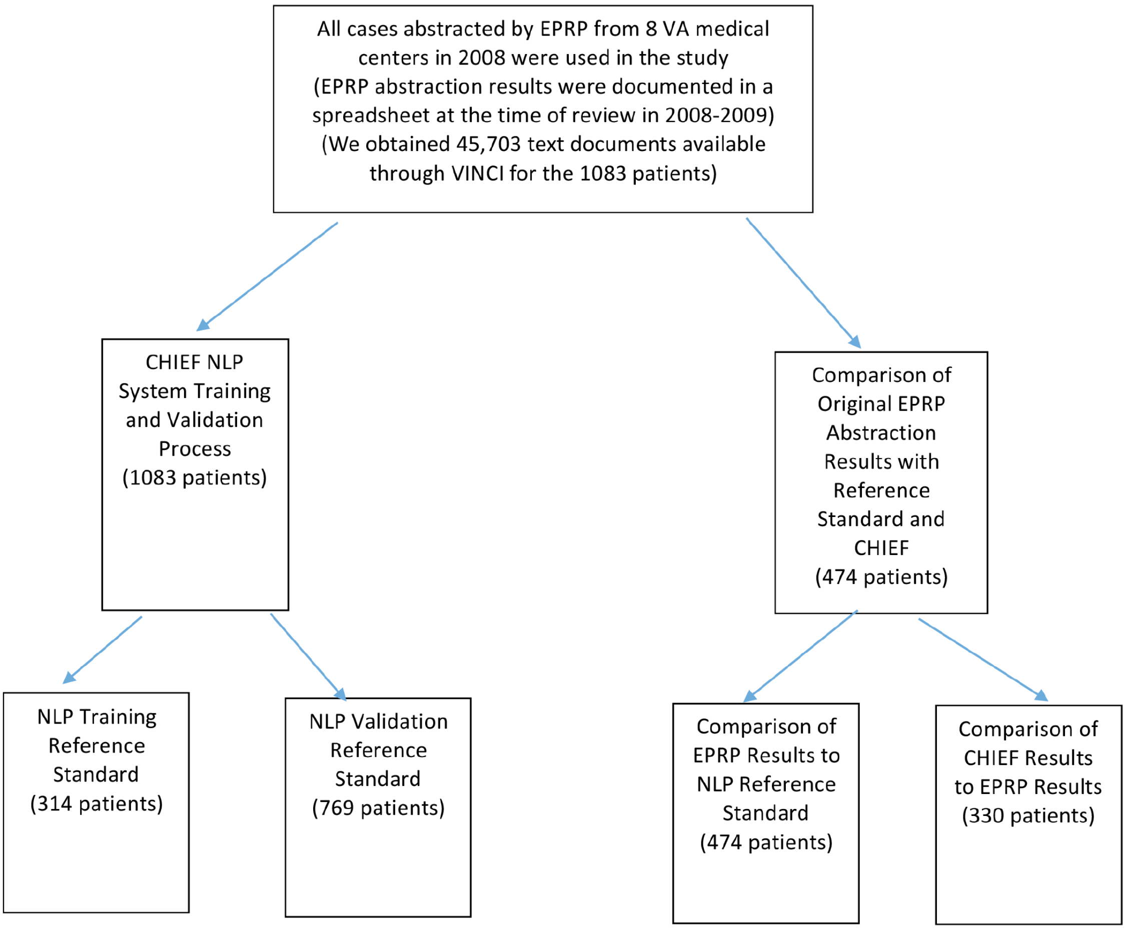
This study was approved by the University of Utah and the Tennessee Valley Healthcare System Institutional Review Boards (IRBs). Informed patient consent was waived for text document use. The IRB approved informed consent with a waiver of documentation of consent for the key informant and SME interviews.

Results

Documents Obtained for the Research

We retrieved 45,703 free-text (TIU) documents from 1083 patients (314 in the training set and 769 in the test set). Using a systematic sample (every tenth document), we mapped the document title names to the following documents types in our corpus: history and physical, progress notes, cardiology consult, echocardiogram, pharmacy (medication reconciliation), pharmacy (other), consult (other), discharge summary, nursing note, and other (general). After mapping and during annotation, we found that EF was most commonly found in the assessment and current history sections of any note in which these sections were used (eg, history and physicals, progress notes, cardiology consults). Medications (ACEI/ARB) were most commonly found in the assessment and medication sections, LVEF was most commonly found in the echocardiogram results and assessment sections, and RNM were most commonly found in the assessment section [51]. Please see Figure 3 for the data capture strategy used in the research.

Figure 3. Data capture strategy.


All cases abstracted by EPRP from 8 VA medical centers in 2008 were used in the study
(EPRP abstraction results were documented in a spreadsheet at the time of review in 2008-2009)
(We obtained 45,703 text documents available through VINCI for the 1083 patients)
CHIEF NLP System Training and Validation Process (1083 patients)
Comparison of Original EPRP Abstraction Results with Reference Standard and CHIEF (474 patients)
NLP Training Reference Standard (314 patients)
NLP Validation Reference Standard (769 patients)
Comparison of EPRP Results to NLP Reference Standard (474 patients)
Comparison of CHIEF Results to EPRP Results (330 patients)

Reference Standard Development

The IAA was found to be 91% in a pairwise comparison of all concepts within the documents in the project corpus for the RS [52]. In consultation with the cardiologist as needed, the annotation team was able to agree on this consensus RS for all patient-level classifications.

Performance of the Natural Language Processing System

We developed the CHIEF NLP system (see Figure 1). When evaluated at the information extraction level, CHIEF extracted relevant mentions and values of LVEF, medications, and RNM with a range of *high* to *fair* recall. HF medications were extracted with recall of 97.8-99.7% and precision of 96-97.8%; mentions of LVEF were extracted with recall of 97.8-98.6% and precision of 98.6-99.4%, and RNM were less common and more difficult to extract, only reaching fair accuracy with 31.1-40.4% recall and 24.7-32.1% precision [53]. As explained earlier, this extracted information was then combined at the patient level using a set of rules. At the patient level, as shown in Table 1 and compared with the RS, CHIEF achieved almost 99% F-measure, SN, and PPV for classifying a patient admission as meeting the CHI19 measure. The SP was 85.5%. The CHIEF could also classify whether the performance measure was met, assess LVEF, and determine whether the EF was below 40%, with F-measures of 98.80%, 99.52%, and 95.52%, respectively. However, the system identified more false positives in medication prescriptions (PPV, <90%; F-measure, 93.53%). For concept extraction, we used machine learning-based approaches (fully automated) rather than rule-based or just keyword searching, and for classification we used several rules (as depicted in Figure 2). Note that for classification we programmed three sets of rules based on decisions depicted in Figure 2. The full set of rules can be found in Multimedia Appendix 1.

The lowest performance of the NLP system was measured for RNM (ie, ACEIs or ARBs), with an SN of 26.9%, SP of 99.4%, and PPV of 90.7%. This performance level was affected by RNM being the least structured, most varied, and least common of all of the concepts evaluated, with only 145 patients in our testing corpus having RNM. When we restricted our analysis to patients with an EF <40% (according to the RS) who were discharged without prescriptions for ACEIs or ARBs (n=70), performance increased slightly (SN, 33%; SP, 92.3%; PPV, 95%; and F-measure, 49%). However, when evaluating hospitalizations for which it was critical for the system determination for patient classification (n=37; EF <40% and discharged without prescriptions according to system output), the RS found more RNM than the NLP system found, but the SN of the system increased to 78%. The PPV was relatively unchanged at 81.8%, for an F-measure of 80%.

Concordance of Reference Standard and External Peer Review Program Results

Table 2 provides a comparison of the human-annotated RS and the NLP system output to the EPRP review findings. Of the 1083 patients, only 474 patient abstractions had the equivalent data elements to those we captured with CHIEF and were classified as meeting or not meeting the measure. Only 10 patients were present in the EPRP data who did not meet the CHI19 measure. Based on this finding, the SN is the only relevant metric, and with only 10 patients we could not get a precise estimate of any other metrics.

Table 1. Performance of the Congestive Heart Failure Information Extraction Framework (CHIEF) system for each patient compared to the reference standard established by human review (patient-level classification CHI19).

Patient-Level Classification	Sensitivity Estimate % (95% CI)	Positive Predictive Value Estimate % (95% CI)	F-measure
Measure CHI19[a] met	98.9 (97.8, 99.5)	98.7 (97.6, 99.4)	98.8
Left Ventricular Systolic Function assessed	100.0 (99.5, 100.0)	99.0 (98.0, 99.6)	99.5
EF[b] <40%	96.8 (94.6, 98.3)	95.1 (92.6, 97.0)	96.0
ACEI[c] or ARB[d]	99.2 (98.1, 99.7)	88.5 (85.8, 90.8)	93.5
Reason not on medications	26.9 (20.0, 34.9)	90.7 (77.9, 97.4)	41.5

[a] CHI19: Congestive Heart Failure Inpatient Measure 19; LVEF >40 on ACEI/ARB at discharge.

[b] EF: ejection fraction.

[c] ACEI: angiotensin-converting enzyme inhibitor.

[d] ARB: angiotensin-receptor blocker.

Table 2. Sensitivity of patient-level classification of the reference standard and Congestive Heart Failure Information Extraction Framework (CHIEF) based on External Peer Review Program (EPRP) review.

Patient-Level Classification/Sensitivity	Number of Patients in Agreement with EPRP Review	Number of Patients with Corresponding EPRP Review	Sensitivity Estimate % (95% CI)
Classification in Reference Standard	469	474	98.95 (97.56, 99.66)
Classification from CHIEF	325	330	98.48 (96.50, 99.51)

Table 3. The External Peer Review Program (EPRP) quality measurement designation of patients in the training and test sets.

Measure	Data Present	Measure Met	Total n (%)	Number in Test n (%)	Number in Training n (%)
CHI10[a]	No	N/A	74 (6.83)	61 (7.93)	13 (4.14)
	Yes	No	4 (0.36)	4 (0.13)	0 (0.00)
		Yes	1005 (92.79)	704 (91.54)	301 (95.85)
CHI19[b]	No	N/A	600 (55.40)	433 (56.30)	167 (53.18)
	Yes	No	9 (0.83)	6 (0.78)	3 (0.31)
		Yes	474 (43.76)	330 (42.91)	144 (45.85)
CHI20[c]	No	N/A	768 (70.91)	546 (71.09)	222 (28.90)
	Yes	No	9 (0.83)	6 (0.78)	3 (0.96)
		Yes	306 (28.25)	217 (28.22)	89 (28.34)
No data on any measure			54 (4.99)	42 (5.46)	12 (3.82)
Total sample size			1083 (100.00)	769 (100.00)	314 (100.00)

[a] CHI10: Congestive Heart Failure Inpatient Measure 10; inpatient left ventricle function assessed at discharge.

[b] CHI19: Congestive Heart Failure Inpatient Measure 19; LVEF >40 on ACEI/ARB at discharge.

[c] CHI20: Congestive Heart Failure Inpatient Measure 20; LVEF >40 on ACEI/ARB prior to inpatient admission.

We compared the EPRP data with both our RS developed with SMEs, and with the results of CHIEF. When we compared the RS to the EPRP patient classifications using the EPRP findings as truth for patients meeting CHI19 in applicable cases in both the training and test sets (n=474), we found the SN of the RS to be 99.0%. We also compared the NLP results for hospitalizations in the NLP test set (n=330) to the hospitalizations for whom the EPRP provided results for CHI19 using the EPRP findings as truth for patients meeting CHI19 in applicable cases, and found an SN of 98.5% for the CHIEF. Human annotators classified 100% of cases as meeting or not meeting the measure. However, we found that there were no EPRP results for CHI19 for 55.4% of the patients assessed, even though other measures (such as CHI10 or CHI20) might have been completed, making these EPRP results noncomparable to our results. The CHIEF processed and classified 100% of patients in the test set, with 92.1% meeting CHI19. Meeting the measure required that the case was eligible for the performance measure and that the patient data showed that the case satisfied the performance required by CHI19 (see Table 3).

Concordance Between the Reference Standard, Natural Language Processing Output, and Structured Data

We found that the agreement (based on Cohen's kappa) between the PBM data and the RS for RNM was 0.326, and the agreement between the PBM and the NLP system output was 0.221. Both results were interpreted as *fair* agreement [54]. We determined that the low kappa result was due to the PBM data not capturing the reasons why ACEI and ARB were not prescribed, as well as the text documents. When we performed the same calculations for laboratory and ICD-9-CM data for RNM, the laboratory data compared with the RS and NLP output provided kappa values of 0.2083 and 0.1373, respectively. The ICD-9-CM codes indicated only five patients with RNM and showed no agreement with the RS or NLP system output. Similar to the PBM data, clinical text documents are a better data source to capture reasons not to prescribe than laboratory results and ICD-9-CM data. A kappa statistic was calculated as an aggregate measure using the laboratory results and the ICD-9-CM codes as well, but did not differ from the kappa statistic for the laboratory results alone.

Summary Findings from Interviews

Key informants and SMEs provided valuable insights about the design of the CHIEF system and the related development and validation methods. The development team held regular meetings with key informants one to two times per year to review design decisions, such as the capture of concepts to approximate the data elements of the measure. For example, the quality measure assesses whether the patient had left ventricular systolic function assessed. The design team used the presence of an EF in the record of the patient to mean that left ventricular systolic function was assessed. Similarly, there are multiple mentions of the EF in a given echocardiogram report. The design team worked with SMEs to determine the most clinically relevant mention to use in the classification algorithm, and targeted the mention in the section of the report that is a narrative summary by the reviewing cardiologist to extract. Last, the key informants agreed that the research team could use a limited document set, rather that the entire medical record for a given patient discharge, to extract and classify the patient's documentation as to whether or not the measure was met.

Interview respondents also discussed several areas related to how the automated NLP processes are potentially aligned with organizational goals and clinical needs. Respondents noted three potential benefits: (1) use of an automated quality measurement system could improve the efficiency of data capture and thus provide it more quickly; (2) an automated system that facilitates redeployment of resources to emerging areas is aligned with VA organizational goals and strategies; and (3) an NLP system and the resulting data could be used for clinical purposes, in addition to use in quality measurement.

An automated system has the potential to provide consistent data sources for measurement and new data regarding EF to the VA primary care almanac; it could also serve as a data source for primary care teams, VA dashboards, and clinical decision support (CDS). The system could also provide data organized in a summarized, longitudinal manner, and assist cohort and registry development.

The use of an automated quality measurement process for measuring HF quality appears to be aligned with VA organizational goals, could support the current VA culture of measurement and feedback, and provide needed data for accountability. An automated system could also facilitate meaningful use certification, further electronic quality measurement, and assist real-time (rather than retrospective) measurement.

Key informants and SMEs also suggested specific clinical uses for the NLP system and the resulting data, as follows: HF guideline and quality measurement training for providers, automated review and documentation of LVF, identification of patients needing transitional management and palliative care, summarization of clinical findings and treatment to assist clinician decision-making, and identification and contacting of patients with gaps in evidence-based care to aid quality improvement efforts (care coordination).

The interviews provided important information about the automated NLP system and its potential clinical uses. Further research is needed to identify potential technical and organizational barriers to the use of such an NLP system in the VA, as this would help determine the next steps in potential implementation.

Discussion

Principal Results

In this paper we report the formative evaluation of the use of the CHIEF system that integrates core algorithms reported previously [53], in addition to rules derived from existing HF guidelines, to generate a final CHI19 classification. The CHIEF processed 100% of the patients in the test set, with 92.1% of patients classified as meeting the CHI19 measure. Use of the CHIEF could potentially reduce or eliminate the need for routine human review of HF charts for the similar measures of CHF7 (HF-Outpatient LVF documented) and CHF14 (HF-Outpatient LVEF Less Than 40 on ACEI or ARB). CHF is a prevalent condition and CHIEF is an application that could provide an automated first review for HF patients to assess guideline-concordant care, and this data could potentially populate the existing EPRP dashboard automatically rather than through human review. During this process those patients who do not meet the measure could be identified; this would potentially allow a redeployment of human resources to evaluate why the care was not guideline-concordant and evaluate other quality of care issues. For example, more human resources could be used to assess patients who are at high risk for readmission, or who are frail and need additional care coordination.

The CHIEF also provides essential data that could be used in a dashboard to facilitate the identification of patients in a given provider's panel who may need additional medications such as an ACEI or ARB therapy or other medications, as guidelines are updated. Although the EPRP abstractors have access to the entire medical record for each patient they review, they focus only on finding the required data elements within the measure, rather than on a broad quality review in which other quality of care issues may be found. We obtained good results with the CHIEF using a limited document set. These findings suggest that the CHIEF is highly reliable and that its use could reduce or eliminate the expense associated with human review of HF patient records.

Limitations

There are several limitations to this work. First, it is likely that some clinical information was not documented in the patient charts and therefore could not be captured by the NLP system. However, we believe the impact of this missing information is minimal, given the importance and longstanding use of the HF quality measurement. Second, although the CHIEF performed well using VA text notes, it might not perform as well in non-VA settings. After training on new documents, we expect that it will perform similarly. Third, documents from only eight medical centers were used in this research; therefore, the CHIEF might under-perform initially when used with documents from other VA medical centers.

Comparison with Prior Work

This work builds on prior research in which we developed a system for concept extraction using a rule-based method. In the current CHIEF system, we used machine learning-based methods (sequential tagging) [40]. Our informatics work is also complementary to other uses of the NLP system in cases of patients identified as having HF or classified as having a preserved or reduced EF [55,56], for the purposes of identifying patients for potential inclusion in research and those appropriate for treatment in primary care notes [57]. The relevance and importance of NLP tools in clinical practice are increasing. As such, testing and evaluating the implementation and deployment of NLP tools in clinical practice settings is an important next step.

Use of the CHIEF is also aligned with the current VA strategic plan for 2013-2018 that sets forth the principal that all initiatives be data-driven and evidence-based to help VA improve service delivery. The CHIEF has delivered promising results that could help achieve the goals of improving performance, advancing innovation, and increasing operational effectiveness and accountability in the VA, as well as in other health care organizations [58]. While CHIEF is not currently being implemented in the VA, we will seek potential implementation in VA and other settings.

Our work is important because some clinical information related to quality measures can only be found in text. Text data is not structured, so transformation of clinical text documents in a systematic, standardized process could result in its incorporation in a data warehouse across an enterprise, which would allow the use of the National Quality Forum information model designed for EHR-based quality measures, and facilitate the use of algorithms across institutions [59].

Due to the increasing availability of EHRs and the development of NLP techniques, many systems and techniques have been, and continue to be, developed to encode narrative data for a variety of uses such as: assessing the incidence rates of adverse events, evaluating the success of preventive interventions, benchmark performance across hospitals, determining cardiovascular risk factors, providing smoking cessation, providing real-time quality metrics for colonoscopies (in terms of identification of adenomas and sessile serrated adenomas), developing retrospective clinical data for use in cardiovascular research using NLP, and identifying ventilator-associated events (VAEs) and quality reporting and research in VAEs [60-62].

The Institute of Medicine envisioned a health care delivery system that would improve the quality of care and reduce costs. To accomplish this goal, it is important to create effective CDS delivered to clinicians through EHRs at the point of care [63]. The data captured from text, once transformed to structured data in the enterprise CDW, could be used in CDS.

Our methods complement other systems that identify hospitalized patients with HF in which machine learning approaches are used. Importantly, the complexity of implementation of these systems is well known and supports the assessment of barriers and facilitators for potential implementation [62,64]. The use of EHRs to automate publicly reported quality measures is receiving increasing attention, and is one of the promises of EHR implementation. Kaiser Permanente has fully or partly automated 6 of 13 the joint commission measure sets, resulting in an automated surgical site infection reporting process which reduced Kaiser Permanente's manual effort by 80%, resulting is savings of US $2 million [65]. The VA could potentially realize reduced expenses associated with increased automation and decreased manual review of medical records for HF quality measurement.

The use of NLP for quality measures also adds to the capture of large amounts of clinical data from EHRs. The next step is to transform health care *big data* into actionable knowledge for quality improvement and research that helps to improve patient care, and potentially limit health care costs, with the aim of developing infrastructure with real-time data to support decision making [62-64,66,67]. The products of this NLP pipeline could potentially impact a number of clinical areas, including personalized CDS (eg, the suggestion to administer ACEIs/ARBs when inappropriately not administered), and could both facilitate appropriate care by promoting CDS use and prevent provider fatigue by reducing the incidence of false-positive notifications [53]. Our work is also in alignment with the recent description of the use of *big data analytics* in the VA, because the extracted data from our system has been scientifically evaluated for accuracy and reliability, and builds on the significant data resources in the CDW [33].

Conclusions

The CHIEF system accurately classified patients for the CHI19 performance measure, with high SN and PPV. HF is an increasingly prevalent condition among patients within the VA. Our results demonstrate that automated methods using NLP can improve the efficiency and accuracy of data collection and facilitate more complete and timely data capture at the time of discharge, at a potentially reduced cost. These tools also have applications in clinical care delivery and are aligned with US national strategic initiatives to use EHR data for quality improvement.

Acknowledgments

We thank the Office of Performance Measurement and the VA interviewees for assisting us in this research. JHG is a research health scientist at the VA Salt Lake City Health Care System. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs, the United States Government, or the academic affiliate organizations. This work was supported by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, IDEAS 2.0 Center, Health Services Research and Development project #IBE 09-069, and by HIR 08-374 (Consortium for Healthcare Informatics Research). MEM was also supported in part by VA HSR&D CDA 08-020.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Three sets of rules to classify the patient.

[PDF File (Adobe PDF File), 22KB - medinform_v6i1e5_app1.pdf]

References

1. Benjamin E, Blaha M, Chiuve S, Cushman M, Das S, Deo R, American Heart Association Statistics CommitteeStroke Statistics Subcommittee. Heart disease and stroke statistics-2017 update: a report from the American Heart Association. Circulation 2017 Mar 07;135(10):e146-e603 [FREE Full text] [doi: 10.1161/CIR.0000000000000485] [Medline: 28122885]

2. VA Office of Research and Development, Health Services Research and Development Service. Quality Enhancement Research Initiative (QUERI) Chronic Heart Failure Fact Sheet. Palo Alto, California; 2014. URL: http://www.queri.research.va.gov/about/factsheets/chf_factsheet.pdf [accessed 2017-09-27] [WebCite Cache ID 6tn9Tcdo8]
3. Heidenreich P, Albert N, Allen L, Bluemke D, Butler J, Fonarow G, American Heart Association Advocacy Coordinating Committee, Council on Arteriosclerosis, Thrombosis and Vascular Biology, Council on Cardiovascular Radiology and Intervention, Council on Clinical Cardiology, Council on Epidemiology and Prevention, Stroke Council. Forecasting the impact of heart failure in the United States: a policy statement from the American Heart Association. Circ Heart Fail 2013 May;6(3):606-619 [FREE Full text] [doi: 10.1161/HHF.0b013e318291329a] [Medline: 23616602]
4. Centers for Disease Control and Prevention. Heart Failure Fact Sheet. 2017. URL: http://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_heart_failure.htm [accessed 2017-09-27] [WebCite Cache ID 6tn9Eolnz]
5. American Heart Association. 2017. Heart disease and stroke statistics 2017 at-a-glance URL: https://www.heart.org/idc/groups/ahamah-public/@wcm/@sop/@smd/documents/downloadable/ucm_491265.pdf [accessed 2017-09-27] [WebCite Cache ID 6tn9bbJXe]
6. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Colvin MM, et al. 2017 ACC/AHA/HFSA focused update of the 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association Task Force on clinical practice guidelines and the Heart Failure Society of America. J Am Coll Cardiol 2017 Aug 08;70(6):776-803. [doi: 10.1016/j.jacc.2017.04.025] [Medline: 28461007]
7. Blecker S, Paul M, Taksler G, Ogedegbe G, Katz S. Heart failure-associated hospitalizations in the United States. J Am Coll Cardiol 2013 Mar 26;61(12):1259-1267 [FREE Full text] [doi: 10.1016/j.jacc.2012.12.038] [Medline: 23500328]
8. Ponikowski P, Anker SD, AlHabib KF, Cowie MR, Force TL, Hu S, et al. Heart failure: preventing disease and death worldwide. ESC Heart Fail 2014 Sep;1(1):4-25. [doi: 10.1002/ehf2.12005] [Medline: 28834669]
9. Agency for Healthcare Research and Quality. HCUP Fast Stats - Most Common Diagnoses for Inpatient Stays. Rockville, MD; 2017. URL: https://www.hcup-us.ahrq.gov/faststats/NationalDiagnosesServlet [accessed 2017-09-05] [WebCite Cache ID 6vZ4Bxe2L]
10. Polanczyk C, Newton C, Dec G, Di Salvo T. Quality of care and hospital readmission in congestive heart failure: an explicit review process. J Card Fail Dec 2001;7(4):a.
11. Pierre-Louis B, Rodriques S, Gorospe V, Guddati AK, Aronow WS, Ahn C, et al. Clinical factors associated with early readmission among acutely decompensated heart failure patients. Arch Med Sci 2016 Jun 01;12(3):538-545 [FREE Full text] [doi: 10.5114/aoms.2016.59927] [Medline: 27279845]
12. Komajda M, Lapuerta P, Hermans N, Gonzalez-Juanatey JR, van Veldhuisen DJ, Erdmann E, et al. Adherence to guidelines is a predictor of outcome in chronic heart failure: the MAHLER survey. Eur Heart J 2005 Aug;26(16):1653-1659. [doi: 10.1093/eurheartj/ehi251] [Medline: 15827061]
13. Luthi JC, Lund MJ, Sampietro-Colom L, Kleinbaum DG, Ballard DJ, McClellan WM. Readmissions and the quality of care in patients hospitalized with heart failure. Int J Qual Health Care 2003 Oct;15(5):413-421. [Medline: 14527985]
14. Basoor A, Doshi NC, Cotant JF, Saleh T, Todorov M, Choksi N, et al. Decreased readmissions and improved quality of care with the use of an inexpensive checklist in heart failure. Congest Heart Fail 2013;19(4):200-206 [FREE Full text] [doi: 10.1111/chf.12031] [Medline: 23910702]
15. Jha AK, Perlin JB, Kizer KW, Dudley RA. Effect of the transformation of the Veterans Affairs Health Care System on the quality of care. N Engl J Med 2003 May 29;348(22):2218-2227. [doi: 10.1056/NEJMsa021899] [Medline: 12773650]
16. Kupersmith J, Francis J, Kerr E, Krein S, Pogach L, Kolodner R, et al. Advancing evidence-based care for diabetes: lessons from the Veterans Health Administration. Health Aff (Millwood) 2007 Apr;26(2):156-168.
17. Blecker S, Agarwal SK, Chang PP, Rosamond WD, Casey DE, Kucharska-Newton A, et al. Quality of care for heart failure patients hospitalized for any cause. J Am Coll Cardiol 2014 Jan 21;63(2):123-130 [FREE Full text] [doi: 10.1016/j.jacc.2013.08.1628] [Medline: 24076281]
18. Centers for Medicare and Medicaid Services. 2017. Hospital Compare URL: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/hospitalqualityinits/hospitalcompare.html [accessed 2017-09-27] [WebCite Cache ID 6tn9pwtx1]
19. Payne VL, Hysong SJ. Model depicting aspects of audit and feedback that impact physicians' acceptance of clinical performance feedback. BMC Health Serv Res 2016 Jul 13;16:260 [FREE Full text] [doi: 10.1186/s12913-016-1486-3] [Medline: 27412170]
20. HealthIT.gov. 2017. EHR incentives & certification, meaningful use definition & objectives URL: http://www.healthit.gov/providers-professionals/meaningful-use-definition-objectives [accessed 2017-09-27] [WebCite Cache ID 6tn9xCdsO]
21. Goulet J, Erdos J, Kancir S, Levin F, Wright S, Daniels S, et al. Measuring performance directly using the veterans health administration electronic medical record: a comparison with external peer review. Med Care 2007 Jan;45(1):73-79 [FREE Full text] [doi: 10.1097/01.mlr.0000244510.09001.e5] [Medline: 17279023]
22. Gray K, Sockolow P. Conceptual models in health informatics research: a literature review and suggestions for development. JMIR Med Inform 2016;4(1):A. [Medline: 26912288]
23. Carbonell JG, Hayes PJ. Natural language understanding. In: Shapiro SC, editor. Encyclopedia of Artificial Intelligence. Indianapolis: Wiley; Jan 2007:73-79.

24. McGinn T. Putting meaning into meaningful use: a roadmap to successful integration of evidence at the point of care. JMIR Med Inform 2016 May 19;4(2):e16 [FREE Full text] [doi: 10.2196/medinform.4553] [Medline: 27199223]
25. Goldstein MK. Using health information technology to improve hypertension management. Curr Hypertens Rep 2008 Jun;10(3):201-207. [Medline: 18765090]
26. Kawamoto K, Houlihan C, Balas E, Lobach D. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005 Apr 02;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
27. Bright T, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux R, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med 2012 Jul 03;157(1):29-43. [doi: 10.7326/0003-4819-157-1-201207030-00450] [Medline: 22751758]
28. Syrowatka A, Krömker D, Meguerditchian A, Tamblyn R. Features of computer-based decision aids: systematic review, thematic synthesis, and meta-analyses. J Med Internet Res 2016 Jan 26;18(1):e20 [FREE Full text] [doi: 10.2196/jmir.4982] [Medline: 26813512]
29. Sirajuddin A, Osheroff J, Sittig D, Chuo J, Velasco F, Collins D. Implementation pearls from a new guidebook on improving medication use and outcomes with clinical decision support. Effective CDS is essential for addressing healthcare performance improvement imperatives. J Healthc Inf Manag 2009;23(4):38-45 [FREE Full text] [Medline: 19894486]
30. Khorasani R, Hentel K, Darer J, Langlotz C, Ip I, Manaker S, et al. Ten commandments for effective clinical decision support for imaging: enabling evidence-based practice to improve quality and reduce waste. AJR Am J Roentgenol 2014 Nov;203(5):945-951. [doi: 10.2214/AJR.14.13134] [Medline: 25341131]
31. Pharmacy Benefits Management Strategic Healthcare Group, Medical Advisory Panel. PBM-MAP clinical practice guideline for the pharmacologic management of chronic heart failure in primary care practice. Publication no. 00-0015. Washington, DC: Veterans Health Administration, Department of Veteran Affairs; 2007. URL: https://www.healthquality.va.gov/guidelines/cd/chf/chf_full_text.pdf [accessed 2018-01-08] [WebCite Cache ID 6wKbc656r]
32. VA Informatics and Computing Infrastructure (VINCI). 2017. URL: https://www.hsrd.research.va.gov/for_researchers/vinci/ [accessed 2017-09-27] [WebCite Cache ID 6tnA446rDVA]
33. VA Corporate Data Warehouse (CDW). 2017. URL: https://www.hsrd.research.va.gov/for_researchers/vinci/cdw.cfm [accessed 2017-09-27] [WebCite Cache ID 6vZ4xsAyF]
34. Bhatia R, Tu J, Lee D, Austin P, Fang J, Haouzi A, et al. Outcome of heart failure with preserved ejection fraction in a population-based study. N Engl J Med 2006 Jul 20;355(3):260-269. [doi: 10.1056/NEJMoa051530] [Medline: 16855266]
35. Knowtator. 2009 Jul 17. URL: http://knowtator.sourceforge.net/ [accessed 2018-01-08] [WebCite Cache ID 6wJ1KfdS1]
36. Ogren PV. A Protégé plug-in for annotated corpus construction. In: Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Stroudsburg: Association for Computational Linguistics; 2006 Presented at: 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology; June 4-9, 2006; New York p. 273-275 URL: https://dl.acm.org/citation.cfm?id=1225785&picked=prox&preflayout=tabs [doi: 10.3115/1225785.1225791]
37. Ashton C, Kuykendall D, Johnson M, Wray NP. An empirical assessment of the validity of explicit and implicit process-of-care criteria for quality assessment. Med Care 1999 Aug;37(8):798-808. [Medline: 10448722]
38. Apache UIMA. 2017. URL: http://uima.apache.org [accessed 2017-09-27] [WebCite Cache ID 6tnAFGnBQ]
39. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat Lang Eng 1999;10(3-4):327-348. [doi: 10.1017/S1351324904003523]
40. Garvin J, DuVall SL, South B, Bray B, Bolton D, Heavirland J, et al. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. J Am Med Inform Assoc 2012;19(5):859-866 [FREE Full text] [doi: 10.1136/amiajnl-2011-000535] [Medline: 22437073]
41. South B, Shen S, Leng J, Forbush T, DuVall S, Chapman W. A prototype tool set to support machine-assisted annotation. In: BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. Montreal: Association for Computational Linguistics; 2012 Jun 01 Presented at: Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP) 139; 2012; Montreal, QC p. 130-139.
42. Meystre S, Kim Y, Garvin J. Comparing methods for left ventricular ejection fraction clinical information extraction. In: Proceedings of the AMIA Summit on Clinical Research Informatics. Bethesda: American Medical Informatics Association; 2012 Mar 21 Presented at: Proceedings of the AMIA Summit on Clinical Research Informatics; 2012; San Francisco URL: https://knowledge.amia.org/amia-55142-cri2012a-1.644955?qr=1
43. Kim Y, Garvin J, Heavirland J, Meystre S. Improving heart failure information extraction by domain adaptation. Stud Health Technol Inform 2013;192:185-189. [Medline: 23920541]
44. Kim Y, Garvin J, Heavirland J, Meystre S. Medication prescription status classification in clinical narrative documents. In: AMIA Annual Symposium Proceedings. Bethesda: American Medical Informatics Association; 2014 Nov 15 Presented at: AMIA Annual Symposium 2014; 2014; Washington, DC URL: https://knowledge.amia.org/56638-amia-1.1540970?qr=1
45. Gobbel G, Reeves R, Jayaramaraja S, Giuse D, Speroff T, Brown S, et al. Development and evaluation of RapTAT: a machine learning system for concept mapping of phrases from medical narratives. J Biomed Inform 2014 Apr;48:54-65 [FREE Full text] [doi: 10.1016/j.jbi.2013.11.008] [Medline: 24316051]

46. Rycroft-Malone J. The PARIHS framework--a framework for guiding the implementation of evidence-based practice. J Nurs Care Qual 2004;19(4):297-304. [Medline: 15535533]
47. Kitson A, Rycroft-Malone J, Harvey G, McCormack B, Seers K, Titchen A. Evaluating the successful implementation of evidence into practice using the PARiHS framework: theoretical and practical challenges. Implement Sci 2008 Jan 07;3:1 [FREE Full text] [doi: 10.1186/1748-5908-3-1] [Medline: 18179688]
48. Sittig D, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care 2010 Oct;19 Suppl 3:i68-i74 [FREE Full text] [doi: 10.1136/qshc.2010.042085] [Medline: 20959322]
49. Van Rijsbergen CJ. Information Retrieval, 2nd edition. London: Butterworths; Nov 1979:374-375.
50. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20(1):37-46. [doi: 10.1177/001316446002000104]
51. Garvin J, Heavirland J, Weaver A. Determining section types to capture key clinical data for automation of quality measurement for inpatients with heart failure. In: Poster Session Proceeding American Medical Informatics Association Annual Symposium. Bethesda: American Medical Informatics Association; 2012 Presented at: The American Medical Informatics Association Annual Symposium; November 3, 2012; Washington, DC URL: https://knowledge.amia.org/amia-55142-a2012a-1.636547?qr=1
52. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. J Biomed Inform 2002 Apr;35(2):99-110. [Medline: 12474424]
53. Meystre SM, Kim Y, Gobbel GT, Matheny ME, Redd A, Bray BE, et al. Congestive heart failure information extraction framework for automated treatment performance measures assessment. J Am Med Inform Assoc 2017 Apr 01;24(e1):e40-e46. [doi: 10.1093/jamia/ocw097] [Medline: 27413122]
54. Landis J, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
55. Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, et al. An information extraction framework for cohort identification using electronic health records. In: AMIA Joint Summits Translational Science Proceedings. Bethesda: American Medical Informatics Association; 2013 Presented at: American Medical Informatics Association Joint Summits Translational Science; March 18th; San Francisco, CA p. 149-153 URL: https://knowledge.amia.org/amia-55142-tbi2013a-1.649951?qr=1
56. Pakhomov S, Weston S, Jacobsen S, Chute C, Meverden R, Roger V. Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care 2007 Jun;13(6 Part 1):281-288 [FREE Full text] [Medline: 17567225]
57. Byrd R, Steinhubl S, Sun J, Ebadollahi S, Stewart WF. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. Int J Med Inform 2014 Dec;83(12):983-992 [FREE Full text] [doi: 10.1016/j.ijmedinf.2012.12.005] [Medline: 23317809]
58. Department of Veterans Affairs, Veterans Health Administration. 2013-2018 strategic plan URL: https://www.va.gov/health/docs/VHA_STRATEGIC_PLAN_FY2013-2018.pdf [accessed 2017-09-27] [WebCite Cache ID 6tnANCEkO]
59. Thompson WK, Rasmussen LV, Pacheco JA, Peissig PL, Denny JC, Kho AN, et al. An evaluation of the NQF Quality Data Model for representing Electronic Health Record driven phenotyping algorithms. In: AMIA Annual Symposium Proceedings. https://knowledge.amia.org/amia-55142-a2012a-1.636547?qr=1: American Medical Informatics Association; 2012 Presented at: AMIA Annual Symposium Proceedings 2012; Nov 3, 2012; Chicago, IL p. 911-920 URL: http://europepmc.org/abstract/MED/23304366
60. Rochefort CM, Buckeridge DL, Tanguay A, Biron A, D'Aragon F, Wang S, et al. Accuracy and generalizability of using automated methods for identifying adverse events from electronic health record data: a validation study protocol. BMC Health Serv Res 2017 Feb 16;17(1):147 [FREE Full text] [doi: 10.1186/s12913-017-2069-7] [Medline: 28209197]
61. Khalifa A, Meystre S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. J Biomed Inform 2015 Dec;58 Suppl:S128-S132 [FREE Full text] [doi: 10.1016/j.jbi.2015.08.002] [Medline: 26318122]
62. Lan H, Thongprayoon C, Ahmed A, Herasevich V, Sampathkumar P, Gajic O, et al. Automating quality metrics in the era of electronic medical records: digital signatures for ventilator bundle compliance. Biomed Res Int 2015;2015:396508 [FREE Full text] [doi: 10.1155/2015/396508] [Medline: 26167484]
63. Jones JB, Stewart WF, Darer JD, Sittig DF. Beyond the threshold: real-time use of evidence in practice. BMC Med Inform Decis Mak 2013 Apr 15;13:47 [FREE Full text] [doi: 10.1186/1472-6947-13-47] [Medline: 23587225]
64. Blecker S, Katz SD, Horwitz LI, Kuperman G, Park H, Gold A, et al. Comparison of approaches for heart failure case identification from electronic health record data. JAMA Cardiol 2016 Dec 01;1(9):1014-1020 [FREE Full text] [doi: 10.1001/jamacardio.2016.3236] [Medline: 27706470]
65. Raju GS, Lum PJ, Slack RS, Thirumurthi S, Lynch PM, Miller E, et al. Natural language processing as an alternative to manual reporting of colonoscopy quality metrics. Gastrointest Endosc 2015 Sep;82(3):512-519 [FREE Full text] [doi: 10.1016/j.gie.2015.01.049] [Medline: 25910665]

66. Patterson OV, Freiberg MS, Skanderson M, Brandt CA, DuVall SL. Unlocking echocardiogram measurements for heart disease research through natural language processing. BMC Cardiovasc Disord 2017 Jun 12;17(1):151 [FREE Full text] [doi: 10.1186/s12872-017-0580-8] [Medline: 28606104]

67. Ross MK, Wei W, Ohno-Machado L. “Big data” and the electronic health record. Yearb Med Inform 2014 Aug 15;9:97-104 [FREE Full text] [doi: 10.15265/IY-2014-0003] [Medline: 25123728]

Abbreviations

ACEI: angiotensin-converting enzyme inhibitor
ARB: angiotensin-receptor blocker
CHF7: HF-Congestive Heart Failure Outpatient Measure 7
CHF14: Congestive Heart Failure Outpatient Measure 14
CHI10: Congestive Heart Failure Inpatient Measure 10
CHI19: Congestive Heart Failure Inpatient Measure 19
CHI20: Congestive Heart Failure Inpatient Measure 20
CDS: clinical decision support
CHIEF: Congestive Heart Failure Information Extraction Framework
CDW: Corporate Data Warehouse
EF: ejection fraction
EHR: electronic health record
EPRP: External Peer Review Program
HF: heart failure
IAA: interannotator agreement
ICD-9-CM: International Classification of Diseases, 9th Revision, Clinical Modification
LVEF: left ventricular ejection fraction
LVF: left ventricular failure
NLP: natural language processing
PBM: Pharmacy Benefit Management
PPV: positive predictive value
PARiHS: Promoting Action on Research Implementation in Health Sciences
RNM: reasons no medications
RS: Reference Standard
SN: sensitivity
SP: specificity
SME: subject matter expert
TIU: text integration utilities
UIMA: Unstructured Information Management Architecture
VA: United States Department of Veterans Affairs
VAE: ventilator-associated event

Edited by G Eysenbach; submitted 09.10.17; peer-reviewed by M Brodnick, A Meguerditchian, A Syrowatka, I Kedan, S Doan, J Zheng; comments to author 03.11.17; revised version received 08.12.17; accepted 10.12.17; published 15.01.18.

Please cite as:
Garvin JH, Kim Y, Gobbel GT, Matheny ME, Redd A, Bray BE, Heidenreich P, Bolton D, Heavirland J, Kelly N, Reeves R, Kalsy M, Goldstein MK, Meystre SM
Automating Quality Measures for Heart Failure Using Natural Language Processing: A Descriptive Study in the Department of Veterans Affairs
JMIR Med Inform 2018;6(1):e5
URL: http://medinform.jmir.org/2018/1/e5/
doi:10.2196/medinform.9150
PMID:29335238

Original Paper

Potential Application of Digitally Linked Tuberculosis Diagnostics for Real-Time Surveillance of Drug-Resistant Tuberculosis Transmission: Validation and Analysis of Test Results

Kamela Charmaine Ng[1], MSc; Conor Joseph Meehan[1], PhD; Gabriela Torrea[1], PhD; Léonie Goeminne[2], BSc; Maren Diels[1], BSc; Leen Rigouts[1,3], MSc, PhD; Bouke Catherine de Jong[1], MD, MSc, PhD; Emmanuel André[2,4], MD, MSc, PhD

[1]Mycobacteriology Unit, Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium

[2]Pôle de Microbiologie Médicale, Institut de Recherche Expérimentale et Clinique, Université Catholique de Louvain, Brussels, Belgium

[3]Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium

[4]Laboratory of Clinical Bacteriology and Mycology, Katholieke Universiteit Leuven, Leuven, Belgium

Corresponding Author:
Kamela Charmaine Ng, MSc
Mycobacteriology Unit
Department of Biomedical Sciences
Institute of Tropical Medicine
Nationalestraat 155
Antwerp, 2000
Belgium
Phone: 32 (0) 33455345
Fax: 32 (0) 32476333
Email: kng@itg.be

Abstract

Background: Tuberculosis (TB) is the highest-mortality infectious disease in the world and the main cause of death related to antimicrobial resistance, yet its surveillance is still paper-based. Rifampicin-resistant TB (RR-TB) is an urgent public health crisis. The World Health Organization has, since 2010, endorsed a series of rapid diagnostic tests (RDTs) that enable rapid detection of drug-resistant strains and produce large volumes of data. In parallel, most high-burden countries have adopted connectivity solutions that allow linking of diagnostics, real-time capture, and shared repository of these test results. However, these connected diagnostics and readily available test results are not used to their full capacity, as we have yet to capitalize on fully understanding the relationship between test results and specific *rpoB* mutations to elucidate its potential application to real-time surveillance.

Objective: We aimed to validate and analyze RDT data in detail, and propose the potential use of connected diagnostics and associated test results for real-time evaluation of RR-TB transmission.

Methods: We selected 107 RR-TB strains harboring 34 unique *rpoB* mutations, including 30 within the rifampicin resistance–determining region (RRDR), from the Belgian Coordinated Collections of Microorganisms, Antwerp, Belgium. We subjected these strains to Xpert MTB/RIF, GenoType MTBDR*plus* v2.0, and Genoscholar NTM + MDRTB II, the results of which were validated against the strains' available *rpoB* gene sequences. We determined the reproducibility of the results, analyzed and visualized the probe reactions, and proposed these for potential use in evaluating transmission.

Results: The RDT probe reactions detected most RRDR mutations tested, although we found a few critical discrepancies between observed results and manufacturers' claims. Based on published frequencies of probe reactions and RRDR mutations, we found specific probe reactions with high potential use in transmission studies: Xpert MTB/RIF probes A, Bdelayed, C, and Edelayed; Genotype MTBDR*plus* v2.0 WT2, WT5, and WT6; and Genoscholar NTM + MDRTB II S1 and S3. Inspection of probe reactions of disputed mutations may potentially resolve discordance between genotypic and phenotypic test results.

Conclusions: We propose a novel approach for potential real-time detection of RR-TB transmission through fully using digitally linked TB diagnostics and shared repository of test results. To our knowledge, this is the first pragmatic and scalable work in response to the consensus of world-renowned TB experts in 2016 on the potential of diagnostic connectivity to accelerate efforts to eliminate TB. This is evidenced by the ability of our proposed approach to facilitate comparison of probe reactions between different RDTs used in the same setting. Integrating this proposed approach as a plug-in module to a connectivity platform will increase usefulness of connected TB diagnostics for RR-TB outbreak detection through real-time investigation of suspected RR-TB transmission cases based on epidemiologic linking.

(JMIR Med Inform 2018;6(1):e12) doi:10.2196/medinform.9309

KEYWORDS

tuberculosis; drug resistance; rifampicin-resistant tuberculosis; rapid diagnostic tests; Xpert MTB/RIF; Genotype MTBDRplus v2.0; Genoscholar NTM + MDRTB II; RDT probe reactions; rpoB mutations; validation and analysis; real-time detection

Introduction

Tuberculosis (TB) has the highest mortality of any infectious disease and is the principal cause of death related to antimicrobial resistance [1]. Efforts to control TB are continuously hampered by complex regimens and low treatment success of drug-resistant cases [1,2]. Multidrug-resistant TB (MDR-TB), defined as resistance to the first-line drugs rifampicin and isoniazid, remains an urgent public health crisis, as only about 39% of notified, confirmed, and previously treated people with TB were tested for rifampicin resistance in 2016, and only 1 in 5 received treatment, of whom only half were cured [1].

Rifampicin resistance is an epidemiologically and clinically important surrogate marker for MDR-TB, as most rifampicin-resistant strains are also resistant to isoniazid [2-4]. Rifampicin resistance-conferring mutations are primarily situated at codon positions 426 to 452 [2] within the 81-bp rifampicin resistance–determining region (RRDR) of the *Mycobacterium tuberculosis* RNA polymerase β subunit (*rpoB*) gene [4-6]. Consequently, commercially available molecular tests developed for rapid detection of MDR-TB only capture mutations in the RRDR, with or without mutations associated with isoniazid resistance [2,3,6].

Recognizing the immediate need to rapidly detect rifampicin-resistant TB (RR-TB), the World Health Organization has recommended implementation of the following rapid diagnostic tests (RDTs) as primary tools for detection: Xpert MTB/RIF (Cepheid, Sunnyvale, CA, USA), GenoType MTBDR*plus* v2.0 (Hain Lifescience GmbH, Nehren, Germany), and Genoscholar NTM + MDRTB II (NIPRO Corporation, Osaka, Japan). Xpert MTB/RIF is the most widely deployed RDT globally, implemented as the initial diagnostic tool by 28 out of 48 high-burden countries for patients with pulmonary TB symptoms by the end of 2016 [1]. It uses real-time polymerase chain reaction and molecular beacon technology, involving probes specifically binding to wild-type sequences, whereas the other RDTs are line probe assays, which rely on hybridization and comprise both wild-type and mutant probes. The specific features and limitations of the RDTs, along with key recommendations from the World Health Organization, have been previously described [6]. The global integration and scale-up of these RDTs in TB diagnostics has dramatically improved detection of MDR-TB [1,4,6-8].

Large gaps in detection and treatment of RR-TB can result in resistant strains circulating within populations [7,9]. Particularly in high-burden settings, transmission was found to be the predominant cause of the globally rising rates of RR-TB and MDR-TB [8,10,11], with an estimated 600,000 new cases of RR-TB arising in 2016, of which 490,000 were MDR-TB [1,8]. To achieve the global targets of ending TB by 2030 [1], there is an urgent need to shift from the current inefficient paper-based investigation of drug-resistant cases to fully digitized surveillance allowing for real-time detection of transmission hotspots.

When Xpert MTB/RIF was rolled out in 2010, there was no system in place to systematically extract, interpret, and employ test results for surveillance. Accumulating data stored in local machines were clearly underused; portions of data even get corrupted and are not used at all. For a few years after, the gap in TB diagnostics implementation was addressed by the emergence of diagnostic eHealth solutions, particularly the development of connectivity solutions. Xpert MTB/RIF machines are now being linked, along with generated test results, to connectivity platforms such as DataToCare (Savics, Brussels, Belgium) and GxAlert (SystemOne, Springfield, MA, USA). These connectivity platforms automate the collection of diagnostic test results from each health facility in a particular setting. They consolidate all data into a built-in analytics system that automates extraction of useful information from raw test data. Generated information is then securely shared with different stakeholders—clinicians, national TB control programs, and patients—through text messaging, email, or a national database [12].

Despite this technological advancement, readily available and accumulating TB diagnostic test results have not been used to date for disease surveillance, due to the limited understanding of the correlation between RDT probe reactions and specific *rpoB* mutations. In accordance with the consensus of global TB experts on the potential of diagnostic connectivity for TB elimination published in 2016 [12], we believe that RDT results coupled with appropriate analysis may present crucial clinical and public health information that could be employed as a molecular epidemiologic tool in field conditions to trace RR-TB transmission hotspots in high-burden TB settings. They could also aid in resolving the discordance between phenotypic drug susceptibility testing and RDT results, primarily for disputed mutations, which are often missed in *Mycobacterium* growth indicator tube phenotypic drug susceptibility testing [13]. In this work, we aimed to analyze RDT data in detail, visualize test results representing specific RRDR mutations, and propose a novel approach of fully using digitally linked TB diagnostics and readily available test results for real-time monitoring of RR-TB transmission.

Methods

Source and Overview of Strains

We selected 107 RR-TB strains harboring 34 unique *rpoB* mutations, including 30 within the RRDR (Multimedia Appendix 1), from the Belgian Coordinated Collections of Microorganisms and the World Health Organization Tropical Disease Research mycobacteria collection in the Institute of Tropical Medicine, Antwerp, Belgium. We prepared thermolysates of these strains as previously described, stored them at –20°C [14], and subjected them to RDTs, namely Xpert MTB/RIF, GenoType MTBDR*plus* v2.0, and Genoscholar NTM + MDRTB II.

Xpert MTB/RIF Assay G4 Version 5

We measured the DNA concentration of the thermolysates through an Invitrogen Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) following the manufacturer's instructions. The calculated weight of DNA per *M tuberculosis* bacillus served as the divisor for each DNA concentration, the quotient of which was the equivalent colony-forming units (CFUs) per milliliter of each thermolysate.

We subjected thermolysates with initial concentrations of 10^8 to 10^9CFU/mL to 1:2 and 10-fold serial dilutions until we obtained 10^6 to 10^7CFU/mL, respectively. We then dispensed a 1:2 mixture of diluted thermolysate and sample reagent in the Xpert MTB/RIF cartridge following the manufacturer's instructions.

GenoType MTBDR*plus* v2.0 and Genoscholar NTM + MDRTB II

We subjected undiluted thermolysates to the amplification and hybridization steps of GenoType MTBDR*plus* v2.0 and Genoscholar NTM + MDRTB II following the manufacturers' instructions.

Analysis and Visualization of Rapid Diagnostic Test Data

We validated RDT results against the strains' *rpoB* gene sequences, allowing for correlation between probe reactions and specific *rpoB* mutations. We then performed a comprehensive comparison and analysis of *rpoB* mutants and associated RDT results using the consensus numbering system based on *M tuberculosis* strain H37Rv [2]. We visualized results of the analysis using Geneious version 10.1.2 (Biomatters Limited) [15] and Affinity Designer 1.5.3 (Serif (Europe) Ltd) [16]. The percentage average reproducibility of the RDTs was analyzed through mutation profiles found in more than one strain.

Results

All strains with mutations in the RRDR were correctly identified as rifampicin resistant, except for one strain with the S428R+H445R mutation that was not initially detected as *M tuberculosis* by Xpert MTB/RIF but was confirmed as rifampicin-resistant *M tuberculosis* on retesting; a strain with the D435F mutation and another strain with the L430P mutation were identified as rifampicin susceptible by GenoType MTBDR*plus* v2.0 even after a repeat test. On the probe level, each test appeared to be reproducible, ranging between 87.9% and 97.8% (Multimedia Appendix 2). Among the RDTs, Genoscholar NTM + MDRTB II yielded the highest concordance between coverage claimed by manufacturers and that observed experimentally.

The alignment of mutant and wild-type RRDR sequences in Multimedia Appendix 1 shows the nucleotide and amino acid change for each mutation profile captured by *rpoB* sequencing.

Multimedia Appendix 1 provides supporting information for Figure 1, a visualization of the probe reactions demonstrating the ability of the RDTs to capture the majority of documented RRDR mutations.

Figure 1 was generated based on observed probe reactions for each individual mutation. Hence, strains with double or triple mutations were scored twice or thrice. For instance, from a strain with the double mutation L430P+H445Q, the probe reaction for L430P was deduced as Xpert MTB/RIF probe A, GenoType MTBDR*plus* v2.0 probe WT2, and Genoscholar NTM + MDRTB II probe S1, whereas the probe reaction for H445Q was Xpert MTB/RIF probe D; for GenoType MTBDR*plus* v2.0, was WT7; and for Genoscholar NTM + MDRTB II, was S4. As expected, we found a strong correlation between observed and claimed probe reactions of RRDR mutations (Figure 1, black and green, represented by low-prevalence mutations L430P, D435G, S441L or S441Q, and L452P), although we noted delayed reactions in Xpert MTB/RIF, denoting partial inhibition of fluorescence of a particular molecular beacon [3], shown in green. It is interesting to note that some mutations were missed by one probe but captured by another, such as 435 mutations missed by Xpert MTB/RIF probe C but captured by probe B, and a codon 437 mutation in the line probe assays (Figure 1, blue mutations). End-probe mutations correctly identified by both probes in GenoType MTBDR*plus* v2.0 were Q432E (WT2 and WT3) and S441Q or S441L (WT5 and WT6). Critically, there were individual mutations completely missed by the RDTs, namely M434T and N437D by Xpert MTB/RIF, S428R and D435F by GenoType MTBDR*plus* v2.0, and M4343V by Genoscholar NTM + MDRTB II (Figure 1, red mutations). T444T is a silent mutation appropriately undetected by GenoType MTBDR*plus* v2.0, but it was captured by Xpert MTB/RIF and Genoscholar NTM + MDRTB II.

We gathered the observed probe reactions representing each of the 30 RRDR mutations tested (Figure 2).

For instance, Xpert MTB/RIF probe E, GenoType MTBDR*plus* v2.0 probes WT8 and MUT3, and Genoscholar NTM + MDRTB II probes S5 and R5 correspond with mutation S450L, with the highest worldwide prevalence [17-19].

Figure 1. Overview of rifampicin-resistant tuberculosis rapid diagnostic test (RDT) probe reactions. The observed results for each rifampicin resistance–determining region mutation are overlaid on claimed probe coverage (light gray) of (A) Xpert MTB/RIF, (B) GenoType MTBDR*plus* v2.0, and (C) Genoscholar NTM + MDRTB II. Mutations yielding the expected probe reactions are in black and green, with delayed XpertMTB/RIF results in green, mutations missed by one probe but captured by another probe in blue, and mutations that were not at all captured by the RDT in red. Probe reactions overlaid on each other are in a striped pattern for greater visibility.

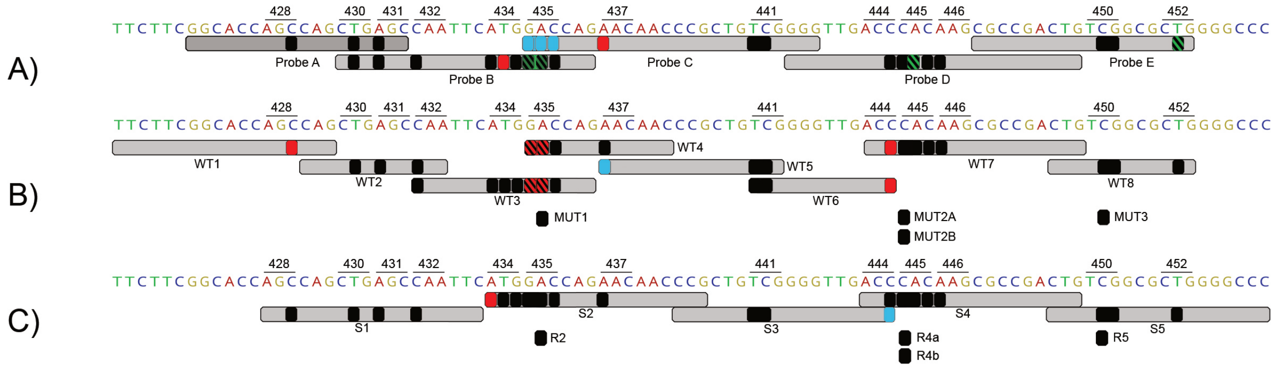
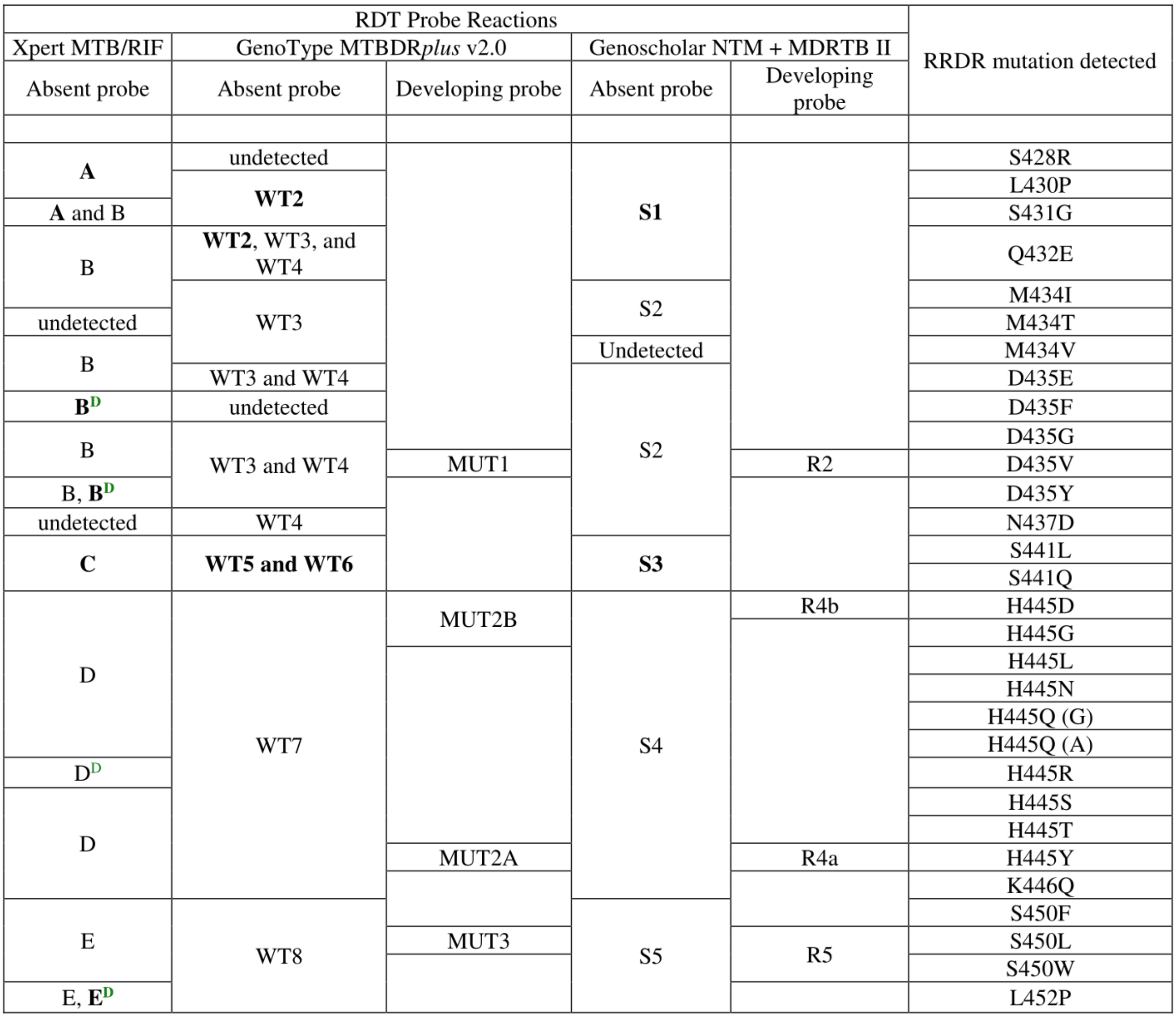


Figure 2. Rapid diagnostic test (RDT) probe reactions corresponding with specific rifampicin resistance–determining region (RRDR) mutations. Probe reactions with high potential use for transmission studies are bolded, while delayed probe reactions are tagged with superscript D shown in green.



RDT Probe Reactions					RRDR mutation detected
Xpert MTB/RIF	GenoType MTBDR*plus* v2.0		Genoscholar NTM + MDRTB II		
Absent probe	Absent probe	Developing probe	Absent probe	Developing probe	
A	undetected		**S1**		S428R
	WT2				L430P
A and B					S431G
B	**WT2**, WT3, and WT4				Q432E
	WT3		S2		M434I
undetected					M434T
B			Undetected		M434V
	WT3 and WT4		S2		D435E
B^{D}	undetected				D435F
B	WT3 and WT4				D435G
		MUT1		R2	D435V
B, **B**^{D}					D435Y
undetected	WT4				N437D
C	**WT5 and WT6**		**S3**		S441L
					S441Q
D	WT7	MUT2B	S4	R4b	H445D
					H445G
					H445L
					H445N
					H445Q (G)
					H445Q (A)
D^{D}					H445R
D					H445S
					H445T
		MUT2A		R4a	H445Y
					K446Q
E	WT8		S5		S450F
		MUT3		R5	S450L
					S450W
E, **E**^{D}					L452P

Finally, we observed a possible association of delayed Xpert MTB/RIF probe reactions with weaker fluorescence in end-probe codons and specific nucleotide substitution types. This was exemplified by guanine-to-thymine transversion in disputed mutation D435Y resulting in 53.8% of delays (averageΔCt8) for probe B and thymine-to-cytosine transition in end-probe disputed mutation L452P resulting in 25% delayed results (averageΔCt6) for probe E (Multimedia Appendices 3-8).

Discussion

Principal Findings and Comparison With Prior Work

Our laboratory validation of rifampicin-resistant strains and visualization of results revealed the specific relationship between unique RDT probe reactions and the majority of documented *rpoB* mutations. We also found RDT probe reactions representing low-frequency mutations in particular settings that have a high potential for use in evaluating transmission. We are proposing a novel approach to the optimal use of readily available diagnostic data through the shift to fully automated and digitized surveillance for real-time detection of RR-TB transmission hotspots.

To our knowledge, this work is the first pragmatic and scalable response to the consensus of world-renowned TB experts, published in 2016 [12], that diagnostic connectivity has the potential to eliminate TB. This is evidenced by the ability of our proposed approach to facilitate comparison of probe reactions between different RDTs used in the same setting, potentially linked to an open source software platform, by the Connected Diagnostics Initiative [12], for instance. Our findings affirm the important role of diagnostic connectivity platforms in accelerating TB control efforts and highlight the importance of fully understanding the relationship between TB diagnostic test results and drug resistance-conferring mutations. Our proposed approach will maximize the use of connected RDTs and associated data shared in repositories through a plug-in module that will automatically translate RDT results to useful information for RR-TB outbreak investigations. This will potentially aid laboratory personnel, clinicians, and national TB control programs in closing detection gaps of RR-TB cases by revealing abnormal figures or trends in particular settings, as well as real-time probable horizontal transmission and epidemiologic linkage between patients.

The comprehensive analysis of claimed and observed probe reactions revealed the ability of the RDTs to detect documented RRDR mutations. Most of the RRDR mutations analyzed yielded the expected result, concordant with manufacturers' claims. We observed the highest concordance between observed and claimed probe coverage in Genoscholar NTM + MDRTB II, exemplified, for instance, by probe S2 detecting mutation N437D, which was missed by Xpert MTB/RIF probe C and GenoType MTBDR*plus* v2.0 probe WT5 (Figure 2,Multimedia Appendix 1, and Figure 1). Remarkably, the mutations missed by one RDT probe but captured by another were situated in the regions where the probes overlap. This affirms the role of overlapping probes in improving the usefulness of RDTs for RR-TB detection. Critically, mutations M434T missed by Xpert MTB/RIF and M434V missed by Genoscholar NTM + MDRTB II are not yet catalogued in the Tuberculosis Drug Resistance Mutation Database, nor in RefSeq [20,21]; hence, their clinical relevance is yet to be determined. Alarmingly, missed mutations, including N437D (Xpert MTB/RIF), S428R, and D435F (GenoType MTBDR*plus* v2.0), are known to confer rifampicin resistance [22]. These undetected rifampicin resistance-conferring mutations within the RRDR, together with those outside the RRDR, most notably the I491F mutation [2], may be untraceably circulated through chains of transmission, as samples with these mutations alone will be falsely classified as rifampicin susceptible, rendering treatment ineffective. Additionally, the impact of silent mutations such as T444T must be further assessed to ensure that recommended probes generate results with high specificity for rifampicin resistance. Accordingly, we suggest that manufacturers review and modify claims of mutation coverage based on the findings of this work. Additionally, for future versions of the line probe assays, manufacturers might consider synthesizing mutation probes that would also capture low-frequency mutations. These proposed modifications would improve identification of underlying mutations and thus increase the usefulness of linked RDTs for detection of rifampicin resistance and for epidemiologic surveillance.

We also gathered probe reactions representing low- to high-prevalence RRDR mutations [4,12,13,18,19,22] in the most comprehensive comparison of *rpoB* mutants and associated RDT probe reactions to date and propose their potential application for transmission studies. The use of RDT probe reactions in defined geographic settings for evaluating transmission is grounded in the need to easily deduce the underlying RRDR mutation to identify genotype clustering in time and space. When probe reactions between 2 strains correspond, even between distinct RDTs, the prevalence of probe reactions and associated RRDR mutations in a given setting must be considered when assessing the probability of a transmission event taking place. This probability will be relatively higher for probe reactions representing low-frequency mutations. This is explained by the principle that, in a normal distribution, the probability of plotting a low-frequency mutation in the tail region is low. Fitting it in the exact same location for the second time has much lower probability. Hence, when low-frequency probe reactions representing less-prevalent mutations correspond between 2 strains, an epidemiologic link might exist between the 2, and the occurrence of transmission can possibly be suspected. For example, Xpert MTB/RIF probe C, with the lowest frequency in Pakistan and Nigeria [20,21], is a potentially valid marker of S441L/Q transmission. To differentiate between the 2 mutations, RDT probe reactions must be combined with complementary genotyping test results. This proposed approach is evidenced by mutations L452P [23] and D435G [18] detected in strains epidemiologically linked to an extensively drug-resistant TB outbreak in KwaZulu-Natal, South Africa. Accordingly, we assert that Xpert MTB/RIF probes A, Bdelayed, C, and Edelayed; GenoType MTBDR*plus* v2.0 probes WT2, WT5, and WT6; and Genoscholar NTM + MDRTB II probes S1 and S3, which have a low frequency in specific settings [20,21] and detect less-prevalent mutations (emphasized in Figure 2), have a high potential for use in transmission studies. On the contrary, the probability of detecting a highly prevalent probe reaction in the distribution is high, but fitting it into the exact same point the second time would have lower probability. It would be unlikely that this is a random event solely attributable to high-frequency mutations being enriched by evolutionary convergence [24]. Therefore, when this is observed, further investigation of suspected cases of transmission is necessary. Particularly in settings of low RR-TB incidence, probe reactions corresponding with highly prevalent mutations such as Xpert MTB/RIF probes E, D, and B [20,21], if repeatedly obtained in the same health facility, may be potentially useful for finding an epidemiologic link between strains. The potential usefulness of RDT probe reactions for transmission studies lies in accounting for the frequency of probe reactions and RRDR mutations associated with RR-TB prevalence in a specific geographic setting. Linking this novel approach to a connectivity software can potentially bring about real-time reporting of suspected RR-TB transmission cases, which national TB control programs and public health officials may then investigate and confirm or exclude based on epidemiologic linking.

In addition to analyzing routine RDT data for evaluating transmission in detail, we found that RDT probe reactions corresponding with disputed mutations, supplemented with *rpoB* gene sequences linked to connectivity software, may assist in resolving discordance between phenotypic and genotypic rifampicin susceptibility results. Our observed probe reactions for disputed mutations were concordant between Xpert MTB/RIF and GenoType MTBDR*plus* v2.0, and were consistent with previous reports [13]. In contrast with discordant Xpert MTB/RIF and *Mycobacterium* growth indicator tube results due to the disputed mutation L430P [13,25], reliable probe reactions between RDTs were observed in this work. Consequently, detecting Xpert MTB/RIF probe A, GenoType MTBDR*plus* v2.0 probe WT2, or Genoscholar NTM + MDRTB II probe S1 would suggest an L430P mutation, despite obtaining a rifampicin-susceptible *Mycobacterium* growth indicator tube result. For other cases, disputed mutation D435Y can be distinguished from other variations at position 435, particularly from D435V, by a higher proportion of delayed results for probe B (Multimedia Appendix 3). Another delayed reaction observed was for probe E, associated with the disputed mutation L452P. Delayed results [3] for Xpert MTB/RIF probes B and E could be attributed to a reduced intensity of fluorescence in the end-probe codons. Alternatively, the specific nucleotide substitution, associated with the average proportion of delayed results and ΔCts (Multimedia Appendix 3), may have caused the delays. These inferred associations may aid in better differentiation of Xpert MTB/RIF probe reactions that have a high potential use for assessing transmission and resolving discordance between phenotypic and genotypic drug susceptibility results. On the contrary, disputed mutations H445L and H445N associated with high in vitro resistance [26] could not be distinguished from undisputed mutations in codon 445; complementary *rpoB* sequences would be very beneficial in this case. We strongly suggest that each specific mutation and RDT result be assessed case-by-case for discordant results when compared with phenotypic drug susceptibility testing and their potential for transmission clustering studies.

Scope and Limitation

The limitation of the study was the lack of strains with RRDR mutations outside labeled codon positions in Figure 1. However, to our knowledge, the uncovered mutations have not yet been encountered in clinical isolates and thus are likely to be extremely rare. This work is not a field evaluation of the potential of connected TB diagnostic data, but rather a pragmatic laboratory validation and analysis of test results.

Conclusions

We propose a novel approach for potential real-time detection of RR-TB transmission through fully using connected TB diagnostics and a shared repository of test results. To our knowledge, this is the first pragmatic and scalable work in response to the consensus of world-renowned TB experts in 2016 on the potential of diagnostic connectivity to accelerate efforts to eliminate TB [12]. This is evidenced by the ability of our proposed approach to facilitate comparison of probe reactions between different RDTs used in the same setting.

While less-prevalent RDT probe reactions corresponding with low-frequency mutations would be most suitable for evaluating transmission, combining this approach with complementary genotyping tests such as membrane-based spoligotyping, MIRU-VNTR, deep targeted sequencing, or whole genome sequencing on cultured isolates [9] would increase the resolution of the approach in identifying RR-TB transmission hotspots. Supplementing RDT test results with spoligotyping data, for instance, not only illustrates data pooling [12], but also enhances the discriminatory power of this combined approach to confirm suspected cases of TB transmission. This proposed prospective approach, inspired by the SpoNC tool, which combined *pncA* sequencing and spoligotyping data for detection of pyrazinamide-resistant TB transmission clusters [9], will be validated against conventional approaches through strains with specific mutations and originating from documented clusters.

Integrating the combined approach as an additional module in the connectivity platform will not only automate the analysis of shared RDT results and its translation into crucial clinical and public health information, but also allow more precise real-time estimation of transmitted RR-TB proportions in a population. Detection of RR-TB transmission hotspots would initiate a timely outbreak response and appropriate investigation of suspected transmission cases based on epidemiologic linking that would prevent further spread of TB.

Acknowledgments

We thank the laboratory technicians of the Institute of Tropical Medicine, Antwerp, Belgium, Mycobacteriology Unit for their assistance. NIPRO Corporation provided the claimed probe coverage of Genoscholar NTM + MDRTB II but did not influence the interpretation of results. No financial support was received from the manufacturers during the conduct of the study. KCN was supported by Erasmus Mundus Joint Doctorate Fellowship grant 2016-1346, and BCdJ, LR, and CJM were supported by European Research Council-INTERRUPTB starting grant 311725. The content is solely the responsibility of the authors and does not necessarily represent the views of the funders.

Authors' Contributions

EA, BCdJ, and KCN designed the study. LR and MD selected the strains. KCN performed all tests. KCN, CJM, GT, LG, LR, BCdJ, and EA substantially contributed to data analysis. KCN, CJM, LG, and EA were responsible for data visualization. KCN drafted the initial manuscript. KCN, CJM, GT, LR, BCdJ, and EA critically reviewed several versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Alignment of *Mycobacterium tuberculosis* strain H37Rv and rifampicin resistance–determining region mutant sequences showing alternative nucleotides and amino acids for each mutation profile analyzed.

[PNG File, 1MB - medinform_v6i1e12_app1.png]

Multimedia Appendix 2

Reproducibility of RDT probe profiles determined through mutation profiles found in multiple strains.

[PDF File (Adobe PDF File), 37KB - medinform_v6i1e12_app2.pdf]

Multimedia Appendix 3

Mutations (with wild-type and mutant nucleotide bases) that returned delayed result for Xpert MTB/RIF with corresponding nucleotide substitution type, capturing probe, location of codon with respect to claimed probe coverage, Ct, ΔCt, and average ΔCt values, and percentage of delayed result considering all strains tested.

[PDF File (Adobe PDF File), 38KB - medinform_v6i1e12_app3.pdf]

Multimedia Appendix 4

Ct curve of mutation D435Y captured by probe B with single transversion substitution type, Ct value 27.9 and ΔCt 6.

[PDF File (Adobe PDF File), 52KB - medinform_v6i1e12_app4.pdf]

Multimedia Appendix 5

Ct curve of mutation D435F captured by probe B with double transversion substitution type, Ct value 38.1 and ΔCt 16.8.

[PDF File (Adobe PDF File), 51KB - medinform_v6i1e12_app5.pdf]

Multimedia Appendix 6

Ct curve of mutation D435V captured by probe B with single transversion substitution type, Ct value 38.6 and ΔCt 17.

[PDF File (Adobe PDF File), 51KB - medinform_v6i1e12_app6.pdf]

Multimedia Appendix 7

Ct curve of mutation H445R captured by probe D with single purine transition substitution type, Ct value 27.5 and ΔCt 8.2.

[PDF File (Adobe PDF File), 51KB - medinform_v6i1e12_app7.pdf]

Multimedia Appendix 8

Ct curve of mutation L452P captured by probe E with single pyrimidine transition substitution type, Ct value 24.5 and ΔCt 6.

[PDF File (Adobe PDF File), 52KB - medinform_v6i1e12_app8.pdf]

References

1. World Health Organization. Global tuberculosis report 2017. Geneva, Switzerland: WHO; 2017. URL: http://apps.who.int/iris/bitstream/10665/259366/1/9789241565516-eng.pdf [accessed 2018-02-08] [WebCite Cache ID 6x58XZeiJ]

2. Andre E, Goeminne L, Cabibbe A, Beckert P, Kabamba MB, Mathys V, et al. Consensus numbering system for the rifampicin resistance-associated rpoB gene mutations in pathogenic mycobacteria. Clin Microbiol Infect 2017 Mar;23(3):167-172 [FREE Full text] [doi: 10.1016/j.cmi.2016.09.006] [Medline: 27664776]
3. Lawn SD, Nicol MP. Xpert MTB/RIF assay: development, evaluation and implementation of a new rapid molecular diagnostic for tuberculosis and rifampicin resistance. Future Microbiol 2011 Sep;6(9):1067-1082. [doi: 10.2217/fmb.11.84] [Medline: 21958145]
4. Hoffmann H, Hillemann D, Rigouts L, Deun AV, Kranzer K. How should discordance between molecular and growth-based assays for rifampicin resistance be investigated? Int J Tuberc Lung Dis. Int J Tuberc Lung Dis 2017;21:721-726. [doi: 10.5588/ijtld.17.0140] [Medline: 28633695]
5. Chakravorty S, Kothari H, Aladegbami B, Cho EJ, Lee JS, Roh SS, et al. Rapid, high-throughput detection of rifampin resistance and heteroresistance in Mycobacterium tuberculosis by use of sloppy molecular beacon melting temperature coding. J Clin Microbiol 2012 Jul;50(7):2194-2202 [FREE Full text] [doi: 10.1128/JCM.00143-12] [Medline: 22535987]
6. Pai M, Nicol MP, Boehme CC. Tuberculosis diagnostics: state of the art and future directions. Microbiol Spectr 2016 Oct;4(5):1-15. [doi: 10.1128/microbiolspec.TBTB2-0019-2016] [Medline: 27763258]
7. Cox H, Dickson-Hall L, Ndjeka N, van't Hoog A, Grant A, Cobelens F, et al. Delays and loss to follow-up before treatment of drug-resistant tuberculosis following implementation of Xpert MTB/RIF in South Africa: a retrospective cohort study. PLoS Med 2017;14:1-19. [doi: 10.1371/journal.pmed.1002238] [Medline: 28222095]
8. Fox GJ, Schaaf HS, Mandalakas A, Chiappini E, Zumla A, Marais BJ. Preventing the spread of multidrug-resistant tuberculosis and protecting contacts of infectious cases. Clin Microbiol Infect 2017 Mar;23(3):147-153. [doi: 10.1016/j.cmi.2016.08.024] [Medline: 27592087]
9. Walker TM, Merker M, Kohl TA, Crook DW, Niemann S, Peto TEA. Whole genome sequencing for M/XDR tuberculosis surveillance and for resistance testing. Clin Microbiol Infect 2017 Mar;23(3):161-166. [doi: 10.1016/j.cmi.2016.10.014] [Medline: 27789378]
10. Said HM, Kushner N, Omar SV, Dreyer AW, Koornhof H, Erasmus L, et al. A novel molecular strategy for surveillance of multidrug resistant tuberculosis in high burden settings. PLoS One 2016 Jan;11(1):e0146106 [FREE Full text] [doi: 10.1371/journal.pone.0146106] [Medline: 26752297]
11. Kendall E, Fofana M, Dowdy D. Burden of transmitted multidrug resistance in epidemics of tuberculosis: a transmission modelling analysis. Lancet Respir Med 2015 Dec;3(12):963-972. [doi: 10.1016/S2213-2600(15)00458-0] [Medline: 26597127]
12. Andre E, Isaacs C, Affolabi D, Alagna R, Brockmann D, de Jong B, et al. Connectivity of diagnostic technologies: improving surveillance and accelerating tuberculosis elimination. Int J Tuberc Lung Dis 2016;20:999-1003. [doi: 10.5588/ijtld.16.0015] [Medline: 27393530]
13. Van Deun A, Aung KJM, Hossain A, de Rijk P, Gumusboga M, Rigouts L, et al. Disputed rpoB mutations can frequently cause important rifampicin resistance among new tuberculosis patients. Int J Tuberc Lung Dis 2015 Feb;19(2):185-190. [doi: 10.5588/ijtld.14.0651] [Medline: 25574917]
14. Rigouts L, Gumusboga M, de Rijk WB, Nduwamahoro E, Uwizeye C, de Jong B, et al. Rifampin resistance missed in automated liquid culture system for Mycobacterium tuberculosis isolates with specific rpoB mutations. J Clin Microbiol 2013 Aug;51(8):2641-2645 [FREE Full text] [doi: 10.1128/JCM.02741-12] [Medline: 23761146]
15. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 2012 Apr 27;28(12):1647-1649. [doi: 10.1093/bioinformatics/bts199] [Medline: 22543367]
16. Serif (Europe) Ltd. Affinity - Professional creative software. 2017. URL: https://affinity.serif.com/en-us/ [accessed 2017-10-30] [WebCite Cache ID 6ubjFgkLG]
17. Gagneux S. The competitive cost of antibiotic resistance in Mycobacterium tuberculosis. Science 2006 Jun 30;312(5782):1944-1946. [doi: 10.1126/science.1124410] [Medline: 16809538]
18. Cohen K, Abeel T, Mcguire A, Desjardins C, Munsamy V, Shea T, et al. Evolution of extensively drug-resistant tuberculosis over four decades: whole genome sequencing and dating analysis of Mycobacterium tuberculosis isolates from KwaZulu-Natal. PLoS Med 2015;12:1-22. [doi: 10.1371/journal.pmed.1001880] [Medline: 26418737]
19. Georghiou SB, Seifert M, Catanzaro D, Garfein RS, Valafar F, Crudu V, et al. Frequency and distribution of tuberculosis resistance-associated mutations between Mumbai, Moldova, and Eastern Cape. Antimicrob Agents Chemother 2016 Jul;60(7):3994-4004 [FREE Full text] [doi: 10.1128/AAC.00222-16] [Medline: 27090176]
20. Ullah I, Shah AA, Basit A, Ali M, khan A, Ullah U, et al. Rifampicin resistance mutations in the 81 bp RRDR of rpoB gene in Mycobacterium tuberculosis clinical isolates using Xpert MTB/RIF in Khyber Pakhtunkhwa, Pakistan: a retrospective study. BMC Infect Dis 2016 Aug 12;16(413):1-6. [doi: 10.1186/s12879-016-1745-2] [Medline: 27519406]
21. Ochang EA, Udoh UA, Emanghe UE, Tiku GO, Offor JB, Odo M, et al. Evaluation of rifampicin resistance and 81-bp rifampicin resistant determinant region of rpoB gene mutations of Mycobacterium tuberculosis detected with XpertMTB/Rif in Cross River State, Nigeria. Int J Mycobacteriol 2016 Dec;5 Suppl 1:S145-S146 [FREE Full text] [doi: 10.1016/j.ijmyco.2016.09.007] [Medline: 28043515]

22. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis drug resistance mutation database. PLoS Med 2009 Feb 10;6(2):e2 [FREE Full text] [doi: 10.1371/journal.pmed.1000002] [Medline: 19209951]
23. Ioerger TR, Koo S, No E, Chen X, Larsen MH, Jacobs WR, et al. Genome analysis of multi- and extensively-drug-resistant tuberculosis from KwaZulu-Natal, South Africa. PLoS One 2009 Nov 05;4(11):e7778 [FREE Full text] [doi: 10.1371/journal.pone.0007778] [Medline: 19890396]
24. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat Genet 2013 Oct;45(10):1183-1189 [FREE Full text] [doi: 10.1038/ng.2747] [Medline: 23995135]
25. Gurbanova E, Mehdiyev R, Blondal K, Tahirli R, Mirzayev F, Hillemann D, et al. Mitigation of discordant rifampicin-susceptibility results obtained by Xpert Mycobacterium tuberculosis/rifampicin and Mycobacterium growth indicator tube. Microb Drug Resist 2017 Apr 27:1-8. [doi: 10.1089/mdr.2016.0149] [Medline: 28447869]
26. Van Deun A, Barrera L, Bastian I, Fattorini L, Hoffmann H, Kam KM, et al. Mycobacterium tuberculosis strains with highly discordant rifampin susceptibility test results. J Clin Microbiol 2009 Nov;47(11):3501-3506 [FREE Full text] [doi: 10.1128/JCM.01209-09] [Medline: 19759221]

Abbreviations

CFU: colony-forming unit
MDR-TB: multidrug-resistant tuberculosis
RDT: rapid diagnostic test
RRDR: rifampicin resistance–determining region
RR-TB: rifampicin-resistant tuberculosis
TB: tuberculosis

Edited by G Eysenbach; submitted 01.11.17; peer-reviewed by M Macaraig, R Pankomera; comments to author 20.12.17; revised version received 09.01.18; accepted 09.01.18; published 27.02.18.

Please cite as:
Ng KC, Meehan CJ, Torrea G, Goeminne L, Diels M, Rigouts L, de Jong BC, André E
Potential Application of Digitally Linked Tuberculosis Diagnostics for Real-Time Surveillance of Drug-Resistant Tuberculosis Transmission: Validation and Analysis of Test Results
JMIR Med Inform 2018;6(1):e12
URL: http://medinform.jmir.org/2018/1/e12/
doi:10.2196/medinform.9309
PMID:29487047

Original Paper

Secure and Efficient Regression Analysis Using a Hybrid Cryptographic Framework: Development and Evaluation

Md Nazmus Sadat[1], BCompSc; Xiaoqian Jiang[2], PhD; Md Momin Al Aziz[1], MSc; Shuang Wang[2], PhD; Noman Mohammed[1], PhD

[1]Department of Computer Science, University of Manitoba, Winnipeg, MB, Canada
[2]Department of Biomedical Informatics, University of California San Diego, La Jolla, CA, United States

Corresponding Author:
Md Nazmus Sadat, BCompSc
Department of Computer Science
University of Manitoba
E2 EITC
Winnipeg, MB, R3T2N2
Canada
Phone: 1 858 375 6047
Email: sadat@cs.umanitoba.ca

Abstract

Background: Machine learning is an effective data-driven tool that is being widely used to extract valuable patterns and insights from data. Specifically, predictive machine learning models are very important in health care for clinical data analysis. The machine learning algorithms that generate predictive models often require pooling data from different sources to discover statistical patterns or correlations among different attributes of the input data. The primary challenge is to fulfill one major objective: preserving the privacy of individuals while discovering knowledge from data.

Objective: Our objective was to develop a hybrid cryptographic framework for performing regression analysis over distributed data in a secure and efficient way.

Methods: Existing secure computation schemes are not suitable for processing the large-scale data that are used in cutting-edge machine learning applications. We designed, developed, and evaluated a hybrid cryptographic framework, which can securely perform regression analysis, a fundamental machine learning algorithm using somewhat homomorphic encryption and a newly introduced secure hardware component of Intel Software Guard Extensions (Intel SGX) to ensure both privacy and efficiency at the same time.

Results: Experimental results demonstrate that our proposed method provides a better trade-off in terms of security and efficiency than solely secure hardware-based methods. Besides, there is no approximation error. Computed model parameters are exactly similar to plaintext results.

Conclusions: To the best of our knowledge, this kind of secure computation model using a hybrid cryptographic framework, which leverages both somewhat homomorphic encryption and Intel SGX, is not proposed or evaluated to this date. Our proposed framework ensures data security and computational efficiency at the same time.

(JMIR Med Inform 2018;6(1):e14) doi:10.2196/medinform.8286

KEYWORDS

privacy-preserving regression analysis; Intel SGX; somewhat homomorphic encryption

Introduction

Machine learning algorithms are now being widely used in many applications to uncover deep and predictive insights from datasets that are large scale and diverse. For instance, building predictive models from biomedical data is very important in biomedical science. Such predictive models can identify genetic risk factors for a specific disease under study and can guide medical treatment. For instance, Tabaei and Hermana formulated a predictive equation to screen for diabetes [1].

Machine learning thrives on growing datasets. In most of the cases, the more data fed into a machine learning system, the more it can learn and offer the potential to make more accurate prediction. It is often known as “data never hurt in machine learning," as insufficient information cannot lead to powerful learning systems. In the context of health care, building an accurate predictive model depends on the quality and quantity of aggregate clinical data, which come from different hospitals or health care institutions. Consequently, in a real-world scenario, machine learning applications use data from several sources, including genetic and genomic, clinical, and sensor data. Day by day, many new sources of data are becoming available—for instance, data from cell phones [2], wearable sensors [3], and participatory sensing applications [4]. For instance, there are wearable sensing frameworks that collect sensing information regarding heart rate, body temperature, caloric expenditure, etc, to train machine learning models. These models are then used for predictive analysis [4].

Data collection, storage, and processing power of a single institution is not always adequate to handle the large-scale data used in cutting-edge machine learning applications. For rare diseases, individual institutions oftentimes do not have sufficient data to calculate a model to achieve sufficient statistical power. Therefore, data sharing among multiple institutions is required. However, sharing sensitive biomedical data (clinical or genomic) exposes many security and privacy threats [5]. In case of data breach, there is a risk of sensitive personal information leakage. Therefore, in addition to addressing the fundamental goal of information retrieval, privacy-preserving learning also requires the learning algorithm to protect the confidentiality of the sensitive records of individuals. Along with obtaining the approval from an institutional review board, collaborative research on shared biomedical data often needs to satisfy 2 criteria at the same time: (1) permitting access to biomedical data for collaborative research, and (2) maintaining participants' privacy and protecting the confidentiality of their genomic and clinical profile [6]. For this reason, strict policies regarding biomedical data sharing have been enforced and, generally, these policies are different in different regions of the world. For instance, there are several key differences between the US Health Insurance Portability and Accountability Act (HIPAA) and the Canadian Personal Information Protection and Electronic Documents Act (PIPEDA). This difference in the policies and regulations of cross-border biomedical data sharing impedes international research projects greatly [7]. It is imperative to address this problem with practical solutions to promote health science discoveries.

In this paper, we concentrate on secure and efficient computation for a fundamental technique used in numerous learning algorithms called *regression* (see Methods). Regression analysis identifies the correlation among different attributes based on input data. Given a number of high-dimensional data points, regression analysis generates a best-fit line or curve through these points. To evaluate the fit, the value of a target attribute is predicted, which is associated with the given values of input. For instance, the input variables can be an individual's age, weight, sex, body mass index, and glucose level, while the output can be the likelihood to develop diabetes. Although regression analysis is widely used in practice, little work has been done in privacy-preserving regression analysis over a distributed dataset. Our objective was to perform the required computation for regression analysis without exposing any other information of user data.

Prior Works

To ensure the security and privacy of the sensitive data used in learning algorithm, different techniques (eg, garbled circuit [8], homomorphic encryption [9], differential privacy [10], and secure hardware [11]) have been adopted (Multimedia Appendix 1 discusses prior works targeting regression). But each of these techniques has certain shortcomings (eg, computational overhead, communication overhead, storage overhead, reduced data utility, and approximation error), which make these techniques difficult to use in real-world applications.

Wu et al developed a framework, grid binary logistic regression (GLORE) [12], for developing a binary logistic regression model where data are distributed across different data owners. In their proposed approach, instead of sharing patient records, data owners send intermediary results to a central entity. These intermediary results are then used to build a prediction model without sharing patient-level data. However, in their approach, the intermediary results are exchanged in plaintext. If the data size of a data owner is small, then sharing the intermediary results might compromise privacy.

Later, Shi et al incorporated secure multiparty computation in GLORE. Their proposed framework, secure multiparty computation framework for grid logistic regression (SMAC-GLORE) [13], protects the confidentiality of intermediary results beside the patient data. However, SMAC-GLORE cannot handle numbers outside of a predefined range, and it does not scale well (eg, it cannot efficiently handle data with more than 10 covariates). In addition, it uses a Taylor series approximation approach to evaluate the logit function. This approximation causes precision loss in the final output.

Why Hybrid?

There are two obvious but suboptimal solutions in terms of security and efficiency. Existing fully homomorphic encryption (FHE) techniques [14] provide rigorous security, but these solutions are not efficient. In existing homomorphic encryption schemes, with subsequent homomorphic operations, the noise (and size) of the ciphertext grows substantially, which increases computational and storage overheads to a great extent (see Methods, Homomorphic Encryption for details). There are some operations to reduce the size and noise of the ciphertext: *bootstrapping* [9] and *relinearization* [15]. However, these operations are very expensive from the computational point of view. Our proposed framework does not use these expensive operations at all, which enhances the efficiency of the framework greatly.

On the contrary, Software Guard Extensions (SGX; Intel)-based solutions are very efficient but have some security concerns resulting from the recent discovery of side-channel attacks against SGX [16]. We developed our method so that only intermediary results, not individual records, are decrypted inside the secure hardware. Hence, a successful adversary would be unable to compromise the privacy of an individual.

Our proposed hybrid framework uses both techniques and provides a good trade-off in terms of security and efficiency.

Contributions

In this paper, we propose a hybrid cryptographic framework for secure and efficient regression analysis (both linear and logistic). Our proposed framework leverages the best features of two secure computation schemes: somewhat homomorphic encryption (SWHE) and secure hardware (Intel SGX). In this framework, data reside at the data owner's end. We assumed that data are horizontally partitioned, where all the records share same attributes. Inspired by GLORE [12], we formulated the regression problem as decomposable parts. Data owners compute these decomposable intermediary results locally. Then, after encrypting these local results using homomorphic encryption, they send the encrypted intermediary results to an SGX-enabled central server. The central server now combines the intermediary results using a homomorphic addition operation. Then, these aggregate encrypted intermediary results are passed to the secure hardware hosted at the central server. Here, the aggregate intermediary results are decrypted and further computation is performed on plaintext. These computations involve matrix inversion and division, which are hard to handle in existing homomorphic encryption schemes. Finally, model coefficients are computed inside the secure hardware.

We summarize our contributions as follows: (1) We address the limitations of existing secure computation schemes and propose a hybrid secure computation model for performing regression analysis over distributed data, which is more efficient and robust. (2) We designed the framework in such a way that no homomorphic multiplication is necessary, which is an expensive operation. In addition, we do not need any bootstrapping or relinearization operation. (3) In our proposed approach, a significant portion of computation is performed at the data owner's end on plaintext. In computation at a central server, after homomorphic addition operations, further computation is performed inside secure hardware on plaintext. Since most of the operations are performed on plaintext, our proposed approach is very efficient. In addition, due to avoiding any kind of approximation technique, our proposed method does not introduce any precision loss in the final output.

In Multimedia Appendix 1 we introduce major existing secure computation techniques, application of these techniques in regression analysis, and their shortcomings.

Methods

Security Background

Homomorphic Encryption

The idea of an encryption scheme that is capable of performing arbitrary computation on encrypted data was first proposed by Rivest et al [17] in 1978. Since then, several cryptosystems were invented that are homomorphic with respect to either addition or multiplication. Finally, Boneh et al [18] proposed a partially homomorphic cryptosystem that is able to perform 1 multiplication and any number of additions. Table 1 shows a partial list of homomorphic encryption schemes [18-22].

Developing an encryption scheme that supports an arbitrary number of additions and multiplications was an open problem until 2009. Since addition and multiplication operations over integer ring Z_2 form a complete set of operations, this type of encryption scheme supports any polynomial time computation on ciphertext. In 2009, Gentry showed the first construction of an FHE scheme [9] that can do any number of addition and multiplication operations on encrypted data.

To explain FHE, say ciphertext c_i is the encrypted form of plaintext m_i, where m_i and c_i are elements of a ring (the operations of the ring are addition and multiplication). In FHE, if a function *f* consists of addition and multiplication in the ring, then *decryption* $(f\,(c_1,c_2,\ldots,c_n)) = f\,(m_1,m_2,\ldots,m_n)$. Generally, *f* is expressed by an arithmetic circuit over Gallois field(2). This is equivalent to a Boolean circuit with exclusive OR and AND gates.

In the existing FHE schemes, a certain amount of noise needs to be introduced in the ciphertexts to ensure data confidentiality. This noise grows while performing homomorphic operations on ciphertexts. In particular, a homomorphic multiplication operation increases the size of the ciphertext abruptly. For instance, if 2 input ciphertexts have size *M* and *N*, then the output ciphertext will be of size $M+N-1$. If the amount of noise becomes too high, then the ciphertext cannot be decrypted correctly. To perform any number of homomorphic operations, the noise of the ciphertexts needs to be reduced. As mentioned before, this can be done using a method known as *bootstrapping* [9], which is computationally expensive.

In use cases where only a predetermined number of computational operations needs to be done, the costly bootstrapping process can be avoided by using an SWHE scheme [23]. This scheme is often more efficient than using an FHE scheme with bootstrapping. SWHE schemes use a method called *relinearization* [15,24] to reduce the size of the ciphertext.

Table 1. Partial list of homomorphic encryption schemes.

Cryptosystem	Homomorphism
Goldwasser and Micali [19], Paillier [20]	Additive
Rivest et al [21], ElGamal [22]	Multiplicative
Boneh et al [18]	Both

Intel Software Guard Extensions

Intel SGX is a collection of extensions to the Intel architecture that mostly concentrates on the issue of running applications on a remote machine managed by an untrusted party. SGX enables parts of an application to run within secure portions of the central processing unit called *enclaves*. Untrusted entities, including system software, cannot access the enclave. SGX guarantees that the code and information inside an enclave cannot be manipulated from outside the enclave. Two SGX features facilitate provisioning of sensitive data to an enclave: attestation and sealing.

SGX enclaves are generated without privacy-sensitive information. Privacy-sensitive information is provisioned after the enclave has been appropriately instantiated. This process of demonstrating that an application has been correctly instantiated within an enclave is called *attestation* [25].

At the point when an enclave is instantiated, SGX protects its data until they are kept within the enclave. In any case, when the enclave procedure terminates, the enclave will be destroyed and all related data will be lost. So, for later use, data should be stored outside the enclave. *Sealing* is the procedure that is used to store encrypted data to ensure that only the same enclave would be capable of unsealing them back to their previous form.

System Architecture

Our proposed framework has three main entities (Figure 1).

Data Owners

These parties are geographically distributed and possess databases. Data can come from a variety of sources, including cell phones, wearable sensors, and relational databases. Data owners send encrypted intermediary results to the central server so that it can analyze the combined dataset.

Key Manager

This generates and distributes the cryptographic keys that will be used for data encryption and decryption in different stages of our proposed framework. Each data owner gets a public key from the key manager and uses it for encrypting data.

Central Server

The central server maintains communication with all the other entities of the framework. It receives data from the data owners and computes the final result using SWHE and secure hardware.

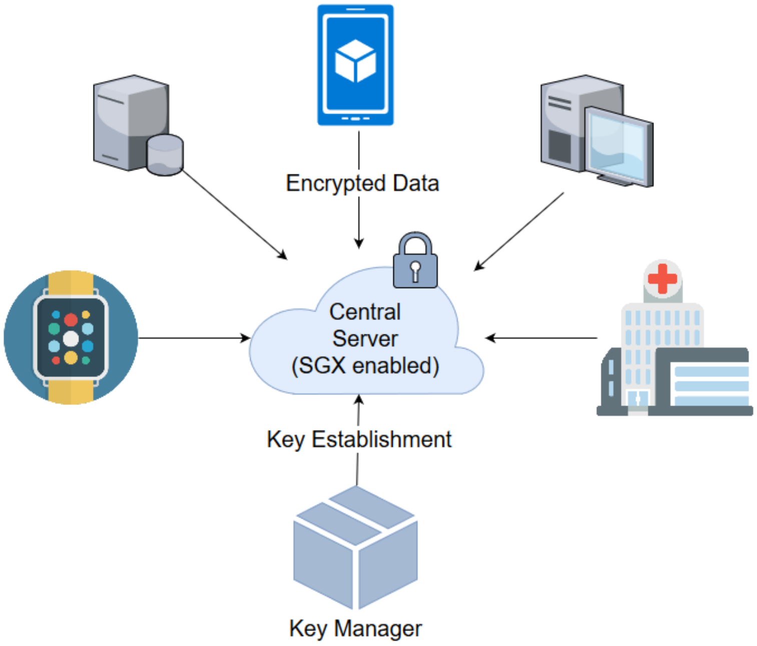
Threat Model

In proposing this framework, our goal was to guarantee the confidentiality of data provided by different data owners. We assume that the central server is a semihonest party (also referred to as honest-but-curious), where it obeys the system protocol but may try to infer sensitive information by analyzing the system logs or received information [26].

We assume that the computation runs in an SGX-enabled central server. SGX architecture enables the central server to perform any computation securely on data provided by different data owners. We assume that the processor of the central server works properly and is not compromised. We trust the design and implementation of SGX and all cryptographic operations performed by it.

In general, side-channel attacks against SGX can be classified into two categories: physical attacks (where the attacker has physical access to the machine) and software attacks (these are launched by any malicious software running in the same machine) [27]. There has been no known successful physical attack against SGX. However, it is possible to exploit a type of software attack known as a *synchronization bug* [28]. Synchronization bugs are possible to exploit because an untrusted operating system can manipulate the thread scheduling of enclaves. However, it is only applicable for multithreaded applications, whereas our application is single threaded.

Figure 1. Block diagram of the system architecture. SGX: Software Guard Extensions.



There is another type of well-known software attack, which is called a *page-fault attack* [16]. As the page tables are maintained in the operating system kernel and operated by the untrusted system software, page table entries can be manipulated to attack enclaves. But, since enclave pages are permission protected, malicious system software cannot compromise their integrity by manipulating them. However, Xu et al [16] showed that, by clearing the present flag in the corresponding page table entries, the malicious software can generate traces of page access from the enclave. Although an adversary can observe access to different enclave pages, enclave memory can be treated as private at page-level granularity (4 kB) [29]. In other words, a different access to an enclave page is indistinguishable to an adversary. Further research is required to better understand the gap between the potential vulnerabilities of SGX and proposed defense mechanisms. Most of the existing defense mechanism have been developed to address the page-fault side-channel attacks [29-31]. However, these mechanisms may not be effective for future attacks. Keeping these attacks in mind, we developed our framework to protect institutional privacy by combining the local inputs of participating institutions without decrypting them, therefore providing a higher layer of protection without introducing too much computational overhead.

We did not consider the aspects of adversarial machine learning through obtained outputs. Adversarial parties may try to infer sensitive attributes of data by model inversion attacks [32,33].

Linear Regression

Suppose we are given a set of paired observations (x_i, y_i) for i=1,2,...,n, and we want to generate the best-fit straight line for these points. This straight line is given by $y=\beta_1+\beta_2x$, for some β_1,β_2. The purpose is to explain the correlation between variable y and x. To evaluate the fit, the value of y is predicted that is associated with a given value of x. In the literature, y is called *the variable to be explained* (or the *dependent* variable) and x is called the *explanatory variable* (the *regressor*, the *covariate*, or the *independent* variable) [34] (pg 79). Consider the following simple linear regression model: $y=\beta_1+\beta_2x+\varepsilon$. Here, ε is the error we make in predicting y. For i=1,...,n, we obtain n equations: $y_1=\beta_1+\beta_2x_1+\varepsilon_1$, $y_2=\beta_1+\beta_2x_2+\varepsilon_2$, and $y_n=\beta_1+\beta_2x_n+\varepsilon_n$.

We can formulate this regression model using the matrix in Figure 2 (a).

Figure 2. Equations used in developing the framework.

(a)
$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

(b)
$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{21} & \cdots & x_{k1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

(c)
$$X^TX = \sum_{i=1}^{n} X_i^TX_i \ , \ X^TY = \sum_{i=1}^{n} X_i^TY_i$$

(d)
$$\begin{aligned} &\operatorname{logit}\left(P(y=1|x_1,x_2,\ldots\ldots,x_k)\right) \\ &= \log\frac{P(y=1|x_1,x_2,\ldots\ldots,x_k)}{1-P(y=1|x_1,x_2,\ldots\ldots,x_k)} \\ &= \beta_1+\beta_2x_2+\ldots.+\beta_kx_k \end{aligned}$$

(e)
$$\beta^{new} \xleftarrow[\beta]{\text{argmin}} (z-X\beta)^TW(z-X\beta)$$

(f)
$$\beta^{new} = \beta^{old} + (X^T\tilde{X})^{-1}X^T(Y-P)$$

(g)
$$X^T\tilde{X} = \sum_{i=1}^{n} X_i^T\tilde{X}_i \ , X^T(Y-P) = \sum_{i=1}^{n} X_i^T(Y_i-P_i)$$

In this way, the simple linear regression function can be represented by a short and simple equation:

The linear regression model with several explanatory variables is known as *multiple linear regression*. This is given by

Here, x_{1i}=1, for i=1,...,n. The function of Equation 2 can also be expressed in matrix form, which is more convenient, as in Figure 2 (b).

It is noteworthy that Equation 1 is also applicable for multiple linear regression.

Using the ordinary least squares estimate technique we can show that $\beta=(X^TX)^{-1}X^TY$ (for details, see Heij et al [34], pg 79).

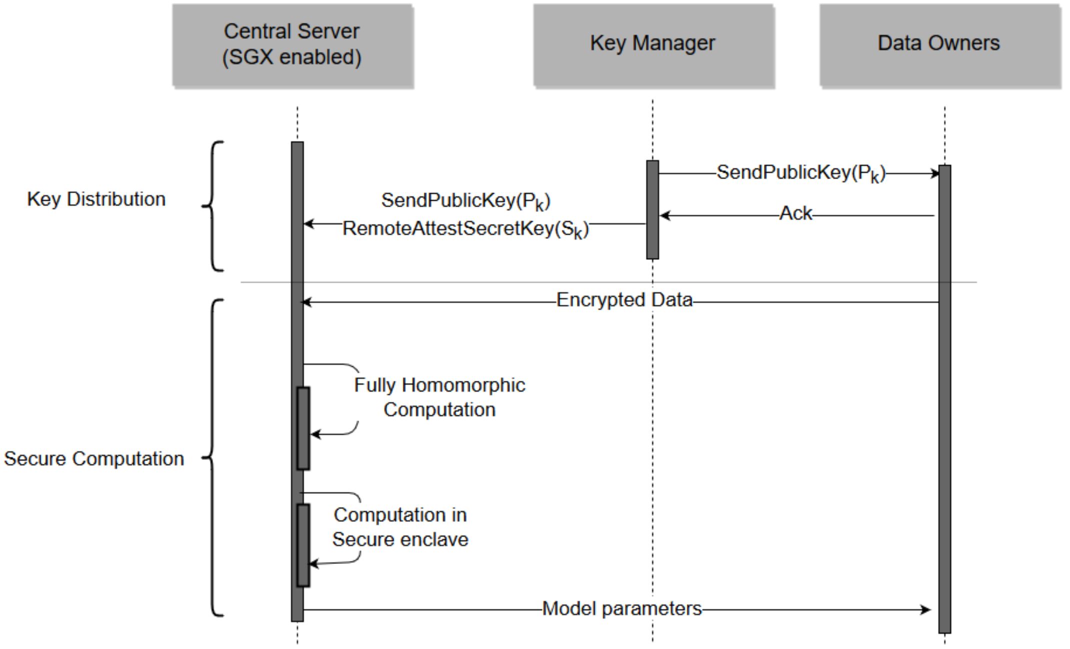
For secure linear regression over distributed data, each data owner D_i computes $X^T{}_iX_i$ and $X^T{}_iY_i$ locally on plaintext. D_i then encrypts $X^T{}_iX_i$ and $X^T{}_iY_i$ using homomorphic encryption. After receiving these intermediary results from all of the data owners, the central server then adds these using homomorphic addition operations to construct X^TY and X^TX (equation from Figure 2 [c]). Further computation is performed inside the enclave after decryption. Textbox 1 shows our secure linear regression algorithm.

Figure 3 illustrates the sequence diagram of our proposed method. At first, the key manager establishes the public key and the private key. The private key is sent to the central server securely using remote attestation. The data owners then encrypt their data with the public key and send the encrypted data to the central server. Finally, the central server computes the model parameters.

Textbox 1. Algorithm 1: secure linear regression.

```
Input: Each data owner $D_i$ provides encrypted $X^T{}_iX_i$ and $X^T{}_iY_i$.
Output: Model parameters ($\beta$)
1. Perform homomorphic addition over $X^T{}_iX_i$ for each data owner $i$.
2. Perform homomorphic addition over $X^T{}_iY_i$ for each data owner $i$.
3. Send $X^TY$ and $X^TX$ to enclave.
4. Inside enclave, decrypt encrypted $X^TY$ and $X^TX$.
5. Inside enclave, compute $(X^TX)^{-1}$.
6. Finally, compute $\beta$ inside enclave.
```

Figure 3. Sequence diagram of our proposed framework. Ack: acknowledge; SGX: Software Guard Extensions.



Logistic Regression

Logistic regression extends the principles of multiple linear regression to the case where the dependent variable y is binary (either 0 or 1). Like in multiple linear regression, the independent variables can be categorical or continuous.

Instead of modeling the dependent variable directly, logistic regression models the probability of the dependent variable. Logistic regression uses the equation of linear regression equation (2). But, in that equation, the value of the dependent variable can fall outside [0, 1]. Therefore, a nonlinear transformation is used, which is called *logit transformation*. The logit function takes any value x and maps it onto a value between 0 and 1. Logit function is given by logit(x)=log[p /(1–p)] as in Figure 2 (d). Therefore, *probability*=(y=1| x_1, x_2,...,x_k) = [exp(β_1+$\beta_2 x_2$+...+$\beta_k x_k$)]/[1+exp(β_1+$\beta_2 x_2$+...+$\beta_k x_k$)] where β_1, β_2,...,B_k are unknown constants analogous to the multiple linear regression model. *Probability*=(y=1| x_1, x_2,...,x_k) denotes the probability that input (x_1, x_2,...,x_k) belongs to default class (y=1).

Logistic regression models are generally fit by maximum likelihood by using the conditional probability of y given x. Here, the Newton-Raphson method is used to solve the coefficients.

Let X represent the matrix of x_i values, Y represent the vector of y_i values, P be the vector of fitted probabilities with the ith element p (x_i;β^{old}), and W be an $n \times n$ diagonal matrix of weights with ith diagonal element p (x_i;β^{old})(1– p [x_i;β^{old}]). Then a Newton step is as follows:

In the second and third steps, the Newton step is expressed as a weighted least squares step, with the response $z= X\beta^{old}+W^{-1}(Y–P)$. This method is also known as iteratively reweighted least squares, since each iteration solves the weighted least squares problem (see Friedman et al [35] for details), as in Figure 2 (e).

In practice, the W matrix is not computed explicitly because its size could be huge. If we have 1000 rows of training data, matrix W would have 1,000,000 cells. For this reason, direct matrix operations with W may be very inefficient. Notice the beta update equation (Equation 3) has a term, WX, which means the matrix product of W and X. Because most of the values in W are zero, most of the matrix multiplication terms are also zero. This allows W times X to be computed directly from P and X, without explicitly constructing W. Several of the mathematical references that describe iteratively reweighted least squares with the Newton-Raphson algorithm for logistic regression use the symbol [X tilde] for the product of W and X. It is generally written as in Figure 2 (f).

For secure logistic regression over distributed data, each data owner D_i computes $X^T{}_i$[X tilde]$_i$ and $X^T{}_i(Y_i– P_i)$ locally on plaintext. D_i then encrypts $X^T{}_i$[X tilde]$_i$ and $X^T{}_i(Y_i– P_i)$ using homomorphic encryption. After receiving these intermediary results from all the data owners, the central server then adds these using homomorphic addition operations to construct X^T[X tilde] and $X^T(Y– P)$ (equation from Figure 2 [g]). Further computation is performed inside the enclave after decryption. After computing β, the central server sends β to all of the data owners. For the next iteration, data owner i computes $X^T{}_i$[X tilde]$_i$ and $X^T{}_i(Y_i– P_i)$ using new β (received from the central server) and sends these intermediary results to the central server. The central server then updates β using newly received $X^T{}_i$[X tilde]$_i$ and $X^T{}_i(Y_i– P_i)$. In this way, iterations continue until parameters converge. Textbox 2 shows our secure logistic regression algorithm.

Implementation

We developed our proposed framework using C++. For SWHE, we used the Simple Encrypted Arithmetic Library (SEAL) [24]. SEAL is an easy-to-use homomorphic encryption library, with no external dependencies. There is another homomorphic encryption framework called HElib [36], but we chose to use SEAL for its simplicity.

Textbox 2. Algorithm 2: secure logistic regression.

Input: Each data owner D_i provides encrypted $X^T{}_i$[X tilde]$_i$ and $X^T{}_i(Y_i– P_i)$, and β is initialized to an all-zero vector.

Output: Model parameters

1. Receive encrypted $X^T{}_i$[X tilde]$_i$ and $X^T{}_i(Y_i– P_i)$ from each data owner D_i.
2. Perform homomorphic addition over $X^T{}_i$[X tilde]$_i$ for each data owner D_i.
3. Perform homomorphic addition over $X^T{}_i(Y_i– P_i)$ for each data owner D_i.
4. Send encrypted X^T[X tilde] and $X^T(Y– P)$ to enclave.
5. Inside enclave, decrypt X^T[X tilde] and $X^T(Y– P)$.
6. Update $\beta^{new}=\beta^{old}+(X^T[X \text{ tilde}])^{-1}X^T(Y– P)$.
7. If the stopping criteria are satisfied, then stop; otherwise, send β to each data owner and go to step 1.

Table 2. Parameters used for the Simple Encrypted Arithmetic Library.

Parameters	Value
Polynomial modulus	$x^{1024}+1$
Plaintext modulus	1<<8
Decomposition bit count	32
No. of coefficients reserved for fractional part	64

Table 3. Size of datasets used for experiments.

Records	Dataset	
	Haberman	Low Birth Weight Study
No. of instances	270	488
No. of features	3	8

Experimental Settings and Dataset

We performed experiments in a machine with an Intel Core i7-6700 (3.40 GHz) processor and 8 GB memory (Intel Corporation, Santa Clara, CA, USA). We used Intel SGX software development kit version 1.7. We simulated 2 data owners and the central server in this machine. Table 2 shows the SEAL parameters.

We performed experiments using Haberman’s survival dataset from the University of California, Irvine, Machine Learning Repository [37] and the Longitudinal Low Birth Weight Study dataset from Hosmer and Lemeshow [38]. The records of the datasets were evenly distributed between the 2 data owners.

Table 3 lists the datasets we used with their sizes.

Results

Table 4 shows the experimental results. For SWHE, most of the computation time was due to homomorphic operations. Our proposed framework avoided expensive homomorphic multiplication by transferring the later phase of computation to the secure hardware. In addition, we needed to decrypt only the intermediary results, not every individual attribute value. Consequently, our proposed framework was more efficient than the solely secure hardware (SWHE)-based technique (where every individual attribute needs to be decrypted) and the SWHE-based technique (which involves many expensive homomorphic multiplication and relinearization operations). Table 4 does not report the results for the SWHE-based technique. However, according to our empirical results, it took more than 2 hours for the Haberman dataset and more than 17 hours for the Low Birth Weight Study dataset for both kinds of regression analyses.

Table 4. Experimental results for computation time.

Regression analyses	Dataset	
	Haberman	Low Birth Weight Study
Linear regression		
Plaintext (ms)	6	25
Proposed method (s)	8.991	39.382
Secure hardware (SWHE[a]) (s)	259.908	880.228
Secure hardware (AES[b]) (s)	4.30	8.54
Logistic regression		
Plaintext (ms)	171	886
Proposed method (s)	27.037	162.544
Secure hardware (SWHE) (s)	264.669	904.718
Secure hardware (AES) (s)	4.65	8.64

[a] SWHE: somewhat homomorphic encryption.

[b] AES: Advanced Encryption Standard.

Table 5. Storage overhead for the secure hardware approach.

Overhead before and after encryption	Dataset	
	Haberman	Low Birth Weight Study
Before encryption (kB)	3.8	28
After encryption (SWHE[a]) (MB)	30.3	123
After encryption (AES[b]) (kB)	36	143

[a] SWHE: somewhat homomorphic encryption.

[b] AES: Advanced Encryption Standard.

We want to emphasize that, although the secure hardware (Advanced Encryption Standard [AES]) method is faster, state-of-the-art attack models targeting SGX show that solely secure hardware-based approaches might expose data from participating institutions to potential attackers (as explained above). Our method, although a little bit slower, preserves such institutional privacy by combining the local inputs without decrypting them; therefore, it offers a stronger security guarantee without imposing too much computation or storage cost. In this way, our proposed hybrid model provides a good trade-off in terms of security and efficiency.

Table 5 shows the storage overhead of the solely secure hardware-based approach. For SWHE, times required to encrypt the datasets were 4.37 minutes for the Haberman dataset and 18.46 minutes for the Low Birth Weight Study dataset. For AES, times required to encrypt the datasets were 14 milliseconds for the Haberman dataset and 38 milliseconds for the Low Birth Weight Study dataset.

Discussion

Comparison With Prior Work

There is a homomorphic encryption-based implementation of linear regression [14], which required 2 days to compute on a dataset containing 51,000 input vectors of 22 features with a key size of 1024 bits. That matrix inversion procedure took 1 day to complete because matrix inversion is a very expensive computational task in homomorphic encryption. In our proposed method, we performed matrix inversion on plaintext in secure hardware, which is much more efficient.

Hall et al [14] proposed an iterative matrix inversion algorithm, which introduces approximation errors when a fixed number of iterations is used. Their method offers a low accuracy of 10^{-3}. Precision can be slightly improved by choosing greater values for the 2 constants used by their method. However, this would require a larger public key, which would introduce significant computation overhead. In contrast, in our proposed method, there is no approximation error: the regression coefficients are completely identical to the plaintext results.

Security Discussions

In the Methods (Threat Model subsection), we discussed the security of SGX, specifically different side-channel attacks on SGX, and how we treat those attacks in our proposed framework. Addressing these attacks, we developed our framework in such a way that it can protect institutional privacy by combining the local inputs of participating institutions without decrypting them. This approach provides a higher layer of security without imposing too much computational cost.

In our proposed method, only intermediate values (eg, X^TY, X^TX) are decrypted inside secure hardware. Even if the hardware is compromised (or, in case of a side-channel attack), it is not possible to retrieve any sensitive attribute from those intermediary results. Hence, our proposed hybrid model not only achieves good performance but also guarantees stronger security than the solely SGX-based techniques. Dowlin et al [24] and Pass et al [25] discussed the security of SEAL and Intel SGX further.

A symmetric cryptosystem like AES requires *n* remote attestations to distribute the key to *n* data owners, which results in much more network communication, which might be prone to attack. In contrast, our proposed framework relies on public-key cryptography, where the data owners use a public key to encrypt their data published by the key manager. In this way, our proposed method reduces the attack surface of the system model, makes key distribution much simpler, and avoids additional communication overhead.

Limitations

There are some limitations of our proposed framework.

First, we did not consider the issue of model privacy. Several works based on differential privacy have addressed inference attacks (eg, model privacy [39]). These solutions are complementary to our proposed method and can be readily incorporated into a single framework.

Second, the central server of our proposed method must be SGX enabled; that is, it must use an Intel processor of sixth generation or later.

Third, since computing coefficients for logistic regression require multiple iterations, all parties must be synchronized until coefficients converge. However, linear regression does not require multiple iterations. So, in this case, parties can be offline just after sending their intermediary results.

Generalizability

Others have addressed training machine learning models (eg, support vector machines [40]) over distributed data [41,42]. Our proposed method can be easily applied to this kind of technique.

Cost of Deployment

The Intel SGX feature is available in all Intel Skylake and Kaby Lake processors. The price of an Intel Skylake or Kaby Lake processor is identical to that of processors from other vendors (having similar configuration). Price ranges from US $42 to US $1207 depending on configuration [43]. Recently, Microsoft started using SGX-capable servers in their Azure confidential computing service [44]. Azure confidential computing is offering the developers the ability to develop applications on top of Intel SGX software development kit. Apparently, there will be no significant additional charge for using this service.

Conclusion

In this age of big data, data need to be analyzed to uncover valuable insights and patterns. But this kind of analysis poses a threat to individual privacy, since data often contain sensitive information. In this paper, we address this data security and privacy issue and propose a hybrid cryptographic framework to overcome the limitations of the existing cryptographic techniques. We think that secure hardware–assisted predictive analysis of biomedical data is very promising for health care and medical research.

In future work, we will investigate the applicability of our proposed method to other learning algorithms such as neural networks, support vector machines, and decision trees.

Acknowledgments

This work was funded in part by the National Human Genome Research Institute (R00HG008175) and the National Institute of Biomedical Imaging and Bioengineering (U01EB023685), the Natural Sciences and Engineering Research Council of Canada Discovery Grants (RGPIN-2015-04147), the National Institute of General Medical Sciences (R01GM118574 and R01GM114612), and the University Research Grants Program from the University of Manitoba, Winnipeg, Manitoba, Canada.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Related works.

[PDF File (Adobe PDF File), 76KB - medinform_v6i1e14_app1.pdf]

References

1. Tabaei BP, Herman WH. A multivariate logistic regression equation to screen for diabetes: development and validation. Diabetes Care 2002 Nov;25(11):1999-2003. [Medline: 12401746]
2. Abdullah S, Murnane E, Matthews M, Kay M, Kientz J, Gay G. Cognitive rhythms: unobtrusive and continuous sensing of alertness using a mobile phone. 2016 Presented at: ACM International Joint Conference on Pervasive and Ubiquitous Computing; Sep 12-16, 2016; Heidelberg, Germany p. 178-189.
3. Rahman T, Czerwinski M, Gilad-Bachrach R, Johns P. Predicting about-to-eat moments for just-in-time eating intervention. 2016 Presented at: 6th International Conference on Digital Health; Apr 11-13, 2016; Montreal, QC, Canada p. 141-150.
4. Ahmadi H, Pham N, Ganti R, Abdelzaher T, Nath S, Han J. Privacy-aware regression modeling of participatory sensing data. 2010 Presented at: 8th ACM Conference on Embedded Networked Sensor Systems; Nov 3-5, 2010; Zurich, Switzerland p. 99-112.
5. El Emam K, Hu J, Mercer J, Peyton L, Kantarcioglu M, Malin B, et al. A secure protocol for protecting the identity of providers when disclosing data for disease surveillance. J Am Med Inform Assoc 2011 May 01;18(3):212-217. [doi: 10.1136/amiajnl-2011-000100] [Medline: 21486880]
6. Council of Canadian Academies. Accessing health and health-related data in Canada: the Expert Panel on Timely Access to Health and Social Data for Health Research and Health System Innovation. Ottawa, ON: Council of Canadian Academies; 2015. URL: http://www.scienceadvice.ca/uploads/eng/assessments%20and%20publications%20and%20news%20releases/health-data/healthdatafullreporten.pdf [accessed 2018-02-20] [WebCite Cache ID 6xNM7ZqMr]
7. Hayden EC. Geneticists push for global data-sharing. Nature 2013 Jun 06;498(7452):16-17. [doi: 10.1038/498017a] [Medline: 23739403]
8. Yao A. Protocols for secure computations. 1982 Presented at: 23rd Annual Symposium on Foundations of Computer Science; Nov 3-5, 1982; Chicago, IL, USA p. 160-164. [doi: 10.1109/SFCS.1982.88]
9. Gentry C. A Fully Homomorphic Encryption Scheme [doctoral thesis]. Stanford, CA: Stanford University; 2009.
10. Dwork C. Differential privacy. 2006 Presented at: 33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006); July 10-14, 2006; Venice, Italy p. 1-12.
11. Hoekstra M, Lal R, Pappachan P, Phegade V, Del Cuvillo J. Using innovative instructions to create trustworthy software solutions. 2013 Presented at: HASP@ ISCA 2013; June 23-24, 2013; Tel-Aviv, Israel p. 11.

12. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. J Am Med Inform Assoc 2012;19(5):758-764 [FREE Full text] [doi: 10.1136/amiajnl-2012-000862] [Medline: 22511014]
13. Shi H, Jiang C, Dai W, Jiang X, Tang Y, Ohno-Machado L, et al. Secure Multi-pArty Computation Grid LOgistic REgression (SMAC-GLORE). BMC Med Inform Decis Mak 2016 Jul 25;16 Suppl 3:89 [FREE Full text] [doi: 10.1186/s12911-016-0316-1] [Medline: 27454168]
14. Hall R, Fienberg SE, Nardi Y. Secure multiple linear regression based on homomorphic encryption. J Off Stat 2011;27(4):669.
15. Laine K, Player R. Simple Encrypted Arithmetic Library-SEAL (v2. 0). Technical report. Redmond, WA: Microsoft Research; 2016 Sep. URL: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/09/sealmanual.pdf [accessed 2018-02-20] [WebCite Cache ID 6xNMHiO5F]
16. Xu Y, Cui W, Peinado M. Controlled-channel attacks: deterministic side channels for untrusted operating systems. 2015 Presented at: IEEE Symposium on SecurityPrivacy; May 18-20, 2015; San Jose, CA, USA p. 640-656.
17. Rivest R, Adleman L, Dertouzos M. On data banks and privacy homomorphisms. Found Secur Comput 1978;4(11):169-180.
18. Boneh D, Goh E, Nissim K. Evaluating 2-DNF formulas on ciphertexts. In: Kilian J, editor. Theory of Cryptography. Cham, Switzerland: Springer International Publishing AG; 2005:325-341.
19. Goldwasser S, Micali S. Probabilistic encryption: how to play mental poker keeping secrl partial information. 1982 Presented at: Fourteenth Annual ACM Symposium on Theory of Computing; May 5-7, 1982; San Francisco, CA, USA p. 365-377.
20. Paillier P. Public-key cryptosystems based on composite degree residuosity classes. In: Stern J, editor. Advances in Cryptology-EUROCRYPT ’99. Cham, Switzerland: Springer International Publishing AG; 1999:223-238.
21. Rivest R, Shamir A, Adleman L. A method for obtaining digital signatures and public-key cryptosystems. Commun ACM Feb 1978;21(2):120-126.
22. Elgamal T. A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Trans Inf Theory Jul 1985;31(4):469-472.
23. Brakerski Z, Gentry C, Vaikuntanathan V. (Leveled) fully homomorphic encryption without bootstrapping. ACM Trans Comput Theory 2014;6(3):13.
24. Dowlin N, Gilad-Bachrach R, Laine K, Lauter K, Naehrig M, Wernsing J. Manual for using homomorphic encryption for bioinformatics. Proc IEEE 2017 Mar;105(3):552-567.
25. Pass R, Shi E, Tramer F. Formal abstractions for attested execution secure processors. In: Coron JS, Nielsen JB, editors. Advances in Cryptology - EUROCRYPT 2017. Cham, Switzerland: Springer International Publishing AG; 2017.
26. Goldreich O. Foundations of Cryptography. Volume 2: Basic Applications. Cambridge, UK: Cambridge University Press; 2009.
27. Fisch B, Vinayagamurthy D, Boneh D, Gorbunov S. IACR Cryptology ePrint Archive. 2016. IRON: functional encryption using Intel SGX URL: https://eprint.iacr.org/2016/1071.pdf [accessed 2018-02-20] [WebCite Cache ID 6xMvkInRB]
28. Weichbrodt N, Kurmus A, Pietzuch P, Kapitza R. Asyncshock: exploiting synchronisation bugs in Intel SGX enclaves. 2016 Presented at: 21st European Symposium on Research in Computer Security; Sep 26-30, 2016; Heraklion, Crete, Greece p. 440-457.
29. Shinde S, Chua Z, Narayanan V, Saxena P. Preventing page faults from telling your secrets. 2016 Presented at: 11th ACM on Asia Conference on Computer and Communications Security; May 30-Jun 3, 2016; Sian, China p. 317-328.
30. Wang W, Chen G, Pan X, Zhang Y, Wang X, Bindschaedler V, et al. Leaky cauldron on the dark land: understanding memory side-channel hazards in SGX. arXiv:1705.07289. 2017 Aug 30. URL: https://arxiv.org/abs/1705.07289 [accessed 2018-02-14] [WebCite Cache ID 6xED5tDYT]
31. Costan V, Lebedev I, Devadas S. Sanctum: minimal hardware extensions for strong software isolation. 2016 Presented at: 25th USENIX Security Symposium; Aug 10-12, 2016; Austin, TX, USA p. 857-874.
32. Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence informationbasic countermeasures. 2015 Presented at: 22nd ACM SIGSAC Conference on Computer and Communications Security; Oct 12-16, 2015; Denver, CO, USA p. 1322-1333.
33. Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T. Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. 2014 Presented at: 23rd USENIX Security Symposium; Aug 20-22, 2014; San Diego, CA, USA p. 17-32.
34. Heij C, de Boer P, Franses P, Kloek T, van Dijk HK. Econometric Methods With Applications in Business and Economics. Oxford, UK: Oxford University Press; 2004.
35. Friedman J, Hastie T, Tibshirani R. The Elements of Statistical Learning. Springer Series in Statistics. Berlin, Germany: Springer; 2001.
36. Halevi S, Shoup V. Algorithms in HElib. In: Garay JA, Gennaro R, editors. Advances in Cryptology - CRYPTO 2014. Cham, Switzerland: Springer International Publishing AG; 2014:554-571.
37. Lichman M. UCI Machine Learning Repository. Irvine, CA: University of California, Irvine, School of Information and Computer Sciences; 2013. URL: http://archive.ics.uci.edu/ml/ [accessed 2018-02-14] [WebCite Cache ID 6xEDJtnq9]
38. Hosmer DJ, Lemeshow S, Sturdivant R. Applied Logistic Regression. New York, NY: John Wiley & Sons; 2013.

39. Abadi M, Chu A, Goodfellow I, McMahan H, Mironov I, Talwar K. Deep learning with differential privacy. 2016 Presented at: ACM SIGSAC Conference on Computer and Communications Security; Oct 24-28, 2016; Vienna, Austria p. 308-318.
40. Vapnik V. The Nature of Statistical Learning Theory. Cham, Switzerland: Springer International Publishing AG; 2013.
41. Yu H, Jiang X, Vaidya J. Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data. 2006 Presented at: ACM Symposium on Applied Computing; Apr 23-27, 2006; Dijon, France p. 603-610.
42. Yu H, Vaidya J, Jiang X. Privacy-preserving SVM classification on vertically partitioned data. In: Ng WK, Kitsuregawa M, Li J, Chang K, editors. Advances in Knowledge Discovery and Data Mining. Cham, Switzerland: Springer International Publishing AG; 2006:647-656.
43. Products formerly Skylake. Santa Clara, CA: Intel Corporation URL: http://ark.intel.com/products/codename/37572/Skylake [accessed 2017-10-11] [WebCite Cache ID 6u8vlW2Om]
44. Russinovich M. Introducing Azure confidential computing. Seattle, WA: Microsoft; 2017 Sep 14. URL: https://azure.microsoft.com/en-us/blog/introducing-azure-confidential-computing/ [accessed 2017-10-11] [WebCite Cache ID 6u8ux4Vmd]

Abbreviations

AES: Advanced Encryption Standard
FHE: fully homomorphic encryption
GLORE: grid binary logistic regression
HIPAA: Health Insurance Portability and Accountability Act
PIPEDA: Personal Information Protection and Electronic Documents Act
SEAL: Simple Encrypted Arithmetic Library
SGX: Software Guard Extensions
SWHE: somewhat homomorphic encryption

Edited by G Eysenbach; submitted 23.06.17; peer-reviewed by K Shimizu, J Bian, L Zhou; comments to author 31.08.17; revised version received 25.10.17; accepted 03.01.18; published 05.03.18.

Please cite as:
Sadat MN, Jiang X, Aziz MMA, Wang S, Mohammed N
Secure and Efficient Regression Analysis Using a Hybrid Cryptographic Framework: Development and Evaluation
JMIR Med Inform 2018;6(1):e14
URL: http://medinform.jmir.org/2018/1/e14/
doi:10.2196/medinform.8286
PMID:29506966

Original Paper

Assessing the Readability of Medical Documents: A Ranking Approach

Jiaping Zheng[1], MS; Hong Yu[1,2,3,4], PhD, FACMI

[1]College of Information and Computer Sciences, University of Massachusetts, Amherst, MA, United States

[2]Center for Healthcare Organization and Implementation Research, Bedford Veterans Affairs Medical Center, Bedford, MA, United States

[3]Department of Computer Science, University of Massachusetts, Lowell, MA, United States

[4]Department of Medicine, University of Massachusetts Medical School, Worcester, MA, United States

Corresponding Author:
Hong Yu, PhD, FACMI
Center for Healthcare Organization and Implementation Research
Bedford Veterans Affairs Medical Center
200 Springs Road
Bedford, MA, 01730
United States
Phone: 1 781 687 2000
Fax: 1 781 687 2000
Email: hong.yu@umassmed.edu

Abstract

Background: The use of electronic health record (EHR) systems with patient engagement capabilities, including viewing, downloading, and transmitting health information, has recently grown tremendously. However, using these resources to engage patients in managing their own health remains challenging due to the complex and technical nature of the EHR narratives.

Objective: Our objective was to develop a machine learning–based system to assess readability levels of complex documents such as EHR notes.

Methods: We collected difficulty ratings of EHR notes and Wikipedia articles using crowdsourcing from 90 readers. We built a supervised model to assess readability based on relative orders of text difficulty using both surface text features and word embeddings. We evaluated system performance using the Kendall coefficient of concordance against human ratings.

Results: Our system achieved significantly higher concordance (.734) with human annotators than did a baseline using the Flesch-Kincaid Grade Level, a widely adopted readability formula (.531). The improvement was also consistent across different disease topics. This method's concordance with an individual human user's ratings was also higher than the concordance between different human annotators (.658).

Conclusions: We explored methods to automatically assess the readability levels of clinical narratives. Our ranking-based system using simple textual features and easy-to-learn word embeddings outperformed a widely used readability formula. Our ranking-based method can predict relative difficulties of medical documents. It is not constrained to a predefined set of readability levels, a common design in many machine learning–based systems. Furthermore, the feature set does not rely on complex processing of the documents. One potential application of our readability ranking is personalization, allowing patients to better accommodate their own background knowledge.

(JMIR Med Inform 2018;6(1):e17) doi:10.2196/medinform.8611

KEYWORDS

electronic health records; readability; comprehension; machine learning

Introduction

Background

Research has demonstrated that actively involving patients in the management of their own health can lead to better outcomes, and potentially lower costs [1,2]. Patient engagement [3]—a concept that includes patient activation, and interventions designed to increase activation and promote positive patient behavior—has thus emerged as an important component of strategies to improve health care. A growing body of evidence has accumulated on better health outcomes and care experiences associated with higher engagement. For example, patients with chronic diseases who have high patient activation measure scores are more likely to practice self-management behaviors and report high medication adherence [4]. High patient activation measure scores are also associated with a high likelihood of clinical indicators (eg, hemoglobin A_{1c}, high-density lipoprotein, and triglycerides) being in the normal range [1].

The use of electronic health record (EHR) systems with patient engagement capabilities, including viewing, downloading, and transmitting health information, has recently grown tremendously. According to data from the US Office of the National Coordinator for Health Information Technology, the percentage of hospitals that enable patients to electronically view, download, and transmit their health information grew almost 7-fold between 2013 and 2015 [5]. In 2015, 95% of hospitals provided their patients with the ability to view their information.

However, actively engaging patients in the management of their own health remains challenging, despite the evidence of better health care outcomes and potentially lower costs. Access to EHRs by itself is not sufficient to motivate patients to be involved because of the complex and technical nature of the EHR. Patients without training in medicine may struggle to process and understand the information buried in the technical language in EHRs. In fact, materials beyond patients' reading abilities are widely reported in the literature [6-10]. The lack of explanation that an expert can provide when reading EHR notes may also engender unnecessary anxiety or confusion [11]. Furthermore, many patients have limited health literacy and are not proficient in completing tasks considered essential to successfully navigate the health system and act on health information [12].

Therefore, assessing the difficulty of EHR notes and integrating appropriate educational assistance in EHR systems may make them more accessible for a layperson without professional training in medicine. In this study, we explored methods to automatically assess the readability levels of clinical narratives in EHRs and other complex documents. An accurate assessment of these documents can be used to match patients' literacy levels, facilitating patient activation and engagement.

Prior Work

The research community has relied on readability formulas to assess a variety of information materials for patients. Numerous readability metrics have been developed to assess the grade level or the number of years of education needed for a person to understand the content. One of the most widely used in the health domain is the Flesch-Kincaid Grade Level [13] (FKGL), which predicts a grade level based on the average sentence length and the average word length. Other similar metrics are the Simple Measure of Gobbledygook, Gunning Fog Index, Coleman-Liau Index, and New Dale-Chall formula. These metrics rely on the assumption that the longer the words and the sentences are, the more difficult the text is. However, this assumption does not hold for EHR narratives, as sentences are usually short and abbreviations are common.

There were also efforts in the health care domain to develop instruments for medical documents. One measurement proposed by Kim et al [14] compared differences in surface text, syntactic features, and semantic features with a known set of easy and difficult documents and reported normalized scores. Another method for health text was based on a naive Bayes classifier [15]. Those authors collected training documents from blogs, patient education documents, and medical journals. They used vocabularies in these documents as features for the classifier. Both of the methods relied on manually curated documents.

Goal of This Work

In this work, we considered measuring readability as a ranking task, where the relative difficulty of documents is compared. Readability in the health domain is often measured with formulas developed to ensure that school textbooks are appropriate for children at a particular school grade level [16]. However, obtaining a grade level often is not the ultimate goal. The document's grade level is usually compared with a person's educational level or another document's grade level in order to find appropriate reading materials. The number of years of education has been challenged as a proxy measure for one's educational experiences when measuring cognitive functions. One study has shown that, in a sample of elderly African Americans, nearly 30% read 3 or more years below their self-reported educational level [17]. Other studies have also advocated the use of reading or literacy ability instead of years of education to account for variance in neuropsychological assessments [18,19].

Therefore, ranking the readability of documents is well suited to applications whose main concern is to match difficulty levels with existing text or to identify easier or more difficult ones, rather than to obtain an absolute score. For example, a patient-facing EHR system may learn from its users' reactions to infer their reading ability and present appropriate educational materials. Such a system can be personalized for an individual user. A user with limited literacy will only see straightforward materials, whereas higher-quality materials that require higher literacy levels can be presented to an advanced user. This personalization is a first step toward user-centered care. To this end, we developed a machine learning model to compare the relative difficulty of documents using data collected from Amazon Mechanical Turk (AMT) users. A demonstration website is available [20].

Methods

Data

We collected difficulty levels on health-related documents from human annotators. We recruited users on AMT (Amazon.com, Inc, Seattle, WA, USA) to read and rate pairs of documents based on their perceived difficulty. We screened AMT users to be from the United States and having an approval rating of at least 95% in prior tasks. Each reader was presented with 20 randomly selected pairs of documents side by side on the computer screen. The readers were requested to rate the readability of the documents on a scale from 1 (easiest to understand) to 10 (most difficult to understand). The setup to show 2 documents helped reduce variation when we assembled the ratings into a complete ranking, as it provided explicit partial ranking, as opposed to implicit order inferred from the difficulty ratings.

The 2 documents in each document pair were of similar length (within a 50-token difference, where a token is a word or term) and comparable difficulty according to FKGL (within 0.5 grade level). We sourced the documents from English Wikipedia articles and deidentified EHR notes written by physicians. The 20 document pairs consisted of 5 pairs of Wikipedia documents, 5 pairs of EHR documents, and 10 pairs of mixed-source documents.

We selected 3 common diseases as topics from the document sources: cancer, diabetes, and hypertension. Wikipedia documents were randomly selected from all article pages up to 3 levels under the disease category page, following the category structure. EHR notes were selected using *International Classification of Diseases, Ninth Revision* codes (140-195 for cancer, 250.00-250.93 for diabetes, and 401.0-401.9 for hypertension). For each disease topic, we collected data from 30 AMT users. In total, 90 AMT users annotated 900 document pairs, with 927 of the documents being unique. Table 1 shows the statistics of the documents annotated by these users.

Machine Learning System

Learning to Rank

We developed a supervised learning system for EHR readability. Traditionally, readability is measured at school grade levels. Formulas that are widely used in the health care domain include the FKGL, Simple Measure of Gobbledygook, Gunning Fog Index, Coleman-Liau Index, and New Dale-Chall formula. They all use a limited number of factors, mostly word and sentence lengths, to estimate a document's grade level. These simple features, however, are not able to fully capture the complexity of medical documents when used alone as in the formulas. For instance, EHR narratives often contain abbreviations and lists, which are treated as short words and sentences, thus lowering the estimated grade level. However, the abbreviations present a great challenge to a layperson's understanding [21,22].

In the machine learning community, many systems were developed to classify documents into a predefined set of readability levels. Such systems can include a multitude of features, including lexical, syntactic, and discourse features. These methods are nevertheless constrained in the granularity that they can estimate, since the predefined difficulty levels are often limited.

In our work, we approached readability as a ranking problem, in which the difficulty levels between documents are compared. This approach overcomes the problems in both the traditional formulas and the classification methods: we are not solely reliant on word and sentence lengths as in the formulas, and our approach can order readability levels for a set of documents.

We trained our ranking system using a pairwise approach. From each user's documents, we generated a training example from any 2 documents that were assigned different difficulty levels.

A support vector machine (SVM) model was learned from the pairwise comparisons of AMT users' assigned document difficulty levels using the SVM$^{\text{rank}}$ package [23]. SVM models normally optimize a hinge loss function based on a binary label for every training example. In the pairwise scenario, the objective is to minimize the number of discordant pairs—that is, pairs that are ordered incorrectly with respect to the true order. More formally, given a set of training examples $\{(\mathbf{x}_i, y_i)\}$, the primal form of the problem is as the equation in Figure 1 shows, where **w** is the weight vector, *C* parameterizes the trade-off between training error and margin size, and ξ is slack variables. Rearranging the first constraint, $\mathbf{w}^T(\mathbf{x}_i–\mathbf{x}_j)>1–\xi_{i,j}$, which is equivalent to a classic SVM problem on the modified input vectors $\mathbf{x}'= \mathbf{x}_i–\mathbf{x}_j$. Therefore, a binary classification SVM optimizer can be used to solve the problem.

In our dataset, we generated pairwise difference vectors **x′** from each AMT user's ratings. The difference vectors were not generated from different users because ratings across users may not form a consistent ranking, as those from a single user do. For example, a vector was generated from 2 documents, A and B, by 1 user, but not from 2 documents from different users.

Table 1. Statistics of documents annotated by readers.

Source and disease	Documents (n)	Sentences (n)	Tokens[a] (n)
Wikipedia			
Cancer	215	2510	46,349
Diabetes	74	1352	33,402
Hypertension	85	2007	45,440
EHR[b] notes			
Cancer	127	2067	37,830
Diabetes	195	6335	81,085
Hypertension	231	6594	90,784
Total	927	20,865	334,890

[a] A token is, loosely, a word or term.

[b] EHR: electronic health record.

Figure 1. The primal form of pairwise ranking.

$$\begin{aligned} \min \quad & \mathbf{w}^2 + C \sum \xi_{i,j} \\ \text{s.t.} \quad & \mathbf{w}^T \mathbf{x}_i \geq \mathbf{w}^T \mathbf{x}_j + 1 - \xi_{i,j}, \forall y_i > y_j \\ & \xi_{i,j} \geq 0, \forall i, j \end{aligned}$$

Features

We employed several types of features, including those from traditional readability formulas. We included average words per sentence, average syllables per word from the FKGL formula, proportion of polysyllabic words (words with more than 3 syllables) from the Gunning Fog Index, and percentage of difficult words from the New Dale-Chall formula. Although these formulas do not correlate well with human perceptions of difficulty [24], these word length–based features are useful at capturing some longer medical jargon (eg, Huntington disease). There is also evidence that the perceived difficulty of a word is correlated with its length [25]. We also included word frequency obtained from the Wikipedia documents and EHR notes, since common words have been found likely to be perceived as easier to understand [25]. We grouped the frequencies into 10 bins and used the percentage of words in each bin as features. Additional features included document length measured in words and sentences. Long documents require more cognitive processing to comprehend, which might translate to higher perceived difficulty. Lastly, we captured language patterns using 2 word embeddings learned separately from Wikipedia documents and deidentified EHR notes. We used Word2vec [26] to learn a 200-dimensional skip-gram embedding.

Results

System Performance

We split the annotated data three ways, into training (60%), development (20%), and test (20%) sets. The 3 disease topics were stratified in the split. Hyperparameters were optimized on the development set. We obtained final test results from a model trained using the optimized parameters.

We evaluated our system using the Kendall coefficient of concordance (*W*) [27], a statistic that measures the agreement between rankings from multiple raters. The coefficient aggregates the ranks assigned to each item from all raters and measures the variance. The variance is then normalized to be between 0 and 1. Higher values represent a high level of concordance. In our experiments, for each AMT user, we ordered his or her documents by their assigned difficulty levels and calculated *W* with the order generated from our system prediction. We then averaged the *W* coefficients of all the users.

Table 2 shows our system's performance, in the row "new system." The next rows show different experiment settings discussed in the next two sections. As a baseline, we evaluated the performance of the widely used FKGL readability formula. The average agreement between this formula and the AMT annotators was .531. Our system achieved an agreement of .734 with the AMT annotators, outperforming the FKGL baseline by 38.3%. The increase is statistically significant as assessed by a Wilcoxon signed rank test at the *P*=.05 level.

We also trained and tested separate models for each of the disease topics following the same process. Our system showed consistent improvement over the baseline across all disease categories. Agreement in the diabetes and hypertension categories increased significantly over the baseline FKGL metric. The cancer category improved substantially, but not significantly, over the baseline. These results suggested that our method is robust across different topics.

User Behavior

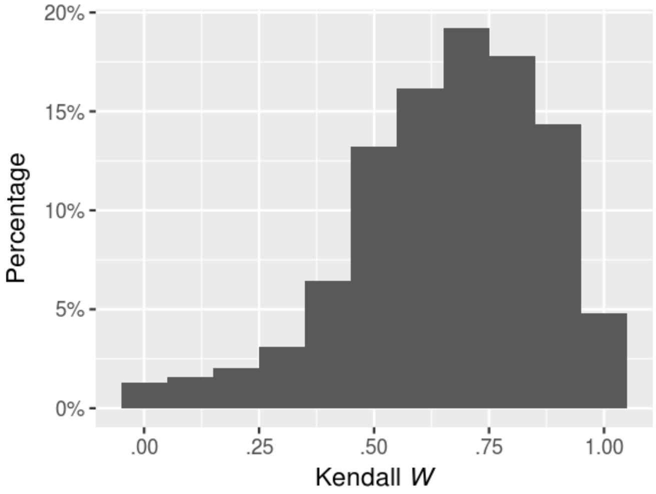
A variety of factors may influence a reader's reading comprehension, which in turn determines his or her judgment on a document's difficulty. We examined the differences in the AMT users' difficulty ratings using the same Kendall *W* coefficient. We calculated *W* for each pair of users' ranked documents. The average concordance between any 2 users was .658. Figure 2 shows the distribution of concordance between any 2 users in our dataset.

While there are pairs of users whose concordance was low, most (851/1299, 65.51%) had a concordance greater than .6. When examined on an individual level, the low concordance can often be attributed to a few users who appeared to disagree with many others. There were 9 users who had a less than .5 concordance with more than 10 other users. Furthermore, 5 of these users' mean concordance with other users was less than .5.

Table 2. System performance (Kendall *W*) compared with baseline for specific disease topics and with partial datasets. Numbers in parentheses are percentage improvements over FKGL (Flesch-Kincaid Grade Level). *P* values are comparisons with FKGL using a Wilcoxon signed rank test.

System	Cancer		Diabetes		Hypertension		All	
System	Kendall *W*	*P* value	Kendall *W*	*P* value	Kendall *W*	*P* value	Kendall *W*	*P* value
FKGL (baseline)	.541		.490		.561		.531	
New system	.656 (+21.3)	.08	.790 (+61.3)	.02	.715 (+27.5)	.03	.734 (+38.3)	<.001
New system with data subsets excluded								
Excluding eccentric users	.694 (+28.3)	.03	.762 (+55.5)	.02	.727 (+29.6)	.03	.722 (+36.0)	<.001
Excluding controversial documents	.650 (+20.1)	.05	.790 (+61.3)	.02	.759 (+35.2)	.02	.737 (+39.0)	<.01

Figure 2. Histogram of Kendall *W* evaluating readability ratings between any 2 Amazon Mechanical Turk users.



To measure a user's conformity in relation to others, we calculated the mean Kendall *W* between individual users and all of their peers. Figure 3 shows the distribution.

Approximately one-third of the users were highly conforming (mean *W*≥.7) with others, whereas 7% (6/90) were eccentric (mean *W*<.5). This result suggests that, despite individual differences in their background knowledge about the subject matter, AMT users still exhibited a consensus on a document's difficulty level. We also noted that our system was able to predict readability orders similar to those of a "regular" user. Our system's mean *W* was highly correlated with a user's conformity (ρ=.85). In contrast, the FKGL formula's predicted grade levels did not show a strong correlation (ρ=–.13) with conformity.

Table 2 (row "–eccentric users") shows the performance of models trained from data excluding eccentric users. All disease topics performed significantly better with our system than with FKGL. Our system's performance on the combined disease topics, also significantly higher than with FKGL, was slightly lower than with the system using the full dataset. This could be due to the large amount of samples removed from training even when we excluded only a small number of users, because the difference vectors were generated from all possible pairwise comparisons. On the individual disease topic level, however, the cancer and hypertension models outperformed our system when trained on the full training data.

Controversial Documents

In addition to annotator differences, another factor that contributes to inconsistent annotations is the nature of the documents. We postulated that some documents may have been challenging for the AMT users. For example, certain types of domain-specific writing may appear easy to understand to some but not all users, leading to inconsistent user ratings. These "controversial" documents would also have confused our system, which attempted to learn from the conflicting human annotation. To highlight the range of AMT users' perceptions of difficulty, Figure 4 shows the maximum difference in ratings assigned by AMT users to documents that were rated by at least two users (n=597).

The mean difference was 3.8, suggesting that users' perceptions of difficulty varied considerably. The 2 sources of documents (Wikipedia and EHR notes) contained approximately the same number of controversial documents (maximum difference >5), and the cancer topic had more such documents than the other 2 topics. We further trained new models after removing controversial documents from the dataset. Table 2 shows the performances of these models in the last row ("Excluding controversial documents"). Performance of 2 categories, cancer and diabetes, remained similar to those of the models trained from the full dataset. The hypertension set increased appreciably.

Feature Ablation

We compared the contribution of the different types of features included in our system. We trained separate models without the word frequency–based features, readability formula features, word length–based features, and word embedding–based features. Table 3 shows the performance of these models.

Excluding word embeddings resulted in the largest decrease in performance. The word frequency–based features did not appear to contribute much to the overall performance. Removing these features resulted in only a 0.1% performance decrease. This could be due to the nature of the word frequency corpus (a general English corpus without any particular emphasis on any domain) we used to calculate these features. The surface text characteristics captured by the formulas showed a moderate contribution, although they were not reliable stand-alone indicators. With the exception of 1 case, the contributions of the features were consistent across different disease topics—word embedding and word length–based features being the highest and word frequency the lowest.

Figure 3. Histogram of individual Amazon Mechanical Turk users' conformity (measured by the mean of Kendall *W* against their peers).

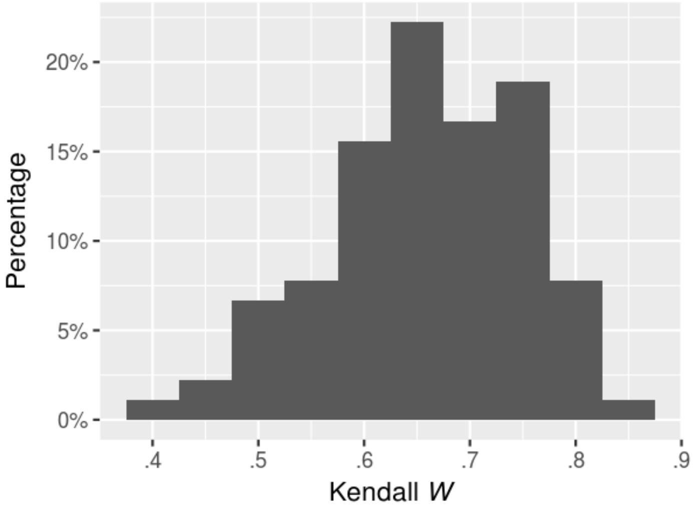


Figure 4. Histogram of maximum differences in Amazon Mechanical Turk users' ratings of documents rated by at least two users.

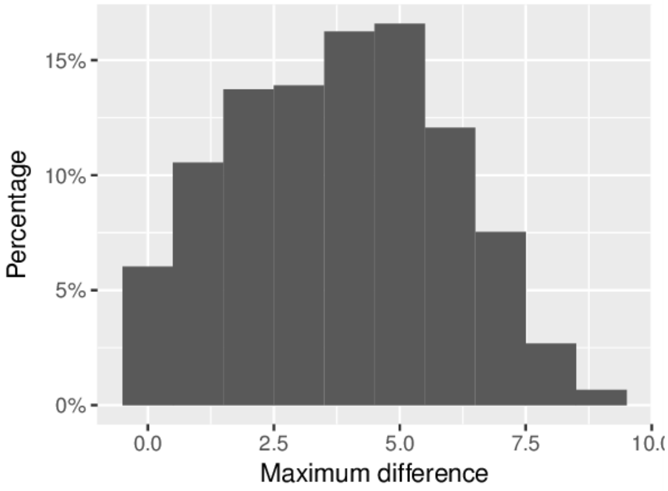


Table 3. Model performance (Kendall *W*) with feature ablation.

Feature set	Cancer	Diabetes	Hypertension	All
Full[a]	.656	.790	.715	.734
Excluded feature				
Frequency	.652	.792	.710	.733
Formula	.648	.789	.709	.728
Length	.636	.785	.702	.716
Embedding	.677	.784	.703	.714

[a] The system with all proposed features included (data from Table 2).

Discussion

Principal Findings

We explored methods to automatically assess the readability levels of clinical narratives. Our ranking-based system using simple textual features and easy-to-learn word embeddings outperformed predictions from applying FKGL. In all of the disease topics we assessed, our method achieved an over 20% increase, with the majority of cases showing higher and statistically significance increases.

Limitations

One limitation of our method is that it may be necessary to prune inconsistent data before training a model. Some users' perceptions of document readability may exhibit a different pattern from others'. Including conflicting data points may result in suboptimal models. A future study direction is to explore the trade-off between expert and crowdsourced annotations.

Another limitation is that we trained our model on AMT users' perceived document difficulty, which can be different from a linguistic perspective.

Comparison With Other Methods

We applied a learning-to-rank approach to readability assessment, whereby we used comparisons of relative difficulty to train a model and, similarly, to predict an order based on document difficulty. Existing machine learning–based systems are usually designed around classification. They are often limited to a few predefined labels [15] or require corpora labeled at distinct levels [14]. The advantage of our approach is that we do not need expert annotation of grade levels on documents, and annotation may be crowdsourced as in our experiments. Acquiring more personalized training examples is also possible without explicit curation, as user actions may be implicitly mined to generate document difficulty comparisons, by using information retrieval methods.

Furthermore, unlike many other machine learning–based methods that require deep natural language processing, such as parsing [28] and discourse analysis [29], our choice of feature set is relatively simple. The surface features from readability formulas and word frequencies were both easy to calculate. Well-established tools also exist to generate word embeddings from large corpora. Therefore, our system could be easily deployed in an EHR system.

Lastly, although traditional readability formulas are very easy to use by nontechnical users, as they do not require training a machine learning model, they are inaccurate in determining the difficulty of complex documents. With simple features and widely available software packages, our proposed method is straightforward to implement.

Conclusions

Patients' access to their EHR notes has increased dramatically according to US national statistics. However, actively engaging patients in the management of their own health remains challenging. Assessing the readability of EHR notes and integrating educational assistance may make these notes more accessible for a layperson without professional training in medicine. To this end, we developed a new machine learning–based method to assess EHR readability from relative orders of text difficulty. We trained a learning-to-rank system to predict relative difficulty levels of given documents, instead of using the traditional classification approach, in which documents are assigned levels from a limited predefined set of values. Our experiments showed that this method significantly outperformed the widely used FKGL formula, and the improvement was consistent across different topics. Our system's average concordance with an individual human user's ratings was higher than the concordance between different human annotators. This method can potentially be personalized to individual users to better accommodate their background knowledge.

Acknowledgments

This work was supported in part by the Investigator Initiated Research grant 1I01HX001457-01 from the Health Services Research & Development Program of the US Department of Veterans Affairs, and in part by the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, USA. The content is solely the responsibility of the authors and does not represent the views of the US Department of Veterans Affairs, the US Government, or the Center for Intelligent Information Retrieval.

Conflicts of Interest

None declared.

References

1. Greene J, Hibbard JH. Why does patient activation matter? An examination of the relationships between patient activation and health-related outcomes. J Gen Intern Med 2012 May;27(5):520-526 [FREE Full text] [doi: 10.1007/s11606-011-1931-2] [Medline: 22127797]
2. Begum N, Donald M, Ozolins IZ, Dower J. Hospital admissions, emergency department utilisation and patient activation for self-management among people with diabetes. Diabetes Res Clin Pract 2011 Aug;93(2):260-267. [doi: 10.1016/j.diabres.2011.05.031] [Medline: 21684030]
3. Hibbard JH, Greene J. What the evidence shows about patient activation: better health outcomes and care experiences; fewer data on costs. Health Aff (Millwood) 2013 Feb;32(2):207-214. [doi: 10.1377/hlthaff.2012.1061] [Medline: 23381511]
4. Mosen DM, Schmittdiel J, Hibbard J, Sobel D, Remmers C, Bellows J. Is patient activation associated with outcomes of care for adults with chronic conditions? J Ambul Care Manage 2007 Mar;30(1):21-29. [Medline: 17170635]
5. Henry J, Pylypchuk Y, Patel V. Electronic capabilities for patients among U.S. non-federal acute care hospitals: 2012-2015. ONC data brief 38. Washington, DC: Office of the National Coordinator for Health Information Technology; 2016 Sep. URL: http://dashboard.healthit.gov/evaluations/data-briefs/hospitals-patient-engagement-electronic-capabilities-2015.php [accessed 2016-10-07] [WebCite Cache ID 6l5V1ZSLl]
6. Agarwal N, Hansberry DR, Sabourin V, Tomei KL, Prestigiacomo CJ. A comparative analysis of the quality of patient education materials from medical specialties. JAMA Intern Med 2013 Jul 8;173(13):1257-1259. [doi: 10.1001/jamainternmed.2013.6060] [Medline: 23689468]

7. Huang G, Fang CH, Agarwal N, Bhagat N, Eloy JA, Langer PD. Assessment of online patient education materials from major ophthalmologic associations. JAMA Ophthalmol 2015 Apr;133(4):449-454. [doi: 10.1001/jamaophthalmol.2014.6104] [Medline: 25654639]
8. Watad A, Bragazzi NL, Brigo F, Sharif K, Amital H, McGonagle D, et al. Readability of Wikipedia pages on autoimmune disorders: systematic quantitative assessment. J Med Internet Res 2017 Jul 18;19(7):e260 [FREE Full text] [doi: 10.2196/jmir.8225] [Medline: 28720555]
9. Brigo F, Otte WM, Igwe SC, Tezzon F, Nardone R. Clearly written, easily comprehended? The readability of websites providing information on epilepsy. Epilepsy Behav 2015 Mar;44:35-39. [doi: 10.1016/j.yebeh.2014.12.029] [Medline: 25601720]
10. Brigo F, Erro R. The readability of the English Wikipedia article on Parkinson's disease. Neurol Sci 2015 Jun;36(6):1045-1046. [doi: 10.1007/s10072-015-2077-5] [Medline: 25596713]
11. Davis GT, Singh H. Should patients get direct access to their laboratory test results? An answer with many questions. JAMA 2011 Dec 14;306(22):2502-2503. [doi: 10.1001/jama.2011.1797] [Medline: 22122864]
12. Koh HK, Brach C, Harris LM, Parchman ML. A proposed 'health literate care model' would constitute a systems approach to improving patients' engagement in care. Health Aff (Millwood) 2013 Feb;32(2):357-367. [doi: 10.1377/hlthaff.2012.1205] [Medline: 23381529]
13. Flesch R. A new readability yardstick. J Appl Psychol 1948 Jun;32(3):221-233. [Medline: 18867058]
14. Kim H, Goryachev S, Rosemblat G, Browne A, Keselman A, Zeng-Treitler Q. Beyond surface characteristics: a new health text-specific readability measurement. AMIA Annu Symp Proc 2007 Oct 11:418-422 [FREE Full text] [Medline: 18693870]
15. Leroy G, Miller T, Rosemblat G, Browne A. A balanced approach to health information evaluation: a vocabulary-based naïve Bayes classifier and readability formulas. J Am Soc Inf Sci 2008 Jul;59(9):1409-1419. [doi: 10.1002/asi.20837]
16. Redish J. Readability formulas have even more limitations than Klare discusses. ACM J Comput Doc 2000 Aug 01;24(3):132-137. [doi: 10.1145/344599.344637]
17. O'Bryant SE, Lucas JA, Willis FB, Smith GE, Graff-Radford NR, Ivnik RJ. Discrepancies between self-reported years of education and estimated reading level among elderly community-dwelling African-Americans: analysis of the MOAANS data. Arch Clin Neuropsychol 2007 Mar;22(3):327-332 [FREE Full text] [doi: 10.1016/j.acn.2007.01.007] [Medline: 17336494]
18. Manly JJ, Jacobs DM, Touradji P, Small SA, Stern Y. Reading level attenuates differences in neuropsychological test performance between African American and white elders. J Int Neuropsychol Soc 2002 Mar;8(3):341-348. [Medline: 11939693]
19. Manly JJ, Schupf N, Tang M, Stern Y. Cognitive decline and literacy among ethnically diverse elders. J Geriatr Psychiatry Neurol 2005 Dec;18(4):213-217. [doi: 10.1177/0891988705281868] [Medline: 16306242]
20. Zheng J, Yu H. Ranking readability demo. 2018. URL: http://bio-nlp.org/readability-ranking [accessed 2018-03-15] [WebCite Cache ID 6xwCIPM6x]
21. Keselman A, Slaughter L, Smith CA, Kim H, Divita G, Browne A, et al. Towards consumer-friendly PHRs: patients' experience with reviewing their health records. AMIA Annu Symp Proc 2007:399-403 [FREE Full text] [Medline: 18693866]
22. Pyper C, Amery J, Watson M, Crook C. Patients' experiences when accessing their on-line electronic patient records in primary care. Br J Gen Pract 2004 Jan;54(498):38-43 [FREE Full text] [Medline: 14965405]
23. Joachims T. Training linear SVMs in linear time. New York, NY: ACM; 2006 Presented at: 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 20-23, 2006; Philadelphia, PA, USA p. 217-226. [doi: 10.1145/1150402.1150429]
24. Zheng J, Yu H. Readability formulas and user perceptions of electronic health records difficulty: a corpus study. J Med Internet Res 2017 Mar 02;19(3):e59 [FREE Full text] [doi: 10.2196/jmir.6962] [Medline: 28254738]
25. Leroy G, Kauchak D. The effect of word familiarity on actual and perceived text difficulty. J Am Med Inform Assoc 2014 Feb;21(e1):e169-e172 [FREE Full text] [doi: 10.1136/amiajnl-2013-002172] [Medline: 24100710]
26. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013 Presented at: Workshop at ICLR; May 2-4, 2013; Scottsdale, AZ, USA.
27. Kendall MG, Smith BB. The problem of m rankings. Ann Math Stat 1939;10(3):275-287.
28. Schwarm S, Ostendorf M. Reading level assessment using support vector machines and statistical language models. Stroudsburg, PA: Association for Computational Linguistics; 2005 Presented at: 43rd Annual Meeting on Association for Computational Linguistics; Jun 25-30, 2005; Ann Arbor, MI, USA p. 523-530. [doi: 10.3115/1219840.1219905]
29. Feng L, Jansche M, Huenerfauth M, Elhadad N. A comparison of features for automatic readability assessment. 2010 Presented at: 23rd International Conference on Computational Linguistics (COLING); Aug 23-27, 2010; Beijing, China p. 287-284.

Abbreviations

AMT: Amazon Mechanical Turk
EHR: electronic health record

FKGL: Flesch-Kincaid Grade Level
SVM: support vector machine

Edited by G Eysenbach; submitted 03.08.17; peer-reviewed by G Leroy, N Adams, F Brigo; comments to author 20.11.17; revised version received 12.01.18; accepted 31.01.18; published 23.03.18.

Please cite as:
Zheng J, Yu H
Assessing the Readability of Medical Documents: A Ranking Approach
JMIR Med Inform 2018;6(1):e17
URL: http://medinform.jmir.org/2018/1/e17/
doi:10.2196/medinform.8611
PMID:29572199

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org